PyraNet: A Lightweight Neural Decoder for Diverse BCI Paradigms

Anonymous authors
Paper under double-blind review

Abstract

Neural signal decoding relies on learning effective representations of brain activity. Existing models have largely followed two paths: compact decoders and large pretrained foundation models. The former perform well on certain target paradigms but struggle on others. The latter seek broad generalizability through increased capacity and cross-dataset pretraining but often fail to deliver performance commensurate with their size. This raises a fundamental question: Must general-purpose neural decoding rely on scaling up model size, or can broad applicability be achieved with a lightweight architecture informed by domain knowledge? To address this question, we propose PyraNet, a lightweight, neurophysiology-informed architecture. Specifically, we devise an encoder in which learned spatial filters capture discriminative cross-channel patterns, while grouped and dilated convolutions integrate these patterns over multiple temporal receptive fields to model coupled spatiotemporal dependencies within a small parameter budget. We further introduce a dataset-level adaptive grid selector that automatically configures feature aggregation using the same penalty factors across datasets, without requiring manual granularity selection. After a shared warm-up, it selects the grid with the lowest penalized validation loss and retains it for subsequent training and inference. We evaluated PyraNet against six compact models and three advanced foundation models across eleven electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG) datasets spanning five decoding paradigms: motor imagery (MI), P300 detection, steady-state visual evoked potential (SSVEP) decoding, mental arithmetic (MA) state classification, and speech decoding. PyraNet demonstrated consistently strong performance across all eleven datasets, with the best average rank of 1.8 and a mean balanced accuracy (BAcc) of 66.93%, exceeding the runner-up by 3.1 percentage points. On the nine EEG datasets shared with the foundation models, PyraNet achieved a mean BAcc of 77.55%, outperforming the strongest foundation model by 3.1 percentage points, with approximately one-thousandth as many parameters.

1 Introduction

A brain-computer interface (BCI) establishes a direct communication pathway between the brain and external devices (Nicolas-Alonso & Gomez-Gil, 2012). Electroencephalography (EEG) non-invasively measures electrical potentials at the scalp and captures the temporal dynamics of brain activity at millisecond resolution (Michel & Murray, 2012). Magnetoencephalography (MEG) measures the magnetic fields generated by neural activity, whereas electrocorticography (ECoG) records electrical activity directly from the cortical surface (Hämäläinen et al., 1993). Since the seminal P300 speller enabled real-time character selection from event-related potentials (Farwell & Donchin, 1988), non-invasive neural decoding has expanded to motor imagery (MI) (Pfurtscheller & Neuper, 2001), steady-state visual evoked potential (SSVEP) decoding (Cheng et al., 2002), and mental arithmetic (MA) state classification (Zyma et al., 2019). More recently, deep learning-based studies have begun to retrieve perceived speech segments and decode individual words from EEG and MEG recordings (Défossez et al., 2023; d'Ascoli et al., 2025), extending non-invasive neural decoding toward speech-decoding tasks.

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.

End-to-end deep learning has shifted neural decoding from handcrafted feature extraction toward representation learning directly from raw recordings (Schirrmeister et al., 2017). Convolutional neural network (CNN) architectures, including DeepConvNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), and FBCNet (Mane et al., 2021), as well as CNN–Transformer hybrids such as EEG Conformer (Conformer) (Song et al., 2023) and EEG-Deformer (Deformer) (Ding et al., 2025), have achieved strong performance in their target tasks and paradigms. However, these models typically perform well on a specific set of tasks but degrade substantially on others.

EEG foundation models, on the other hand, use self-supervised pretraining on large and heterogeneous EEG datasets to learn representations shared across BCI paradigms (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). However, recent benchmarks show that these models often perform poorly under linear probing, vary substantially across paradigms, and do not consistently benefit from larger parameter counts (Xiong et al., 2025; Yang et al., 2026; Liu et al., 2026). These results suggest that the representations learned by current foundation models do not yet generalize consistently across BCI paradigms.

To identify the fundamental features that support decoding across BCI paradigms, we examine the characteristic neural patterns involved in each paradigm. MI is associated with event-related desynchronization/synchronization (ERD/ERS) of μ and β rhythms over sensorimotor regions (Pfurtscheller & Lopes da Silva, 1999). SSVEPs exhibit periodic responses at stimulus frequencies and their harmonics (Norcia et al., 2015). P300 is a transient positive response to attended target stimuli (Farwell & Donchin, 1988). Although these patterns differ, each is characterized by its spatial distribution, spectral content, and temporal dynamics. Spatial, spectral, and temporal structure therefore forms a fundamental yet discriminative basis for neural decoding and should be captured jointly to support generalization across BCI paradigms.

Motivated by this perspective, we propose PyraNet, a lightweight, neurophysiology-informed architecture for broadly applicable neural decoding. PyraNet first learns discriminative combinations of signals from all recording channels and then applies multi-scale grouped and dilated temporal convolutions to jointly process the learned spatial patterns over different receptive fields, capturing coupled spatiotemporal patterns and diverse frequency responses. These parallel branches span different temporal scales and form a temporal pyramid, from which PyraNet takes its name. Since each learned spatial pattern already integrates signals from all recording channels, grouping restricts each temporal filter to a subset of these patterns while retaining global spatial information. Dilation further expands the temporal receptive fields without increasing kernel size, enabling expressive coupled feature extraction within a small parameter budget.

PyraNet further addresses a key design choice in neural decoding concerning how finely to aggregate learned features. Existing architectures generally set this granularity through fixed pooling kernels or statistical windows. EEGNet employs fixed average pooling (Lawhern et al., 2018), Conformer constructs tokens using a fixed temporal pooling kernel and stride (Song et al., 2023), and FBCNet computes variance over manually specified non-overlapping windows (Mane et al., 2021). Overall, these architectures fix aggregation granularity in advance rather than adapt it to the dataset or decoding paradigm. This choice matters because our empirical results show that aggregation granularity has a substantial impact on decoding performance. Fine-grained representations better preserve rapid variations and periodic dynamics, which can be important for SSVEP decoding, but may also introduce redundant feature dimensions and increase classifier complexity. Conversely, coarse-grained representations sacrifice temporal detail while integrating information over longer intervals, which can better characterize sustained neural modulations such as MI-related ERD/ERS. To balance this trade-off, PyraNet introduces a dataset-level adaptive grid selector that compares candidate granularities after a shared warm-up and applies a penalty to finer grids, selecting a fixed grid for subsequent training and inference.

Our main contributions are summarized as follows:

- We propose a lightweight model that effectively captures spatiotemporal patterns. After learning discriminative combinations of signals from all recording channels, it uses a pyramidal structure of grouped and dilated temporal convolutions to jointly process subsets of spatial patterns over different receptive fields. The resulting model uses only about one-thousandth as many parameters as the leading foundation model while delivering superior overall performance.

- We identify feature granularity as a critical, task-dependent factor in neural decoding. Based on this observation, we introduce a dataset-level adaptive grid selector that automatically selects an appropriate granularity after a shared warm-up stage, without requiring manual granularity selection.
- We conducted extensive experiments on eleven EEG, MEG, and ECoG datasets spanning MI, P300, SSVEP, speech decoding, and MA state classification. Comparisons with six compact models and three pretrained EEG foundation models demonstrated the broad applicability of PyraNet across paradigms and recording modalities.

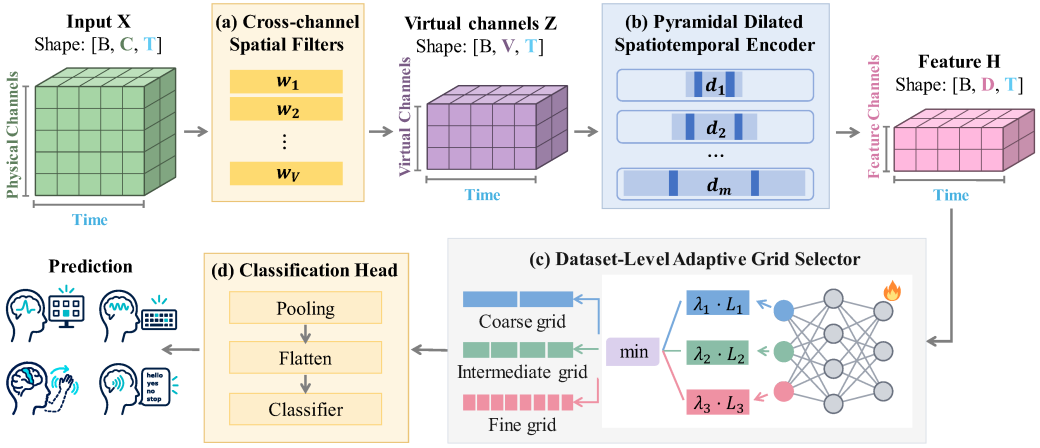

Figure 1: Overview of PyraNet. Raw signals pass through cross-channel spatial filtering, pyramidal dilated encoding, and adaptive grid selection to extract discriminative spatiotemporal representations at an appropriate feature granularity, followed by a simple linear classification head for prediction.

2 Method

We consider supervised neural decoding using a dataset $\mathcal{D} = \{(\mathbf{X}_i, y_i)\}$ of labeled multichannel recordings. Each $\mathbf{X}_i \in \mathbb{R}^{C\times T}$ contains C channels and T time samples at sampling frequency f_s, and $y_i \in \{1, \ldots, Y\}$ denotes one of Y classes. PyraNet has three modules: learned cross-channel filtering, pyramidal dilated spatiotemporal encoding, and dataset-level adaptive selection of temporal feature granularity. The first two modules produce

$$\mathbf{Z}_i = \mathcal{S}_{\theta_s}(\mathbf{X}_i) \in \mathbb{R}^{V\times T}, \qquad \mathbf{H}_i = \mathcal{E}_{\theta_e}(\mathbf{Z}_i) \in \mathbb{R}^{D\times T}, \tag{1}$$

where V and D denote the numbers of virtual channels and encoded features, respectively, and θ_s, θ_e are learnable parameters. The selector then chooses how finely to pool $\mathbf{H}_i$ before classification. Figure 1 summarizes the pipeline, and Appendix B lists the hyperparameters.

2.1 Learned Cross-Channel Filtering

Neural recordings reflect spatially mixed neural activity rather than isolated local sources. A single scalp EEG electrode therefore measures a macroscopic, smoothed version of the local field potential (Buzsáki et al., 2012), and subdural ECoG electrodes capture the same mixed fields. MEG sensors instead measure the weak magnetic fields produced by the underlying neuronal currents, which pass through the skull and scalp with little distortion, yet each sensor still observes a superposition of many distributed sources (Hämäläinen et al., 1993). To extract discriminative spatial patterns from these mixed observations, PyraNet learns a set of spatial filters, each of which forms a weighted combination of signals from all recording channels at each time point. These filters map the original recording channels into virtual channels that emphasize discriminative neural activity, providing spatial representations for subsequent temporal modeling. We apply batch normalization (BN) after spatial filtering:

$$\mathbf{Z}_{i,:,t} = \mathrm{BN}_s(\mathbf{W}_s\mathbf{X}_{i,:,t}), \qquad \mathbf{W}_s \in \mathbb{R}^{V\times C}. \tag{2}$$

The operation combines information across all channels at each time point, without mixing information across time or changing the temporal length. Each resulting virtual channel represents a task-driven mixture of signals from all recording channels rather than a signal from a single recording channel or a predefined scalp region.

2.2 Pyramidal Dilated Spatiotemporal Encoder

EEG encoders such as EEGNet use depthwise temporal convolution within separable convolution blocks to summarize each spatial feature independently before pointwise feature mixing (Lawhern et al., 2018). This temporal operation is efficient, whereas dense temporal convolution jointly models all spatial features over time at a higher parameter cost. PyraNet takes an intermediate route. It uses M parallel temporal branches, each implemented with G groups and producing D/M output features. A filter therefore combines V/G virtual channels over time. Since those virtual channels already combine signals from all recording channels, grouping reduces redundant connections without returning to channel-wise processing, so each filter can still model interactions among several global spatial patterns.

The branches share this grouped structure but use different temporal dilations. We use the Gaussian error linear unit (GELU) as the activation:

$$\mathbf{U}_i^{(m)} = \mathrm{BN}_m\big(\mathrm{Conv}_{k,d_m}^{G}(\mathbf{Z}_i)\big) \in \mathbb{R}^{(D/M)\times T}, \qquad \mathbf{H}_i = \mathrm{GELU}\Big(\mathop{\Big\|}_{m=1}^{M} \mathbf{U}_i^{(m)}\Big). \tag{3}$$

Here, $\mathrm{Conv}_{k,d_m}^{G}$ denotes a temporal convolution with kernel length k, dilation rate d_m, and G groups. Zero padding preserves the temporal length T, and $\|$ denotes concatenation along the feature dimension. The dilation rates are chosen so that the effective receptive fields approximate target temporal windows, whose lengths in samples are determined by the sampling rate. With a fixed kernel length, larger dilation rates cover longer intervals, while smaller rates capture shorter intervals. The branches are ordered from long to short temporal windows, forming a temporal pyramid that captures neural responses at multiple time scales. This design achieves broad temporal coverage with a lightweight parameterization. Grouping and dilation further reduce the parameter cost of multi-scale feature extraction.

2.3 Dataset-Level Adaptive Grid Selector

Existing networks typically aggregate the encoder output features with a fixed average pooling window along the time axis. This approach achieves strong performance on certain paradigms but degrades severely on others, and cannot adapt to a broad range of tasks across paradigms (Lawhern et al., 2018; Song et al., 2023; Mane et al., 2021). Instead, we consider an ordered set of candidate grids $\Gamma = \{g_1, \ldots, g_Q\}$ for feature aggregation, where Q is the number of candidates and g_q is the number of temporal bins of the q-th grid, with $g_1 < \cdots < g_Q$. In our experiments, we use $g_q \in \{8, 32, 128\}$. Feature aggregation controls the temporal detail retained for classification. Fine grids preserve more detail but increase the size of the classification head, while coarse grids reduce its size but may discard useful information. We therefore select one grid to balance these effects.

Each candidate applies adaptive average pooling to the encoded features $\mathbf{H}_i$ and uses an independent linear classification head:

$$\mathbf{A}_{i,g_q} = \mathrm{Pool}_{g_q}(\mathbf{H}_i) \in \mathbb{R}^{D\times g_q}, \qquad \mathbf{z}_{i,g_q} = h_{g_q}\big(\mathrm{Flatten}(\mathrm{Dropout}(\mathbf{A}_{i,g_q}))\big) \in \mathbb{R}^{Y}. \tag{4}$$

Here, $\mathbf{A}_{i,g_q}$ is the pooled feature sequence, $\mathrm{Flatten}$ converts it into a vector of length Dg_q, and h_{g_q} maps this vector to logits $\mathbf{z}_{i,g_q}$ for the Y classes. During the first e_w epochs, the shared encoder and all candidate heads are optimized jointly with equal loss weighting, giving every candidate supervised training before selection.

After the final warm-up update, we compute each candidate's mean per-trial classification loss on the validation set and multiply it by a penalty coefficient that assigns larger penalties to finer grids. A finer grid must therefore provide a sufficiently large loss reduction to offset its larger penalty. The

selected grid is determined by:

$$\overline{\mathcal{L}}_{\mathrm{val},g_q} = \frac{1}{N_{\mathrm{val}}} \sum_{i \in \mathcal{D}_{\mathrm{val}}} \ell(\mathbf{z}_{i,g_q}, y_i),$$
$$g^* = \arg\min_{g_q \in \Gamma} \lambda_{g_q} \overline{\mathcal{L}}_{\mathrm{val},g_q}. \tag{5}$$

Here, $\mathcal{D}_{\mathrm{val}}$ is the validation set, N_{val} is its number of trials, ℓ is the classification loss, and y_i is the label of trial i. The coefficient λ_{g_q} penalizes finer grids, and g^* is the selected grid. The relevant parameters are listed in Appendix B.

Once g^* is selected, the unused heads are removed, and the shared encoder and retained head continue training. All subsequent training and inference use this fixed grid. The selector has no trainable parameters, and test data are not used for selection. The candidate grids and penalty coefficients used in all experiments are listed in Appendix B.

3 Experiments

3.1 Downstream Tasks and Datasets

To evaluate generalizability across diverse BCI paradigms, PyraNet was assessed on eleven datasets spanning five decoding tasks and three recording modalities. Details of the datasets are reported in Table 1 and Appendix C.1.

Table 1: Summary of the eleven datasets.

Task	Dataset	Modality	Native Hz	Channels	Subjects	Sessions	Prediction targets (classes)
MI	BNCI2014_001	EEG	250	22	9	2	Left/right hand, feet, tongue (4)
MI	BNCI2014_002	EEG	512	15	14	1	Right hand, feet (2)
MI	BNCI2014_004	EEG	250	3	9	5	Left/right hand (2)
MI	BNCI2015_001	EEG	512	13	12	2–3	Right hand, feet (2)
P300	BNCI2014_008	EEG	256	8	8	1	Target/non-target (2)
P300	BNCI2014_009	EEG	256	16	10	3	Target/non-target (2)
MA	EEGMAT	EEG	500	19	36	1	Rest / MA (2)
SSVEP	Nakanishi2015	EEG	256	8	9	1	Stimulus frequency (12)
SSVEP	Wang2016	EEG	250	62	34	1	Stimulus frequency (40)
Speech	text-MEG	MEG	1,000	57–61	9	15	Mandarin words (48)
Speech	bouchard2019	ECoG	3,051.76	229–249	4	7/7/14/3	Consonant–vowel syllables (54)

These tasks test whether the same architecture can accommodate different neural dynamics and recording modalities. For MI, participants imagine cued movements, and the task is to classify the imagined movement from EEG. For P300, participants attend to a character in a visual speller, and the task is to distinguish target from non-target EEG responses. For SSVEP, participants fixate on a flickering stimulus, and the task is to identify the attended target from EEG. We obtained the MI, P300, and SSVEP datasets through MOABB (Jayaram & Barachant, 2018).

For MA, the task is to distinguish resting EEG from EEG recorded during mental subtraction, using the publicly available EEGMAT dataset (Zyma et al., 2019). For speech decoding, we used the publicly released text-MEG (Jia et al., 2025) and bouchard2019 (Bouchard & Chang, 2020) datasets. Text-MEG supports word classification from MEG recorded during visually prompted Mandarin word reading. Bouchard2019 supports syllable classification from ECoG recorded during consonant–vowel syllable production.

3.2 Baselines

We compared PyraNet with six compact models: EEGNet (Lawhern et al., 2018), FBCNet (Mane et al., 2021), IFNet (Wang et al., 2023), Conformer (Song et al., 2023), DBConformer (Wang et al., 2026), and Deformer (Ding et al., 2025), as well as three pretrained EEG foundation models: LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), and REVE (El Ouahidi et al., 2025). The compact models were adapted from publicly released code, whereas the foundation models were

evaluated using their public code and publicly released pretrained weights. Since all three foundation models were pretrained on EEG, they were evaluated only on the nine EEG datasets, excluding the MEG and ECoG speech datasets. Model configurations are provided in Appendix C.3.

3.3 EVALUATION

We used leave-one-subject-out (LOSO) evaluation for the nine EEG datasets, training on the remaining subjects and testing on the held-out subject. For the two speech datasets, we used leave-one-session-out evaluation within each subject.

We applied Euclidean alignment (EA), an established preprocessing approach that aligns EEG trials from different subjects in Euclidean space to make them more similar and thereby improve learning performance for a new subject (He & Wu, 2020). EA computes the covariance matrix of each trial, averages these matrices to obtain a reference covariance, and multiplies the signals by its inverse square root. For the nine EEG datasets, EA was applied separately for each subject. For the two speech datasets, whose evaluation is within subjects across sessions, EA was applied separately for each session.

Within each fold, the non-test data were randomly split into fitting and validation sets. The validation fraction is ρ_{val}, with its value and other training settings listed in Appendix C.6. For early stopping, we monitored the area under the receiver operating characteristic curve (AUROC) on the validation set for binary tasks and Cohen's κ for multiclass tasks. Training stopped if this score did not improve for a preset number of epochs. For PyraNet, early stopping began after the warm-up stage. We then restored the model with the highest validation score and evaluated it on the test set. Test labels were used only for evaluation.

For binary tasks, we report BAcc, the area under the precision–recall curve (AUC–PR), and AUROC. For multiclass tasks, we report BAcc, Cohen's κ, and weighted F1. All experiments were repeated with three random seeds. All metrics except κ are expressed as percentages.

Table 2: Mean BAcc (%) on eleven datasets across five paradigms. Standard deviations are reported in Appendix D. Best results are bold, and second-best results are underlined. Equal means are ranked by lower standard deviation.

Paradigm	Dataset	EEGNet	FBCNet	IFNet	Conformer	DBConformer	Deformer	CBraMod	LaBraM	REVE	PyraNet
MI	BNCI2014_001	51.54	47.23	<u>52.49</u>	50.01	49.25	42.53	47.37	43.51	43.99	**53.57**
	BNCI2014_002	74.90	71.88	76.16	<u>76.70</u>	75.34	64.85	69.35	65.52	62.57	**79.46**
	BNCI2014_004	74.24	70.57	**74.47**	72.98	<u>74.37</u>	72.71	73.07	72.45	70.19	74.24
	BNCI2015_001	74.61	72.22	**75.37**	73.95	73.56	71.07	69.50	69.06	69.68	<u>75.03</u>
P300	BNCI2014_008	71.59	53.49	66.17	70.88	<u>71.76</u>	71.52	70.91	71.19	71.40	**71.76**
	BNCI2014_009	<u>81.52</u>	61.12	76.77	78.88	80.52	81.42	79.74	79.53	80.41	**81.52**
SSVEP	Nakanishi2015	92.88	38.17	76.11	81.75	30.31	<u>95.70</u>	95.27	92.18	90.02	**97.10**
	Wang2016	90.56	26.19	66.92	90.42	31.47	**94.15**	92.80	91.92	60.36	<u>94.06</u>
MA	EEGMAT	64.20	71.08	**75.19**	58.21	73.83	61.27	71.94	68.98	<u>73.83</u>	71.20
Speech	text-MEG	6.87	5.30	<u>11.17</u>	7.17	5.92	9.45	-	-	-	**11.44**
	bouchard2019	19.09	2.01	13.63	19.40	14.48	<u>23.60</u>	-	-	-	**26.86**
Overall Average ↑		<u>63.82</u>	47.21	60.40	61.85	52.80	62.57	-*	-*	-*	**66.93**
Overall Rank ↓		<u>4.09</u>	7.91	4.45	5.64	5.00	5.18	6.18*	7.45*	7.27*	**1.82**

*Rank is averaged over all eleven datasets. Missing foundation model results on speech datasets are assigned rank 9. The overall average is calculated only for models with results on all eleven datasets.

3.4 RESULTS

Table 2 shows that PyraNet achieves strong performance across paradigms, with the highest mean BAcc (66.93%) and best mean rank (1.8) among the models evaluated on all eleven datasets. Its strong results extend from MI and evoked-response decoding with EEG to speech decoding with MEG and ECoG. IFNet remains particularly competitive on MI, while its mean BAcc on P300 is 5.17 percentage points below PyraNet. Deformer performs well on P300, SSVEP, and speech

Table 3: Model parameters and single-trial FLOPs on BNCI2014_001. PyraNet's three slash-separated values in Head Parameters and Total Parameters correspond to grids 8, 32, and 128.

Method	Backbone parameters	Head parameters	Total parameters	FLOPs (M)
EEGNet	1,944	1,988	3,932	45
FBCNet	7,200	4,612	11,812	13
IFNet	9,344	2,052	11,396	29
Conformer	195,760	633,252	829,012	137
DBConformer	95,265	132	95,397	18
Deformer	1,030,480	8,964	1,039,444	523
CBraMod	4,883,800	70,404	4,954,204	691
LaBraM	5,819,936	71,204	5,891,140	1104
REVE	70,270,976	228,356	70,499,332	12,713
PyraNet	4,160	1,028/4,100/16,388	5,188/8,260/20,548	8/8/8

decoding, but its mean BAcc on MI is 7.79 percentage points below PyraNet. These comparisons support the broad applicability of PyraNet across paradigms and recording modalities, with separate training for each dataset.

Across the nine EEG datasets shared with the foundation models, PyraNet outperforms all three foundation models on eight datasets. On the BNCI2014_001 dataset, it uses approximately one-thousandth as many parameters as the strongest foundation model and has the lowest FLOPs among all evaluated models (Table 3). These results highlight a favorable balance between decoding performance and computational efficiency. Full evaluation metrics are reported in Appendix D, and further details on parameters and FLOPs are provided in Appendix E.

3.5 Visualization Analysis

Learned cross-channel filtering. Figure 2 shows that the learned spatial filters emphasize different recording channels for MI and SSVEP. On BNCI2014_001, C3 and C4 receive the largest relative weights. These electrodes lie over sensorimotor regions, which show prominent activity changes during MI (Pfurtscheller & Neuper, 2001). On Nakanishi2015, Oz receives the largest weight. This electrode lies over the occipital visual cortex, which generates strong responses to periodic visual stimulation (Norcia et al., 2015). These patterns suggest that cross-channel filtering adapts to task-relevant spatial structure. Analysis details and results for the other datasets are provided in Appendix F.

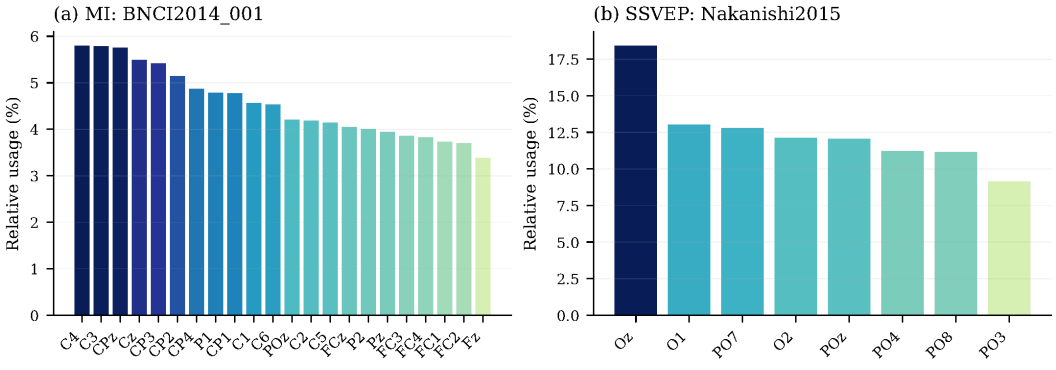

Figure 2: Relative channel usage on (a) BNCI2014_001 and (b) Nakanishi2015, computed from the learned spatial filter weights.

Pyramidal dilated spatiotemporal encoder. Figure 3 compares the activation spectra of the four temporal branches on MI and SSVEP. The branches retain similar broad spectral profiles but differ in their relative emphasis across frequencies. This suggests that the temporal scales provide partially

complementary features while sharing information about the same neural responses. The substantial overlap also shows that the branches do not act as separate, non-overlapping frequency bands. Spectral analysis settings and the remaining datasets are provided in Appendix F.

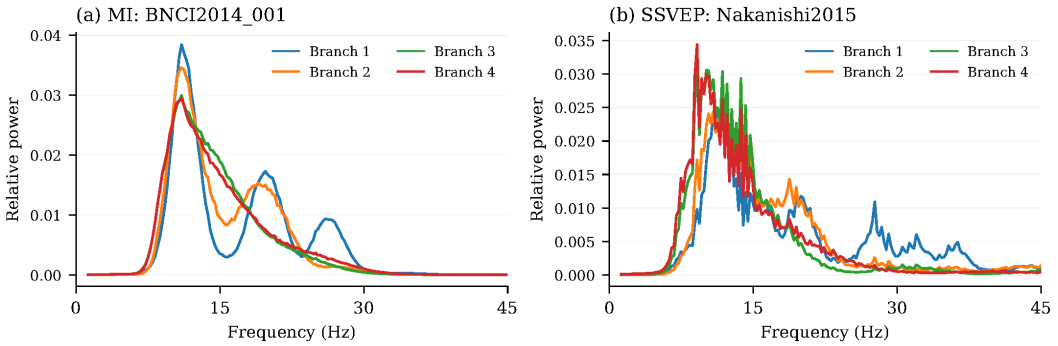

Figure 3: Relative activation power spectra of the four temporal branches on (a) BNCI2014_001 and (b) Nakanishi2015.

Dataset-level adaptive grid selector. As shown in Figure 4, the fine-grid representation on Nakanishi2015 shows clearer class separation than the coarse-grid representation of the same held-out trials. This qualitative observation is consistent with the fixed-grid ablation in Section 3.6 and helps explain why retaining finer temporal detail benefits SSVEP decoding.

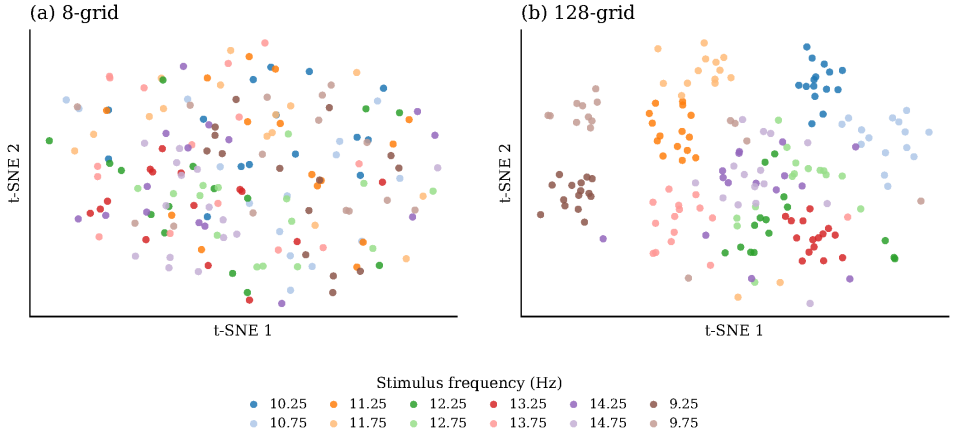

Figure 4: t-SNE visualizations of Nakanishi2015 representations using (a) grid 8 and (b) grid 128. Both panels use the same encoder and held-out trials. Colors indicate stimulus frequencies.

3.6 Ablation and Component Analysis

We assess the pyramidal dilated spatiotemporal encoder and the dataset-level adaptive grid selector. Figure 5a shows that dilation reduces temporal kernel parameters by 82.6%–91.7% while maintaining comparable performance. Figure 5b shows the expected selection behavior: PyraNet favors a coarse grid to limit classifier size when finer grids provide only marginal gains, but selects a fine grid despite the parameter penalty when the performance improvement is substantial. To further evaluate both components, we select IFNet, the strongest baseline among those that can be adapted to incorporate the grid selector. Table 4 compares IFNet with and without the selector to evaluate its effectiveness, and compares IFNet with the grid selector against PyraNet to assess the contribution of the pyramidal dilated spatiotemporal encoder.

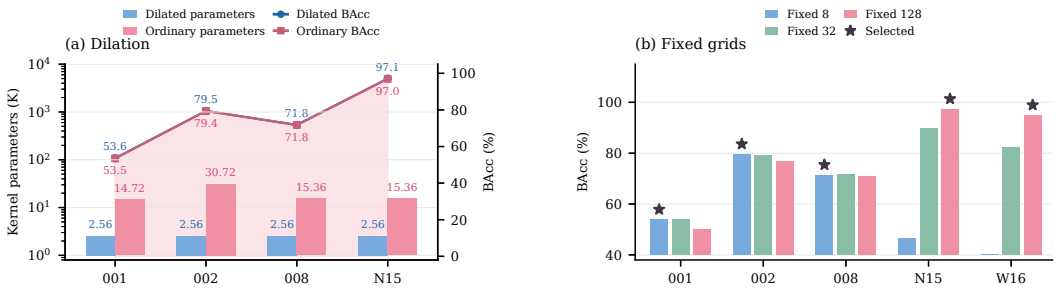

Figure 5: Encoder and grid ablations. (a) Dilated versus ordinary convolution. Bars show temporal kernel parameters and lines show BAcc. (b) Fixed grids. Stars mark grids selected by PyraNet.

Table 4: Comparison of IFNet, IFNet + GS, and PyraNet across eight datasets. GS denotes the grid selector.

Method	BAcc	BNCI2014_001 κ	Weighted F1	BAcc	BNCI2014_002 AUC–PR	AUROC
IFNet	52.49 ± 0.18	0.3665 ± 0.0024	51.27 ± 0.26	76.16 ± 0.90	84.53 ± 0.09	85.14 ± 0.17
IFNet + GS	52.91 ± 0.66	0.3721 ± 0.0088	51.55 ± 0.88	77.17 ± 0.50	84.72 ± 0.18	85.49 ± 0.21
PyraNet	**53.57** ± 0.26	**0.3809** ± 0.0035	**52.81** ± 0.38	**79.46** ± 0.38	**84.95** ± 0.14	**85.68** ± 0.05
Method	BAcc	BNCI2014_008 AUC–PR	AUROC	BAcc	BNCI2014_009 AUC–PR	AUROC
IFNet	66.17 ± 0.35	35.68 ± 0.23	72.56 ± 0.24	76.77 ± 0.23	59.50 ± 0.18	85.68 ± 0.06
IFNet + GS	69.83 ± 0.17	41.95 ± 0.11	76.71 ± 0.15	79.89 ± 0.12	67.49 ± 0.32	88.56 ± 0.09
PyraNet	**71.76** ± 0.12	**44.85** ± 0.14	**78.85** ± 0.12	**81.52** ± 0.14	**70.61** ± 0.39	**89.59** ± 0.09
Method	BAcc	Nakanishi2015 κ	Weighted F1	BAcc	Wang2016 κ	Weighted F1
IFNet	76.11 ± 1.98	0.7394 ± 0.0216	75.62 ± 2.17	66.92 ± 0.33	0.6607 ± 0.0034	65.58 ± 0.37
IFNet + GS	**97.10** ± 0.05	**0.9684** ± 0.0005	97.03 ± 0.08	**94.71** ± 0.17	**0.9458** ± 0.0018	**94.58** ± 0.20
PyraNet	**97.10** ± 0.05	**0.9684** ± 0.0005	**97.04** ± 0.05	94.06 ± 0.08	0.9391 ± 0.0008	93.87 ± 0.06
Method	BAcc	text-MEG κ	Weighted F1	BAcc	bouchard2019 κ	Weighted F1
IFNet	11.17 ± 0.38	0.0928 ± 0.0039	7.23 ± 0.34	13.63 ± 0.02	0.1197 ± 0.0006	12.54 ± 0.16
IFNet + GS	11.01 ± 0.30	0.0912 ± 0.0030	7.46 ± 0.15	8.12 ± 0.31	0.0638 ± 0.0028	7.39 ± 0.25
PyraNet	**11.44** ± 0.06	**0.0956** ± 0.0006	**7.74** ± 0.20	**26.86** ± 0.47	**0.2533** ± 0.0036	**25.43** ± 0.23
Mean BAcc	IFNet 54.93		IFNet + GS 61.34		PyraNet **64.47**	

As shown in Table 4, the grid selector improves IFNet's mean BAcc by 0.71 percentage points on MI, 3.39 on P300, and 24.39 on SSVEP. Across eight datasets, adding the grid selector raises IFNet's mean BAcc by 6.41 percentage points, supporting its effectiveness beyond PyraNet. PyraNet outperforms IFNet with the grid selector on the two speech datasets by an average of 9.59 percentage points. These results suggest that the pyramidal dilated spatiotemporal encoder contributes to broader adaptation across paradigms and recording modalities.

4 Conclusion

In this paper, we propose PyraNet, a lightweight model for broadly applicable neural signal decoding. Specifically, we use learned cross-channel filtering to extract spatial patterns, followed by a pyramidal dilated spatiotemporal encoder that combines grouped and dilated temporal convolutions to model spatiotemporal dependencies over multiple receptive fields with few parameters. We also introduce a dataset-level adaptive grid selector that balances classification loss against a fine-granularity penalty and retains one fixed grid for subsequent training and inference. Across eleven EEG, MEG, and ECoG datasets, PyraNet delivered consistently strong performance, with the best average rank of 1.8 and a mean BAcc of 66.93%, exceeding the runner-up by 3.1 percentage points.

AI USE STATEMENT

We used AI to assist with translation, language polishing, literature searches, the development of a general-purpose experimental framework, and monitoring the progress of long-running experiments for anomalies. The authors have manually reviewed all AI-assisted code and text and verified the retrieved literature. Retrieved sources were checked individually online. Cited publications were downloaded and inspected to verify that the relevant passages supported the corresponding statements. All translated and edited text was reviewed and further revised manually. The authors take full responsibility for the final content of this work, including all AI-assisted text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The model and evaluation procedures are described in the main text, with additional implementation details in Appendix C.6. We provide an anonymized code supplement with this submission, including model implementations, preprocessing code, training and evaluation scripts, configurations and random seeds, environment dependencies, and scripts for the complete experimental pipeline. These materials are intended to support reproduction of the reported experiments.

REFERENCES

Pietro Aricò, Fabio Aloise, Francesca Schettini, Serenella Salinari, Donatella Mattia, and Febo Cincotti. Influence of P300 latency jitter on event related potential-based brain-computer interface performance. *Journal of Neural Engineering*, 11(3):035008, 2014. doi: 10.1088/1741-2560/11/3/035008.

Kristofer E. Bouchard and Edward F. Chang. Human ECoG speaking consonant-vowel syllables. Figshare collection, version 4, 2020. URL `https://figshare.com/collections/Human_ECoG_speaking_consonant-vowel_syllables/4617263`.

Kristofer E. Bouchard, Nima Mesgarani, Keith Johnson, and Edward F. Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495:327–332, 2013. doi: 10.1038/nature11911.

Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010. doi: 10.1109/ICPR.2010.764.

György Buzsáki, Costas A. Anastassiou, and Christof Koch. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13:407–420, 2012. doi: 10.1038/nrn3241.

Gian Emilio Chatrian, Edward Lettich, and Paul L. Nelson. Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG Technology*, 25(2):83–92, 1985. doi: 10.1080/00029238.1985.11080163.

Ming Cheng, Xiaorong Gao, Shangkai Gao, and Dingfeng Xu. Design and implementation of a brain–computer interface with high transfer rates. *IEEE Transactions on Biomedical Engineering*, 49(10):1181–1186, 2002. doi: 10.1109/TBME.2002.803536.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16:10521, 2025. doi: 10.1038/s41467-025-65499-0.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5 (10):1097–1107, 2023. doi: 10.1038/s42256-023-00714-5.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. EEG-Deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2025. doi: 10.1109/JBHI.2024.3504604.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG - adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 22541–22577, 2025. doi: 10.52202/085713-0760. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/20a917f77773ac0fa8bea2bdd6606b66-Paper-Conference.pdf.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012. doi: 10.1109/TNSRE.2012.2189584.

Lawrence A. Farwell and Emanuel Donchin. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988. doi: 10.1016/0013-4694(88)90149-6.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Matti Hämäläinen, Riitta Hari, Risto J. Ilmoniemi, Jukka Knuutila, and Olli V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497, 1993. doi: 10.1103/RevModPhys.65.413.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Vinay Jayaram and Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Zhihong Jia, Hongbin Wang, Yuanzhong Shen, Feng Hu, Jiayu An, Kai Shu, and Dongrui Wu. Magnetoencephalography (MEG) based non-invasive Chinese speech decoding. *Journal of Neural Engineering*, 22(6):066014, 2025. doi: 10.1088/1741-2552/ae1ea2.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

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*, 2026. doi: 10.1093/nsr/nwag466.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A. P. Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain-computer interface. *arXiv preprint arXiv:2104.01233*, 2021. URL https://arxiv.org/abs/2104.01233.

Christoph M. Michel and Micah M. Murray. Towards the utilization of EEG as a brain imaging tool. *NeuroImage*, 61(2):371–385, 2012. doi: 10.1016/j.neuroimage.2011.12.039.

Masaki Nakanishi, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PLOS ONE*, 10(10):e0140703, 2015. doi: 10.1371/journal.pone.0140703.

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

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

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. doi: 10.1016/S1388-2457(99)00141-8.

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

John Polich. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118 (10):2128–2148, 2007. doi: 10.1016/j.clinph.2007.04.019.

Angela Riccio, Luca Simione, Francesca Schettini, Antonella Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, and Donatella Mattia. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7:732, 2013. doi: 10.3389/fnhum.2013.00732.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas D. J. 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. doi: 10.1002/hbm.23730.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

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.

David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R. Müller-Putz. Random forests in non-invasive sensorimotor rhythm brain-computer interfaces: A practical and convenient non-linear classifier. *Biomedical Engineering / Biomedizinische Technik*, 61(1):77–86, 2016. doi: 10.1515/bmt-2014-0117.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. doi: 10.3389/fnins.2012.00055.

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

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Jiaheng Wang, Lin Yao, and Yueming Wang. IFNet: An interactive frequency convolutional neural network for enhancing motor imagery decoding from EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1900–1911, 2023. doi: 10.1109/TNSRE.2023.3257319.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017. doi: 10.1109/TNSRE.2016.2627556.

Ziwei Wang, Hongbin Wang, Tianwang Jia, Xingyi He, Siyang Li, and Dongrui Wu. DBConformer: Dual-branch convolutional transformer for EEG decoding. *IEEE Journal of Biomedical and Health Informatics*, 30(5):4134–4147, 2026. doi: 10.1109/JBHI.2025.3622725.

Wei Xiong, Jiangtong Li, Jie Li, and Kun Zhu. EEG-FM-Bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of EEG foundation models, 2025. URL https://arxiv.org/abs/2508.17742.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? Comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019. doi: 10.3390/data4010014.

A Related Work

End-to-End Neural Decoders End-to-end neural networks learn task-relevant representations directly from neural recordings. Schirrmeister et al. introduced two influential convolutional baselines: DeepConvNet uses stacked convolution–pooling blocks to learn hierarchical features, whereas ShallowConvNet employs squaring and logarithmic pooling to approximate band-power extraction (Schirrmeister et al., 2017). EEGNet substantially reduces model size by combining temporal convolution, depthwise spatial filtering, and separable convolution, and has demonstrated applicability across both event-related and oscillatory BCI paradigms (Lawhern et al., 2018).

Subsequent compact models incorporated stronger neurophysiological priors. FBCNet applies a predefined filter bank, learns depthwise spatial filters within each frequency band, and summarizes temporal activity using windowed variance (Mane et al., 2021). IFNet constructs low and high frequency representations and explicitly models their interaction before temporal aggregation (Wang et al., 2023). Both architectures achieve strong MI decoding performance, although their frequency decomposition and aggregation strategies are primarily designed for this paradigm. CNN–Transformer hybrids extend convolutional feature extraction with attention-based sequence modeling. Conformer combines local convolutional features with self-attention for global temporal dependencies (Song et al., 2023), while Deformer introduces hierarchical coarse-to-fine temporal modeling (Ding et al., 2025).

These models demonstrate the effectiveness of compact architectural priors, but their efficiency is frequently obtained from decoupled or depthwise processing, and most designs are tailored to one or a limited number of BCI paradigms, which limits their generalization across heterogeneous decoding settings. PyraNet follows this lightweight direction while adopting a different coupling strategy that jointly extracts spatial and temporal features in a parameter-efficient manner. Moreover, its adaptive design allows a single architecture to accommodate the distinct temporal and spectral characteristics of MI, P300, SSVEP, MA, and speech decoding, yielding stronger generalization across BCI paradigms.

EEG Foundation Models EEG foundation models seek to alleviate limited labeled data and substantial cross-dataset variability by learning reusable representations from large collections of unlabeled recordings. BIOT represents recordings with different channel sets, durations, and missing values as unified sequences of channel-wise spectral tokens (Yang et al., 2023). LaBraM maps EEG patches into discrete neural codes and performs masked code prediction on approximately 2,500 hours of EEG collected from about 20 datasets (Jiang et al., 2024). EEGPT combines masked modeling with spatiotemporal representation alignment (Wang et al., 2024), whereas CBraMod reconstructs masked EEG patches using parallel spatial and temporal attention together with conditional positional encoding (Wang et al., 2025). REVE uses four-dimensional positional encoding and masked reconstruction to process EEG with varying channel layouts and temporal lengths (El Ouahidi et al., 2025). Despite their architectural differences, these models share the premise that heterogeneous self-supervised pretraining can produce representations transferable across subjects, datasets, and tasks.

Recent standardized evaluations, however, provide mixed evidence for universal transfer. Foundation models can perform well after full fine-tuning in data-rich population-level settings, but often do not significantly outperform compact decoders in data-scarce or subject-specific settings. Moreover, increasing model size does not produce a consistent improvement in downstream performance (Yang et al., 2026; Liu et al., 2026). Frozen-backbone linear probing is also frequently weak, indicating that current pretrained representations are not consistently task-ready or linearly separable (Xiong et al., 2025; Yang et al., 2026; Liu et al., 2026). This does not imply that pretraining is ineffective, as pretrained weights may still provide useful initialization. Nevertheless, performance can change substantially with the feature-aggregation strategy and downstream classification head, making it difficult to attribute full-fine-tuning gains to the pretrained backbone alone (Xiong et al., 2025).

Data heterogeneity likewise does not guarantee uniformly positive transfer. Unified multi-task experiments show non-monotonic outcomes. Joint training over multiple datasets benefits some tasks but can degrade others (Xiong et al., 2025). In addition, some downstream datasets originate from corpora represented during pretraining, even when the underlying recordings are kept disjoint (Jiang

et al., 2024). This is not label leakage, but it differs from transfer to a completely unseen corpus. Transparent reporting of pretraining composition, recording-level overlap, preprocessing, downstream heads, and adaptation protocols is therefore essential for assessing genuine cross-dataset generalization.

B PyraNet Hyperparameters

Table 5 reports hyperparameters of PyraNet. The temporal kernel receptive fields are set according to the sampling frequency of the input signal, targeting approximately $f_s/4$, $f_s/8$, $f_s/16$, and $f_s/32$, corresponding to 0.25, 0.125, 0.0625, and 0.03125 s, respectively. Taking 250 Hz as an example, the four dilated kernels span 65, 33, 17, and 9 samples.

Table 5: Hyperparameters of PyraNet.

Hyperparameter	Symbol	Value
Virtual channels	V	64
Output features	D	32
Temporal branches	M	4
Convolution groups	G	4
Temporal kernel length	k	5
Multi-scale receptive fields (250 Hz)	—	65, 33, 17, 9
Dropout	—	0.5
Candidate grids	Γ	$\{8, 32, 128\}$
Number of candidates	Q	3
Fine-granularity penalty factors	$(\lambda_{g_1}, \lambda_{g_2}, \lambda_{g_3})$	$(1, 2, 8)$
Warm-up epochs	e_w	10

C Downstream Experimental Settings

C.1 Downstream Datasets and Preprocessing

MI datasets. In MI experiments, participants imagine a movement specified by a visual cue without executing it. We used four datasets obtained from the MOABB framework (Jayaram & Barachant, 2018). BNCI2014_001 and BNCI2014_004 are BCI Competition IV datasets 2a and 2b, respectively (Tangermann et al., 2012). BNCI2014_001 covers left hand, right hand, both feet, and tongue imagery. BNCI2014_004 covers left hand and right hand imagery. BNCI2014_002 (Steyrl et al., 2016) and BNCI2015_001 (Faller et al., 2012) cover right hand and both feet imagery.

P300 datasets. We evaluated P300 detection using two P300 datasets, BNCI2014_008 and BNCI2014_009, obtained from MOABB (Jayaram & Barachant, 2018). In these visual speller experiments, participants attend to a selected character while items on the display are flashed. Responses to flashes associated with the attended character are labeled as targets, and the remaining responses as non-targets, forming a binary classification task. BNCI2014_008 contains eight EEG channels from eight participants with amyotrophic lateral sclerosis (Riccio et al., 2013), whereas BNCI2014_009 contains 16 EEG channels from ten healthy participants (Aricò et al., 2014).

SSVEP datasets. We used Nakanishi2015 (Nakanishi et al., 2015) and Wang2016 (Wang et al., 2017), obtained from MOABB. In Nakanishi2015, participants fixate on a cued target among 12 visual stimuli flickering with distinct frequency–phase combinations. Our experiments used the nine subjects available through the MOABB loader, each recorded with eight EEG channels. Wang2016 contains recordings from 34 participants using 62 retained EEG channels and 40 visual targets encoded by distinct frequency–phase combinations. For both datasets, the task is to identify the attended target from its evoked EEG response.

Speech decoding datasets. We used text-MEG (Jia et al., 2025) and bouchard2019 (Bouchard & Chang, 2020) to evaluate speech decoding from MEG and ECoG, respectively. In text-MEG, nine participants read 48 Mandarin words displayed on a screen. Each trial consists of a preparation

period, word presentation and reading, and rest. Each participant completes 15 blocks containing all 48 words, and the task is 48-class word classification using 57–61 retained MEG channels. In bouchard2019, participants produce consonant–vowel syllables, with annotated syllable identities and speech onsets. This dataset contains recordings from four participants, with 229–249 retained ECoG channels, for 54-class syllable classification.

MA dataset. We used EEGMAT (Zyma et al., 2019) to distinguish resting activity from activity during MA. Participants first rest with their eyes closed and then repeatedly subtract a specified number from a larger number mentally. The dataset contains 19 EEG channels from 36 participants. We divided the recordings into consecutive analysis windows and classified each window as rest or arithmetic, forming a binary state-classification task.

Preprocessing. For MI, P300, and SSVEP, all data were obtained from MOABB with its default preprocessing, including channel selection, filtering, and epoch extraction. The analysis windows and filter ranges are listed in Table 6. No additional trial normalization or data augmentation was applied.

For speech decoding, we used MEG channels for text-MEG and ECoG channels for bouchard2019. Text-MEG recordings natively contain 57–61 MEG channels per subject, and bouchard2019 removes electrodes marked bad in any of a subject's sessions. We used high-gamma amplitude envelopes instead of raw broadband waveforms, following prior speech-decoding work (Jia et al., 2025; Bouchard et al., 2013). After common average referencing and filtering, the 70–170 Hz signals underwent robust scaling and clipping before Hilbert envelope extraction and resampling to 400 Hz. Each text-MEG trial contains 640 samples, and each bouchard2019 trial contains 400 samples.

For EEGMAT (Zyma et al., 2019), we retained 19 EEG channels at the native 500 Hz sampling rate. We used the final 60 s of the resting recording and the first 60 s of the arithmetic recording, dividing each into non-overlapping 4 s windows. No additional filtering, trial normalization, or data augmentation was applied during this extraction.

For the pretrained baselines LaBraM, CBraMod, and REVE, we resampled every evaluated EEG dataset to 200 Hz to match their pretrained input rate. Channels were retained or selected according to each model's input representation, and temporal inputs were divided into model specific patches as described below. All three foundation models were pretrained on EEG and were therefore evaluated only on the EEG datasets. The two speech datasets were excluded.

Table 6: Preprocessing details of eleven downstream datasets.

Dataset	Duration (s)	Interval (s)	Bandpass (Hz)
BNCI2014_001 (Tangermann et al., 2012)	4.0	[2,6]	8–32
BNCI2014_002 (Steyrl et al., 2016)	5.0	[3,8]	8–32
BNCI2014_004 (Tangermann et al., 2012)	4.5	[3,7.5]	8–32
BNCI2015_001 (Faller et al., 2012)	5.0	[0,5]	8–32
BNCI2014_008 (Riccio et al., 2013)	1.0	[0,1]	1–24
BNCI2014_009 (Aricò et al., 2014)	0.8	[0,0.8]	1–24
Nakanishi2015 (Nakanishi et al., 2015)	4.15	[0.15,4.3]	7–45
Wang2016 (Wang et al., 2017)	5.0	[0.5,5.5]	7–45
EEGMAT (Zyma et al., 2019)	4.0	Consecutive windows	No additional filter
text-MEG (Jia et al., 2025)	1.6	[0,1.6] from trial start	70–170
bouchard2019 (Bouchard & Chang, 2020)	1.0	[0,1] from speech onset	70–170

C.2 Channel Names and Locations

Table 7 lists the retained input channels for the datasets cited in Table 6. Speech datasets are listed by subject because their retained channels differ across subjects. The channel list is constant across sessions within each speech subject. Subjects with identical channel lists are grouped together. Figure 6 shows the EEG electrode positions in the international 10–10 system (Chatrian et al., 1985). MEG sensor layouts depend on the acquisition system (Hämäläinen et al., 1993), whereas ECoG electrode locations depend on each subject's implantation scheme (Bouchard et al., 2013). Neither

modality shares a universal layout analogous to the EEG 10–10 system, so we do not provide a common position diagram for MEG or ECoG.

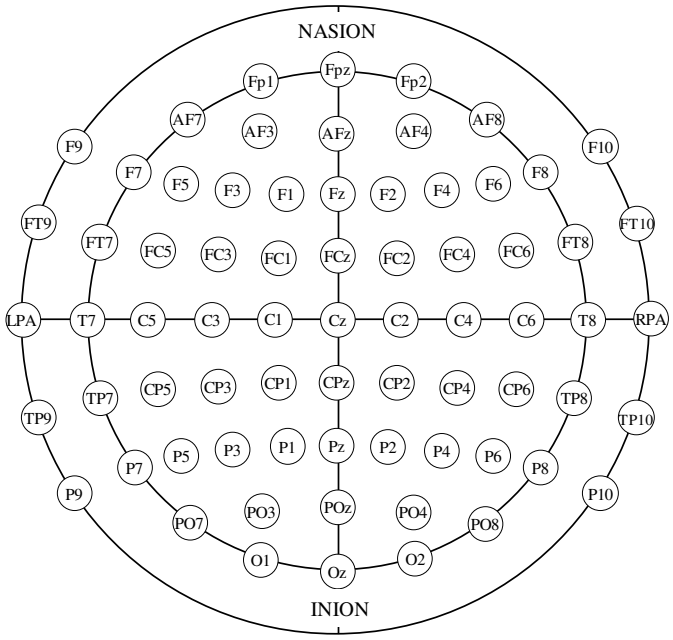

Figure 6: EEG electrode positions in the international 10–10 system.

Table 7: Channel names for all eleven datasets.

Dataset	Subject	C	Channel names
BNCI2014_001	All	22	Fz, FC3, FC1, FCz, FC2, FC4, C5, C3, C1, Cz, C2, C4, C6, CP3, CP1, CPz, CP2, CP4, P1, Pz, P2, POz
BNCI2014_002	All	15	EEG1, EEG2, EEG3, EEG4, EEG5, EEG6, EEG7, EEG8, EEG9, EEG10, EEG11, EEG12, EEG13, EEG14, EEG15
BNCI2014_004	All	3	C3, CZ, C4
BNCI2015_001	All	13	FC3, FCZ, FC4, C5, C3, C1, CZ, C2, C4, C6, CP3, CPZ, CP4
BNCI2014_008	All	8	Fz, Cz, Pz, Oz, P3, P4, PO7, PO8
BNCI2014_009	All	16	Fz, Cz, Pz, Oz, P3, P4, PO7, PO8, F3, F4, FCz, C3, C4, CP3, CPz, CP4
Nakanishi2015	All	8	PO7, PO3, POz, PO4, PO8, O1, Oz, O2
Wang2016	All	62	Fp1, Fpz, Fp2, AF3, AF4, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, M1, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, M2, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO5, PO3, POz, PO4, PO6, PO8, O1, Oz, O2
EEGMAT	All	19	Fp1, Fp2, F3, F4, F7, F8, T3, T4, C3, C4, T5, T6, P3, P4, O1, O2, Fz, Cz, Pz
text-MEG	S01–S06	61	SM1z, SM4z, SM5z, SM7z, SL2z, SL3z, AL1z, SL5z, SR2z, SR3z, AR1z, SR5z, MM2z, MM4z, MM5z, MM6z, MM7z, MM8z, MM9z, ML2z, AL2z, ML4z, ML5z, ML6z, ML7z, ML8z, MR2z, AR2z, MR4z, MR5z, MR6z, MR7z, MR8z, IM1z, AR6z, AL6z, IM5z, IM6z, IM7z, IM8z IL1z, IL3z, IL4z, IL5z, IL6z, IL7z, IR1z, IR3z, IR4z, IR5z, IR6z, IR7z, AM1z, AM3z, AM5z, AL3z, AL5z, AM4z, AR3z, AR5z, AM6z
text-MEG	S07	59	SM1z, SM4z, SM5z, SM7z, SL2z, SL3z, AL1z, SL5z, SR2z, SR3z, AR1z, SR5z, MM2z, MM4z, MM5z, MM6z, MM7z, MM8z, MM9z, ML2z, AL2z, ML4z, ML5z, ML6z, ML7z, ML8z, MR2z, MR4z, MR5z, MR6z, MR7z, MR8z, IM1z, AR6z, IM5z, IM6z, IM7z, IM8z, IL1z, IL3z IL4z, IL5z, IL6z, IL7z, IR1z, IR3z, IR4z, IR5z, IR6z, IR7z, AM1z, AM3z, AM5z, AL3z, AL5z, AM4z, AR3z, AR5z, AM6z

Continued on next page

Table 7 (continued)

Dataset	Subject	C	Channel names
text-MEG	S08	60	SM1z, SM4z, SM5z, SM7z, SL2z, SL3z, AL1z, SL5z, SR2z, SR3z, AR1z, SR5z, MM2z, MM4z, MM5z, MM6z, MM7z, MM8z, MM9z, ML2z, AL2z, ML4z, ML5z, ML6z, ML7z, ML8z, MR2z, MR4z, MR5z, MR6z, MR7z, MR8z, IM1z, AR6z, AL6z, IM5z, IM6z, IM7z, IM8z, IL1z IL3z, IL4z, IL5z, IL6z, IL7z, IR1z, IR3z, IR4z, IR5z, IR6z, IR7z, AM1z, AM3z, AM5z, AL3z, AL5z, AM4z, AR3z, AR5z, AM6z
text-MEG	S09	57	SM1z, SM4z, SM5z, SM7z, SL2z, SL3z, AL1z, SL5z, SR2z, SR3z, AR1z, SR5z, MM2z, MM4z, MM5z, MM6z, MM7z, MM8z, MM9z, ML2z, AL2z, ML4z, ML5z, ML6z, ML7z, ML8z, MR2z, MR4z, MR5z, MR6z, MR7z, MR8z, IM1z, AR6z, IM5z, IM6z, IM7z, IM8z, IL3z, IL4z IL5z, IL6z, IL7z, IR1z, IR3z, IR4z, IR5z, IR6z, IR7z, AM1z, AM5z, AL3z, AL5z, AM4z, AR3z, AR5z, AM6z
bouchard2019	EC2	246	ch000, ch001, ch002, ch003, ch004, ch005, ch006, ch007, ch008, ch009, ch010, ch011, ch012, ch013, ch014, ch015, ch016, ch017, ch018, ch019, ch020, ch021, ch022, ch023, ch024, ch025, ch026, ch027, ch028, ch029, ch030, ch031, ch032, ch033, ch034, ch035, ch036, ch037, ch038, ch039 ch040, ch041, ch042, ch043, ch044, ch045, ch046, ch047, ch049, ch050, ch051, ch052, ch053, ch054, ch055, ch056, ch057, ch058, ch059, ch060, ch061, ch062, ch063, ch064, ch065, ch066, ch067, ch068, ch069, ch070, ch071, ch072, ch073, ch074, ch075, ch076, ch077, ch078, ch079, ch080 ch081, ch082, ch083, ch084, ch085, ch086, ch087, ch088, ch089, ch090, ch091, ch092, ch093, ch094, ch095, ch096, ch097, ch098, ch099, ch100, ch101, ch102, ch103, ch104, ch105, ch106, ch107, ch108, ch109, ch110, ch111, ch116, ch117, ch118, ch119, ch120, ch121, ch122, ch124, ch125 ch126, ch128, ch129, ch130, ch131, ch132, ch133, ch134, ch135, ch136, ch137, ch138, ch139, ch140, ch141, ch142, ch143, ch144, ch145, ch146, ch147, ch148, ch149, ch150, ch151, ch152, ch153, ch154, ch155, ch156, ch157, ch158, ch159, ch160, ch161, ch162, ch163, ch164, ch165, ch166 ch167, ch168, ch169, ch170, ch171, ch172, ch173, ch174, ch175, ch176, ch177, ch178, ch179, ch180, ch181, ch182, ch183, ch184, ch185, ch186, ch187, ch188, ch189, ch190, ch191, ch193, ch194, ch195, ch196, ch197, ch198, ch199, ch200, ch201, ch202, ch203, ch204, ch205, ch206, ch207 ch208, ch209, ch210, ch211, ch212, ch213, ch214, ch215, ch216, ch217, ch218, ch219, ch220, ch221, ch222, ch223, ch224, ch225, ch226, ch227, ch228, ch229, ch230, ch231, ch232, ch233, ch234, ch235, ch236, ch237, ch239, ch241, ch242, ch243, ch244, ch245, ch246, ch247, ch248, ch249 ch250, ch251, ch252, ch253, ch254, ch255
bouchard2019	EC9	249	ch000, ch001, ch002, ch003, ch004, ch005, ch006, ch007, ch008, ch009, ch010, ch011, ch012, ch013, ch014, ch015, ch016, ch017, ch018, ch019, ch020, ch021, ch022, ch023, ch024, ch025, ch026, ch027, ch028, ch029, ch030, ch031, ch032, ch033, ch034, ch035, ch036, ch037, ch038, ch039 ch040, ch041, ch042, ch043, ch044, ch045, ch046, ch047, ch048, ch049, ch050, ch051, ch052, ch053, ch054, ch055, ch056, ch057, ch058, ch059, ch060, ch061, ch062, ch063, ch067, ch068, ch069, ch070, ch071, ch072, ch073, ch074, ch075, ch076, ch077, ch078, ch079, ch080, ch081, ch082 ch083, ch084, ch085, ch086, ch087, ch088, ch090, ch091, ch092, ch093, ch094, ch095, ch096, ch097, ch098, ch099, ch100, ch101, ch102, ch103, ch104, ch105, ch106, ch107, ch108, ch109, ch110, ch111, ch112, ch113, ch114, ch115, ch116, ch117, ch118, ch119, ch120, ch121, ch122, ch123 ch124, ch125, ch126, ch127, ch128, ch129, ch130, ch131, ch132, ch133, ch134, ch135, ch136, ch137, ch138, ch139, ch140, ch141, ch142, ch143, ch144, ch145, ch146, ch147, ch148, ch149, ch150, ch151, ch152, ch153, ch154, ch155, ch156, ch157, ch158, ch159, ch160, ch161, ch162, ch163 ch164, ch165, ch166, ch167, ch168, ch169, ch170, ch171, ch172, ch173, ch174, ch175, ch176, ch177, ch178, ch179, ch180, ch181, ch182, ch183, ch184, ch185, ch186, ch187, ch188, ch189, ch190, ch191, ch192, ch193, ch194, ch195, ch196, ch197, ch198, ch199, ch200, ch201, ch202, ch203 ch204, ch205, ch206, ch207, ch208, ch209, ch210, ch211, ch212, ch213, ch214, ch215, ch216, ch217, ch218, ch219, ch220, ch221, ch222, ch223, ch224, ch225, ch226, ch227, ch228, ch229, ch230, ch231, ch232, ch233, ch234, ch235, ch236, ch237, ch238, ch239, ch240, ch241, ch242, ch243 ch246, ch247, ch248, ch249, ch250, ch251, ch252, ch253, ch254
bouchard2019	GP31	242	ch000, ch001, ch002, ch003, ch004, ch005, ch006, ch007, ch008, ch009, ch010, ch011, ch012, ch013, ch014, ch015, ch016, ch017, ch019, ch020, ch021, ch022, ch023, ch024, ch025, ch026, ch027, ch028, ch029, ch030, ch031, ch032, ch033, ch034, ch035, ch036, ch037, ch038, ch039, ch040 ch041, ch042, ch043, ch044, ch045, ch046, ch047, ch049, ch050, ch051, ch052, ch053, ch054, ch055, ch056, ch057, ch058, ch059, ch060, ch061, ch062, ch063, ch067, ch068, ch069, ch070, ch071, ch072, ch073, ch074, ch075, ch076, ch077, ch078, ch079, ch080, ch081, ch082, ch083, ch084

Continued on next page

Table 7 (continued)

Dataset	Subject	C	Channel names
			ch085, ch086, ch087, ch088, ch089, ch090, ch091, ch092, ch093, ch094, ch095, ch096, ch097, ch098, ch099, ch100, ch101, ch102, ch103, ch104, ch105, ch106, ch107, ch108, ch109, ch110, ch111, ch116, ch117, ch118, ch119, ch120, ch121, ch122, ch124, ch125, ch126, ch128, ch129, ch130
			ch131, ch132, ch133, ch134, ch135, ch136, ch137, ch138, ch139, ch140, ch141, ch142, ch143, ch144, ch145, ch146, ch147, ch148, ch149, ch150, ch151, ch152, ch153, ch154, ch155, ch156, ch157, ch158, ch159, ch160, ch161, ch162, ch163, ch164, ch165, ch166, ch167, ch168, ch169, ch170
			ch171, ch172, ch173, ch174, ch175, ch176, ch177, ch178, ch179, ch180, ch181, ch182, ch183, ch184, ch185, ch186, ch187, ch188, ch189, ch190, ch191, ch192, ch193, ch194, ch196, ch197, ch198, ch199, ch200, ch202, ch203, ch204, ch205, ch206, ch207, ch208, ch209, ch210, ch211, ch212
			ch213, ch214, ch215, ch216, ch217, ch218, ch219, ch220, ch221, ch222, ch223, ch224, ch225, ch226, ch227, ch228, ch229, ch230, ch231, ch232, ch233, ch234, ch235, ch236, ch237, ch238, ch239, ch240, ch241, ch242, ch243, ch244, ch245, ch246, ch247, ch248, ch249, ch250, ch251, ch252
			ch253, ch254
bouchard2019	GP33	229	ch000, ch001, ch002, ch003, ch004, ch005, ch006, ch007, ch008, ch009, ch010, ch011, ch012, ch013, ch014, ch015, ch016, ch017, ch020, ch021, ch022, ch023, ch024, ch025, ch026, ch027, ch028, ch029, ch030, ch031, ch032, ch033, ch034, ch035, ch036, ch037, ch038, ch039, ch040, ch041
			ch042, ch043, ch044, ch045, ch046, ch047, ch049, ch050, ch051, ch052, ch053, ch054, ch055, ch056, ch057, ch058, ch059, ch060, ch061, ch062, ch063, ch064, ch065, ch066, ch067, ch068, ch069, ch070, ch071, ch072, ch073, ch074, ch075, ch076, ch077, ch078, ch079, ch080, ch081, ch082
			ch083, ch084, ch085, ch086, ch087, ch088, ch089, ch090, ch091, ch092, ch093, ch094, ch095, ch096, ch097, ch098, ch099, ch100, ch101, ch102, ch103, ch104, ch105, ch106, ch107, ch108, ch109, ch110, ch111, ch116, ch117, ch118, ch119, ch120, ch121, ch122, ch124, ch125, ch126, ch128
			ch129, ch130, ch131, ch132, ch133, ch134, ch135, ch136, ch137, ch138, ch139, ch140, ch141, ch142, ch143, ch144, ch145, ch146, ch147, ch148, ch149, ch150, ch151, ch152, ch153, ch154, ch155, ch156, ch157, ch158, ch159, ch160, ch161, ch162, ch163, ch164, ch165, ch166, ch167, ch168
			ch169, ch170, ch171, ch172, ch173, ch174, ch175, ch176, ch177, ch178, ch179, ch180, ch181, ch182, ch183, ch184, ch185, ch186, ch187, ch188, ch189, ch190, ch191, ch193, ch194, ch195, ch196, ch197, ch198, ch199, ch200, ch201, ch202, ch203, ch204, ch205, ch206, ch207, ch224, ch225
			ch226, ch227, ch228, ch229, ch230, ch231, ch232, ch233, ch234, ch235, ch236, ch237, ch238, ch239, ch240, ch241, ch242, ch243, ch244, ch245, ch246, ch247, ch248, ch249, ch250, ch251, ch252, ch253, ch254

C.3 BASELINES

We evaluated six compact models and three advanced pretrained EEG foundation models. The following descriptions summarize the main architectural ideas in their original papers.

EEGNet. EEGNet is a compact CNN designed to decode EEG across different BCI paradigms. Depthwise and separable convolutions incorporate established principles for EEG feature extraction while reducing the number of trainable parameters (Lawhern et al., 2018).

FBCNet. FBCNet uses a multi-view representation of EEG and spatial filtering to learn discriminative spectral–spatial features. A variance layer aggregates temporal information, supporting efficient learning from limited MI training data (Mane et al., 2021).

IFNet. IFNet is a lightweight CNN that learns interactions between low and high frequency spectral–spatial representations for MI decoding. It combines these representations through elementwise addition and temporal average pooling (Wang et al., 2023).

Conformer. Conformer combines temporal and spatial convolutions for local feature extraction with self-attention for global temporal dependencies. Its unified CNN–Transformer architecture is designed to capture both local and global EEG features (Song et al., 2023).

DBConformer. DBConformer uses temporal and spatial Conformer branches to model long-range temporal dependencies and inter-channel interactions. A lightweight channel-attention module further refines the spatial representation (Wang et al., 2026).

Deformer. Deformer introduces a hierarchical coarse-to-fine Transformer with a fine-grained temporal-learning branch to capture EEG dynamics at multiple resolutions. Its dense information-purification module integrates refined temporal information from multiple levels (Ding et al., 2025).

CBraMod. CBraMod is an EEG foundation model whose criss-cross Transformer uses parallel attention mechanisms to model spatial and temporal dependencies separately. It combines asymmetric conditional positional encoding with masked EEG patch reconstruction to support transfer across diverse EEG formats and downstream tasks (Wang et al., 2025).

LaBraM. LaBraM learns generic EEG representations by dividing recordings into channel patches and using a vector-quantized neural tokenizer to encode them as discrete neural codes. Its Transformer is pretrained to predict the codes of masked patches and then fine-tuned for downstream tasks (Jiang et al., 2024).

REVE. REVE combines overlapping temporal patches with four dimensional positional encoding derived from electrode coordinates and time. Its Transformer is pretrained with masked EEG reconstruction and supports varying channel layouts and signal durations (El Ouahidi et al., 2025).

Table 8 reports the configurations used in our experiments, based on the models' public implementations.

Table 8: Key hyperparameters of the baselines.

Model	Key parameters
EEGNet	$F_1 = 8$, depth multiplier 2, $F_2 = 16$, dropout 0.25
FBCNet	9 filter bands, 32 spatial filters per band, stride factor 4
IFNet	radix 2, width 64, kernel length 63, patch length 125
Conformer	embedding 40, 10 heads, depth 8, dropout 0.5, pooling 75/15
DBConformer	temporal/spatial widths 40/40, depths 2/2, 8 heads, dropout 0.5
Deformer	64 kernels, depth 4, 16 heads, MLP/head dims 16/16, dropout 0.5
CBraMod	$d_{\text{model}} = 200$, FFN 800, 12 layers, 8 heads, single linear head, dropout 0.5
LaBraM	patch length 200, dynamic patch count, embedding 200, 12 layers, 10 heads
REVE	patch 200, overlap 20, embedding 512, 22 layers, 8 heads, LoRA rank/alpha 16/16
PyraNet	$V = 64$, $D = 32$, $M = 4$, $G = 4$, kernel 5, grids $\{8, 32, 128\}$

C.4 Training Pipeline Details for Foundation Models

This section describes the downstream pipeline used for LaBraM, CBraMod, and REVE. The model specific procedures described below were implemented following the corresponding public code repositories. The only difference is the use of a unified experimental protocol to ensure that all models were evaluated under the same criteria. All three models used the same LOSO folds as the other EEG baselines. Every trial was first resampled to 200 Hz. EA was then applied separately to each subject before the LOSO split. The training subjects were pooled and randomly divided at the trial level into an approximately 80% fitting set and a 20% validation set. Binary tasks used validation AUROC for model selection. Multiclass tasks used validation Cohen's κ. The held-out subject was used only for testing.

C.4.1 LaBraM

Input preparation. The 200 Hz trial was passed to LaBraM. Channel names were converted to a common uppercase format. For a named montage, channels found in the official LaBraM 10–20 position table were retained and assigned their official position indices. Channels outside this table were not passed to the model.

Channel and temporal adaptation. LaBraM used the number of retained channels instead of expanding every dataset to a fixed channel template. The number of temporal patches was determined from the resampled trial length. Each trial was divided into the largest possible number of non-overlapping 200-sample patches. Samples remaining after the last complete patch were discarded, whereas a trial shorter than 200 samples was padded with zeros on the right to form one patch.

Model and task head. We loaded the public LaBraM-base checkpoint and restored all compatible backbone parameters. The original prediction head was removed. The backbone returned the class token and every channel-time token. These tokens were flattened and passed through dropout with probability 0.5 and a linear task head. The backbone and the new task head were fine-tuned together.

Optimization and model selection. LaBraM was trained for 50 epochs with AdamW and batch size 64. The peak learning rate was 5×10^{-4} and the weight decay was 0.05. The learning rate increased during the first five epochs and then followed cosine decay to 10^{-6}. Layerwise learning rate decay was 0.65. Drop-path probability was 0.1 and label smoothing was 0.1. The P300 datasets used cross entropy with class weights because their classes were imbalanced. Training always ran for 50 epochs. The checkpoint with the best validation AUROC or Cohen's κ was restored for testing.

C.4.2 CBraMod

Input preparation. The 200 Hz trial was passed to CBraMod with its retained channels in the order provided by each dataset. We used the actual number of channels in each dataset without mapping them to a fixed channel template, reordering them, or adding channels filled with zeros.

Channel and temporal adaptation. The number of temporal patches was determined from the resampled trial length. Each channel was divided into the largest possible number of non-overlapping 200-sample patches. Samples remaining after the last complete patch were discarded, whereas a trial shorter than 200 samples was padded with zeros on the right to form one patch.

Model and task head. We loaded the public CBraMod checkpoint with strict parameter matching. The pretrained output projection was replaced by an identity operation. All encoder tokens from the retained channels and temporal patches were flattened and passed directly to a single linear classifier. The complete backbone and classifier were fine-tuned together.

Optimization and model selection. CBraMod was trained for 50 epochs with AdamW and batch size 64. The initial learning rate was 10^{-4} and the weight decay was 0.05. Cosine annealing reduced the learning rate to 10^{-6} without a warm-up stage. The gradient norm was clipped at 1.0. Multiclass tasks used label smoothing of 0.1. The P300 datasets instead used cross entropy with class weights. Training always ran for 50 epochs. The checkpoint with the best validation AUROC or Cohen's κ was restored for testing.

C.4.3 REVE

Input preparation. The 200 Hz trial was passed to REVE. REVE used the available dataset channels rather than a fixed channel template. Each channel name was matched to the public REVE position bank and assigned a three dimensional sensor coordinate.

Channel and temporal adaptation. REVE retained the full resampled trial instead of forcing every dataset to 800 samples. It formed temporal patches of 200 samples with an overlap of 20 samples. Trials shorter than 200 samples were padded with zeros on the right to form one patch. During training, Gaussian noise with standard deviation 0.0025 was added to the sensor coordinates. The coordinates were left unchanged during validation and testing.

Model and task head. We loaded the public REVE-Base checkpoint. The task head used a learned class query to attend to all encoder tokens. The resulting context token was concatenated with all encoder tokens. The combined representation was flattened and passed through RMS normalization, dropout, and a linear classifier. The first stage froze the encoder and trained the task head and class query. In the second stage, the pretrained encoder and task head remained frozen, while Low-Rank Adaptation (LoRA) modules were trained in the query, key, value, and output projections of every attention block. The LoRA rank and scaling factor were both 16.

Optimization and model selection. Both stages used the public StableAdamW implementation with $\beta = (0.92, 0.999)$ and $\epsilon = 10^{-9}$. The batch size was 64. Training used mixed precision,

mixup, and gradient clipping at 2.0. During the first stage, the four MI datasets used learning rate 5×10^{-3} and weight decay 10^{-3}. The other datasets used learning rate 10^{-3} and weight decay 10^{-2}. The first stage allowed at most 200 epochs. Its learning rate used one global exponential warm-up over the first ten epochs. After warm-up, the learning rate was reduced by a factor of 0.5 when the validation score did not improve for five epochs. The first stage used early-stopping patience 15 and dropout 0.3. The second stage trained only the LoRA modules for five epochs with learning rate 10^{-4}, weight decay 10^{-2}, and dropout 0.5. A new optimizer was created when the second stage began. The best validation checkpoint from the second stage was restored for testing.

We kept the model specific training operations above but used the common benchmark protocol for data splitting and evaluation. In particular, we used LOSO, subject-wise EA, the random fitting and validation split, AUROC or Cohen's κ for model selection, and one test evaluation after training.

C.5 Metrics

Balanced accuracy (BAcc). BAcc is the average of per-class recall (Brodersen et al., 2010). It gives equal weight to all classes, thereby balancing the impact of misclassifications across majority and minority classes. It is reported as a percentage for all datasets.

Area under the precision–recall curve (AUC–PR). For binary tasks, AUC–PR is the area under the precision–recall curve obtained by varying the classification threshold (Saito & Rehmsmeier, 2015). It reflects the model's ability to identify positive samples while limiting false positives. It is reported as a percentage.

Area under the receiver operating characteristic curve (AUROC). For binary tasks, AUROC is the area under the curve plotting the true positive rate against the false positive rate across different classification thresholds (Fawcett, 2006). It reflects the model's ability to distinguish positive from negative samples. It is reported as a percentage.

Cohen's κ. For multiclass tasks, Cohen's κ measures the agreement between predicted and true labels while accounting for agreement expected by chance (Cohen, 1960). It reflects classification consistency beyond chance level agreement and is particularly useful when class distributions are imbalanced.

Weighted F1. For multiclass tasks, weighted F1 is the average of per-class F1 scores weighted by the number of samples in each class (Sokolova & Lapalme, 2009). It jointly considers precision and recall while accounting for class frequency, allowing performance across classes to be summarized in a single metric. It is reported as a percentage.

Within each seed, EEG results average test subjects, while speech results first average held-out sessions within each subject and then subjects. Reported standard deviations are population standard deviations across three seed means.

C.6 Training Configuration

Table 9 summarizes training hyperparameters. We used AdamW (Loshchilov & Hutter, 2019) for all models except REVE, which used StableAdamW. All experiments used a random 20% validation split and seeds 42, 43, and 44. The supervised models used early stopping based on validation AUROC for binary tasks and Cohen's κ for multiclass tasks. LaBraM and CBraMod completed all 50 epochs and restored the best validation checkpoint. REVE used the two-stage stopping and checkpoint rules described in Appendix C.4.

Table 9: Hyperparameters for downstream task training.

Hyperparameter	Supervised EEG	text-MEG	bouchard2019	LaBraM	CBraMod	REVE
Maximum epochs	100	200	200	50	50	200 LP + 5 LoRA-FT
Early-stopping patience	5	30	30	None	None	15 in LP
Batch size	32	64	32	64	64	64
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	StableAdamW
Learning rate	3×10^{-4}	10^{-3}	10^{-3}	5×10^{-4}	10^{-4}	5×10^{-3}/10^{-3} LP, 10^{-4} FT
Weight decay	10^{-2}	10^{-4}	10^{-4}	5×10^{-2}	5×10^{-2}	10^{-3}/10^{-2} LP, 10^{-2} FT
Learning-rate schedule	Constant	Cosine	Cosine	Warm-up + cosine	Cosine	10-epoch exponential warm-up + plateau in LP
Adaptation	Full	Full	Full	Full	Full	LP + LoRA rank/alpha 16/16
Validation fraction	0.2	0.2	0.2	0.2	0.2	0.2

D Detailed Downstream Results

The following are the detailed results for each paradigm, including the complete set of evaluation metrics. Values are seed means ± standard deviations. Bold and underlined values indicate the best and second-best results, respectively.

D.1 MI

Table 10: Full metrics for the four MI datasets.

Method	BNCI2014_001 BAcc	BNCI2014_001 κ	BNCI2014_001 Weighted F1	BNCI2014_002 BAcc	BNCI2014_002 AUC–PR	BNCI2014_002 AUROC
EEGNet	51.54 ± 1.44	0.3538 ± 0.0191	50.55 ± 1.49	74.90 ± 0.53	82.07 ± 0.33	82.99 ± 0.54
FBCNet	47.23 ± 1.16	0.2964 ± 0.0154	45.76 ± 1.09	71.88 ± 0.46	78.23 ± 0.05	79.25 ± 0.15
IFNet	<u>52.49</u> ± 0.18	<u>0.3665</u> ± 0.0024	<u>51.27</u> ± 0.26	76.16 ± 0.90	<u>84.53</u> ± 0.09	<u>85.14</u> ± 0.17
Conformer	50.01 ± 0.60	0.3335 ± 0.0080	49.05 ± 0.53	<u>76.70</u> ± 0.58	82.72 ± 0.27	83.63 ± 0.13
DBConformer	49.25 ± 0.85	0.3233 ± 0.0113	48.11 ± 0.66	75.34 ± 0.75	83.09 ± 0.39	84.09 ± 0.26
Deformer	42.53 ± 1.29	0.2338 ± 0.0172	38.12 ± 1.86	64.85 ± 1.45	84.18 ± 0.07	84.76 ± 0.13
CBraMod	47.37 ± 0.12	0.2983 ± 0.0016	46.88 ± 0.29	69.35 ± 0.91	74.77 ± 0.29	76.19 ± 0.23
LaBraM	43.51 ± 0.69	0.2468 ± 0.0092	42.89 ± 0.97	65.52 ± 0.37	71.06 ± 0.18	72.50 ± 0.27
REVE	43.99 ± 0.40	0.2533 ± 0.0054	43.66 ± 0.40	62.57 ± 1.31	71.68 ± 0.97	72.29 ± 0.75
PyraNet	**53.57** ± 0.26	**0.3809** ± 0.0035	**52.81** ± 0.38	**79.46** ± 0.38	**84.95** ± 0.14	**85.68** ± 0.05

Method	BNCI2014_004 BAcc	BNCI2014_004 AUC–PR	BNCI2014_004 AUROC	BNCI2015_001 BAcc	BNCI2015_001 AUC–PR	BNCI2015_001 AUROC
EEGNet	74.24 ± 0.28	<u>81.92</u> ± 0.13	<u>82.07</u> ± 0.13	74.61 ± 0.57	81.76 ± 0.71	82.36 ± 0.74
FBCNet	70.57 ± 0.16	76.95 ± 0.16	77.64 ± 0.04	72.22 ± 0.38	82.26 ± 0.46	82.73 ± 0.31
IFNet	**74.47** ± 0.22	81.77 ± 0.03	**82.10** ± 0.08	**75.37** ± 0.61	**84.70** ± 0.22	**85.19** ± 0.27
Conformer	72.98 ± 0.14	80.66 ± 0.29	80.79 ± 0.19	73.95 ± 0.20	82.96 ± 0.65	83.35 ± 0.56
DBConformer	<u>74.37</u> ± 0.03	81.65 ± 0.05	81.93 ± 0.04	73.56 ± 0.22	82.40 ± 0.23	83.02 ± 0.34
Deformer	72.71 ± 0.74	**81.97** ± 0.13	81.93 ± 0.20	71.07 ± 0.90	<u>83.71</u> ± 0.57	<u>84.06</u> ± 0.48
CBraMod	73.07 ± 0.25	80.44 ± 0.35	80.60 ± 0.35	69.50 ± 0.53	75.64 ± 0.42	76.38 ± 0.48
LaBraM	72.45 ± 0.29	79.25 ± 0.13	79.74 ± 0.08	69.06 ± 0.19	75.36 ± 0.13	76.18 ± 0.17
REVE	70.19 ± 0.57	75.97 ± 0.15	76.56 ± 0.26	69.68 ± 0.45	75.47 ± 0.14	76.73 ± 0.25
PyraNet	74.24 ± 0.30	80.84 ± 0.09	81.21 ± 0.07	<u>75.03</u> ± 0.43	83.01 ± 0.19	82.99 ± 0.21

BNCI2014_001. BNCI2014_001 (Tangermann et al., 2012), also known as BCI Competition IV dataset 2a, contains recordings from nine participants performing four-class motor imagery (left hand, right hand, both feet, tongue) across two sessions, recorded with 22 EEG channels at 250 Hz. Each 4.0 s window covers 2–6 s relative to trial onset and is bandpass-filtered to 8–32 Hz. Multiclass metrics are reported.

BNCI2014_002. BNCI2014_002 (Steyrl et al., 2016) contains recordings from fourteen participants performing two-class motor imagery (right hand, both feet) in a single session, recorded with 15 EEG channels at 512 Hz. Each 5.0 s window covers 3–8 s relative to trial onset and is bandpass-filtered to 8–32 Hz. Binary metrics are reported.

BNCI2014_004. BNCI2014_004 (Tangermann et al., 2012), also known as BCI Competition IV dataset 2b, contains recordings from nine participants performing two-class motor imagery

(left/right hand) across five sessions, recorded with 3 EEG channels at 250 Hz. Each 4.5 s window covers 3–7.5 s relative to trial onset and is bandpass-filtered to 8–32 Hz. Binary metrics are reported.

BNCI2015_001. BNCI2015_001 (Faller et al., 2012) contains recordings from twelve participants performing two-class motor imagery (right hand, both feet) across 2–3 sessions, recorded with 13 EEG channels at 512 Hz. Each 5.0 s window covers 0–5 s and is bandpass-filtered to 8–32 Hz. Binary metrics are reported.

D.2 P300 AND SSVEP

Table 11: Full metrics for P300 and SSVEP.

Method	BNCI2014_008 BAcc	BNCI2014_008 AUC–PR	BNCI2014_008 AUROC	BNCI2014_009 BAcc	BNCI2014_009 AUC–PR	BNCI2014_009 AUROC
EEGNet	71.59 ± 0.07	45.44 ± 0.44	78.95 ± 0.12	81.52 ± 0.39	69.97 ± 0.50	**89.65** ± 0.19
FBCNet	53.49 ± 0.33	21.66 ± 0.38	56.26 ± 0.22	61.12 ± 1.08	32.57 ± 0.59	68.31 ± 0.44
IFNet	66.17 ± 0.35	35.68 ± 0.23	72.56 ± 0.24	76.77 ± 0.23	59.50 ± 0.18	85.68 ± 0.06
Conformer	70.88 ± 0.33	44.62 ± 0.41	77.97 ± 0.17	78.88 ± 0.18	63.25 ± 0.37	86.72 ± 0.19
DBConformer	71.76 ± 0.14	47.05 ± 0.42	**79.45** ± 0.15	80.52 ± 0.57	69.53 ± 0.56	89.41 ± 0.26
Deformer	71.52 ± 0.30	46.22 ± 0.27	79.00 ± 0.26	81.42 ± 0.60	70.27 ± 0.31	89.40 ± 0.16
CBraMod	70.91 ± 0.59	45.87 ± 0.54	78.85 ± 0.28	79.74 ± 0.39	69.05 ± 0.45	88.89 ± 0.08
LaBraM	71.19 ± 0.09	**47.35** ± 0.31	79.37 ± 0.24	79.53 ± 0.19	67.94 ± 0.25	88.75 ± 0.08
REVE	71.40 ± 0.06	45.19 ± 0.21	78.65 ± 0.10	80.41 ± 0.16	67.88 ± 0.17	88.56 ± 0.05
PyraNet	**71.76** ± 0.12	44.85 ± 0.14	78.85 ± 0.12	**81.52** ± 0.14	**70.61** ± 0.39	89.59 ± 0.09

Method	Nakanishi2015 BAcc	Nakanishi2015 κ	Nakanishi2015 Weighted F1	Wang2016 BAcc	Wang2016 κ	Wang2016 Weighted F1
EEGNet	92.88 ± 0.81	0.9223 ± 0.0088	92.87 ± 0.70	90.56 ± 0.19	0.9032 ± 0.0019	90.22 ± 0.22
FBCNet	38.17 ± 1.88	0.3255 ± 0.0205	34.64 ± 2.26	26.19 ± 0.40	0.2430 ± 0.0041	20.73 ± 0.30
IFNet	76.11 ± 1.98	0.7394 ± 0.0216	75.62 ± 2.17	66.92 ± 0.33	0.6607 ± 0.0034	65.58 ± 0.37
Conformer	81.75 ± 4.92	0.8009 ± 0.0537	81.07 ± 5.25	90.42 ± 0.15	0.9017 ± 0.0016	90.12 ± 0.18
DBConformer	30.31 ± 1.40	0.2397 ± 0.0152	23.90 ± 2.06	31.47 ± 1.25	0.2972 ± 0.0129	26.54 ± 1.50
Deformer	95.70 ± 0.42	0.9531 ± 0.0046	95.64 ± 0.40	**94.15** ± 0.15	**0.9400** ± 0.0016	**93.94** ± 0.18
CBraMod	95.27 ± 0.28	0.9484 ± 0.0030	95.17 ± 0.27	92.80 ± 0.11	0.9262 ± 0.0012	92.55 ± 0.16
LaBraM	92.18 ± 0.68	0.9147 ± 0.0075	92.10 ± 0.75	91.92 ± 0.25	0.9171 ± 0.0025	91.64 ± 0.23
REVE	90.02 ± 0.71	0.8911 ± 0.0077	89.94 ± 0.71	60.36 ± 0.24	0.5934 ± 0.0025	59.13 ± 0.38
PyraNet	**97.10** ± 0.05	**0.9684** ± 0.0005	**97.04** ± 0.05	94.06 ± 0.08	0.9391 ± 0.0008	93.87 ± 0.06

BNCI2014_008. BNCI2014_008 (Riccio et al., 2013) contains P300 speller recordings from eight participants with amyotrophic lateral sclerosis, using 8 EEG channels at 256 Hz. Participants attend to a target character while items on the display flash. Responses to flashes of the attended character are targets and the remaining responses are non-targets, forming a binary task. Each 1.0 s window covers 0–1 s after the flash and is bandpass-filtered to 1–24 Hz.

BNCI2014_009. BNCI2014_009 (Aricò et al., 2014) contains P300 speller recordings from ten healthy participants, using 16 EEG channels at 256 Hz and the same flashing paradigm with binary target/non-target labels. Each 0.8 s window covers 0–0.8 s and is bandpass-filtered to 1–24 Hz.

Nakanishi2015. Nakanishi2015 (Nakanishi et al., 2015) contains SSVEP recordings from nine participants with 8 EEG channels at 256 Hz. Participants fixate on a cued target among 12 visual stimuli flickering at distinct frequency–phase combinations. The task is 12-class stimulus-frequency identification. Each 4.15 s window covers 0.15–4.3 s and is bandpass-filtered to 7–45 Hz. Multiclass metrics are reported.

Wang2016. Wang2016 (Wang et al., 2017) contains SSVEP recordings from 34 participants with 62 retained EEG channels at 250 Hz. Participants fixate on one of 40 visual targets encoded by distinct frequency–phase combinations. The task is 40-class target identification. Each 5.0 s window covers 0.5–5.5 s relative to trial onset and is bandpass-filtered to 7–45 Hz. Multiclass metrics are reported.

D.3 SPEECH DECODING

Table 12: Full metrics for text-MEG (S01–S06).

Method	S01 BAcc	S01 κ	S01 Weighted F1	S02 BAcc	S02 κ	S02 Weighted F1
EEGNet	6.44 ± 0.51	0.0445 ± 0.0052	4.28 ± 0.46	8.33 ± 0.41	0.0638 ± 0.0042	5.72 ± 0.31
FBCNet	4.26 ± 0.57	0.0222 ± 0.0058	2.69 ± 0.48	4.72 ± 0.34	0.0270 ± 0.0035	2.60 ± 0.23
IFNet	**10.56** ± 1.08	**0.0865** ± 0.0110	6.28 ± 0.67	13.19 ± 0.79	0.1135 ± 0.0081	8.46 ± 0.55
Conformer	6.71 ± 0.62	0.0473 ± 0.0064	3.97 ± 0.59	8.15 ± 1.15	0.0619 ± 0.0117	4.75 ± 0.65
DBConformer	5.46 ± 0.47	0.0345 ± 0.0048	1.79 ± 0.40	6.57 ± 1.07	0.0459 ± 0.0109	2.97 ± 0.48
Deformer	8.29 ± 1.05	0.0634 ± 0.0107	4.65 ± 0.91	11.06 ± 0.66	0.0917 ± 0.0068	7.28 ± 0.49
PyraNet	10.56 ± 1.09	0.0865 ± 0.0111	**6.72** ± 0.60	**14.68** ± 1.18	**0.1286** ± 0.0120	**10.09** ± 1.11

Method	S03 BAcc	S03 κ	S03 Weighted F1	S04 BAcc	S04 κ	S04 Weighted F1
EEGNet	7.50 ± 0.89	0.0553 ± 0.0090	5.04 ± 0.50	5.28 ± 0.90	0.0326 ± 0.0092	3.65 ± 0.71
FBCNet	5.88 ± 0.54	0.0388 ± 0.0055	3.83 ± 0.25	3.52 ± 0.51	0.0147 ± 0.0052	1.80 ± 0.34
IFNet	**14.44** ± 1.01	**0.1262** ± 0.0103	**10.36** ± 0.99	**9.12** ± 1.35	**0.0719** ± 0.0138	5.58 ± 1.09
Conformer	8.19 ± 0.20	0.0624 ± 0.0020	5.00 ± 0.22	4.77 ± 0.43	0.0274 ± 0.0044	2.96 ± 0.12
DBConformer	8.52 ± 1.02	0.0657 ± 0.0104	4.57 ± 0.71	4.63 ± 0.40	0.0260 ± 0.0041	1.98 ± 0.21
Deformer	10.46 ± 0.74	0.0856 ± 0.0075	6.65 ± 0.46	6.11 ± 0.60	0.0411 ± 0.0061	3.95 ± 0.28
PyraNet	12.69 ± 0.17	0.1083 ± 0.0017	8.95 ± 0.14	8.80 ± 0.85	0.0686 ± 0.0087	**6.03** ± 0.73

Method	S05 BAcc	S05 κ	S05 Weighted F1	S06 BAcc	S06 κ	S06 Weighted F1
EEGNet	6.44 ± 0.43	0.0444 ± 0.0044	4.53 ± 0.38	3.61 ± 0.63	0.0156 ± 0.0064	2.28 ± 0.28
FBCNet	5.00 ± 0.41	0.0298 ± 0.0042	3.02 ± 0.15	3.56 ± 0.62	0.0151 ± 0.0064	1.92 ± 0.47
IFNet	7.82 ± 0.62	0.0586 ± 0.0064	4.76 ± 0.62	**6.20** ± 0.43	**0.0421** ± 0.0044	**3.43** ± 0.03
Conformer	6.85 ± 0.62	0.0487 ± 0.0064	4.41 ± 0.71	5.28 ± 0.57	0.0326 ± 0.0058	2.81 ± 0.37
DBConformer	4.58 ± 0.20	0.0255 ± 0.0020	1.40 ± 0.15	3.61 ± 0.74	0.0156 ± 0.0076	1.11 ± 0.50
Deformer	**9.63** ± 0.36	**0.0771** ± 0.0037	**6.30** ± 0.23	4.95 ± 0.95	0.0293 ± 0.0097	2.41 ± 0.50
PyraNet	8.43 ± 0.36	0.0648 ± 0.0037	5.28 ± 0.25	5.69 ± 0.45	0.0369 ± 0.0047	3.42 ± 0.41

Table 13: Full metrics for text-MEG (S07–S09 and Mean).

Method	S07 BAcc	S07 κ	S07 Weighted F1	S08 BAcc	S08 κ	S08 Weighted F1
EEGNet	4.35 ± 0.74	0.0232 ± 0.0075	2.86 ± 0.60	9.68 ± 0.43	0.0776 ± 0.0044	6.96 ± 0.34
FBCNet	5.60 ± 0.13	0.0359 ± 0.0014	3.49 ± 0.13	5.69 ± 0.74	0.0369 ± 0.0076	3.69 ± 0.76
IFNet	**8.10** ± 0.80	**0.0615** ± 0.0081	**4.83** ± 0.33	13.61 ± 0.49	0.1177 ± 0.0051	9.18 ± 0.53
Conformer	4.40 ± 0.24	0.0237 ± 0.0024	2.40 ± 0.13	9.95 ± 0.57	0.0804 ± 0.0058	6.74 ± 0.33
DBConformer	4.86 ± 0.34	0.0284 ± 0.0035	1.83 ± 0.11	6.94 ± 0.11	0.0496 ± 0.0012	3.28 ± 0.07
Deformer	6.48 ± 1.09	0.0449 ± 0.0111	3.44 ± 0.54	13.10 ± 0.43	0.1125 ± 0.0044	9.28 ± 0.36
PyraNet	7.50 ± 0.52	0.0553 ± 0.0053	4.54 ± 0.21	**16.57** ± 0.40	**0.1480** ± 0.0041	**11.87** ± 0.53

Method	S09 BAcc	S09 κ	S09 Weighted F1	Mean BAcc	Mean κ	Mean Weighted F1
EEGNet	10.19 ± 0.57	0.0828 ± 0.0058	7.61 ± 0.40	6.87 ± 0.17	0.0489 ± 0.0018	4.77 ± 0.09
FBCNet	9.44 ± 0.68	0.0752 ± 0.0069	6.16 ± 0.65	5.30 ± 0.06	0.0328 ± 0.0006	3.24 ± 0.08
IFNet	17.45 ± 1.05	0.1570 ± 0.0108	12.16 ± 0.95	11.17 ± 0.38	0.0928 ± 0.0039	7.23 ± 0.34
Conformer	10.23 ± 0.85	0.0832 ± 0.0087	6.95 ± 0.79	7.17 ± 0.24	0.0520 ± 0.0025	4.44 ± 0.20
DBConformer	8.06 ± 0.60	0.0610 ± 0.0061	3.53 ± 0.37	5.92 ± 0.29	0.0391 ± 0.0030	2.49 ± 0.11
Deformer	14.95 ± 0.82	0.1314 ± 0.0084	9.99 ± 0.56	9.45 ± 0.37	0.0752 ± 0.0038	5.99 ± 0.25
PyraNet	**18.06** ± 0.74	**0.1631** ± 0.0076	**12.78** ± 0.96	**11.44** ± 0.06	**0.0956** ± 0.0006	**7.74** ± 0.20

Table 14: Full metrics for bouchard2019.

Method	EC2 BAcc	EC2 κ	EC2 Weighted F1	EC9 BAcc	EC9 κ	EC9 Weighted F1
EEGNet	38.21 ± 0.49	0.3701 ± 0.0055	37.10 ± 0.83	8.54 ± 1.05	0.0718 ± 0.0079	8.27 ± 0.32
FBCNet	1.93 ± 0.07	0.0008 ± 0.0007	1.00 ± 0.11	2.10 ± 0.17	0.0021 ± 0.0010	1.30 ± 0.25
IFNet	31.28 ± 0.39	0.2994 ± 0.0045	29.69 ± 0.47	4.82 ± 0.21	0.0326 ± 0.0036	4.29 ± 0.70
Conformer	40.64 ± 0.81	0.3923 ± 0.0081	38.79 ± 0.87	8.21 ± 0.71	0.0659 ± 0.0062	6.44 ± 0.81
DBConformer	28.76 ± 1.15	0.2727 ± 0.0121	26.65 ± 1.31	7.08 ± 0.94	0.0594 ± 0.0075	4.83 ± 0.65
Deformer	44.41 ± 0.95	0.4312 ± 0.0090	42.71 ± 1.13	10.65 ± 0.25	0.0884 ± 0.0046	8.48 ± 0.24
PyraNet	**49.58** ± 0.48	**0.4850** ± 0.0054	**48.45** ± 0.32	**12.00** ± 1.21	**0.1031** ± 0.0079	**10.39** ± 0.52

Method	GP31 BAcc	GP31 κ	GP31 Weighted F1	GP33 BAcc	GP33 κ	GP33 Weighted F1
EEGNet	18.34 ± 0.53	0.1661 ± 0.0058	17.07 ± 0.57	11.29 ± 1.05	0.0964 ± 0.0101	10.69 ± 0.96
FBCNet	2.34 ± 0.32	0.0042 ± 0.0033	1.26 ± 0.14	1.66 ± 0.20	−0.0018 ± 0.0020	1.24 ± 0.04
IFNet	9.07 ± 0.16	0.0711 ± 0.0014	7.72 ± 0.37	9.35 ± 0.35	0.0757 ± 0.0039	8.45 ± 0.56
Conformer	16.38 ± 0.50	0.1469 ± 0.0048	14.75 ± 0.46	12.36 ± 0.77	0.1062 ± 0.0071	11.41 ± 0.63
DBConformer	16.05 ± 0.34	0.1431 ± 0.0038	14.19 ± 0.36	6.02 ± 1.36	0.0423 ± 0.0135	4.56 ± 1.03
Deformer	19.20 ± 0.49	0.1754 ± 0.0056	17.12 ± 0.58	20.15 ± 0.13	0.1859 ± 0.0023	18.51 ± 0.67
PyraNet	**22.30** ± 0.27	**0.2063** ± 0.0028	**20.86** ± 0.20	**23.55** ± 1.09	**0.2188** ± 0.0110	**22.03** ± 0.73

Method	Mean BAcc	Mean κ	Mean Weighted F1
EEGNet	19.09 ± 0.29	0.1761 ± 0.0022	18.28 ± 0.13
FBCNet	2.01 ± 0.05	0.0013 ± 0.0005	1.20 ± 0.05
IFNet	13.63 ± 0.02	0.1197 ± 0.0006	12.54 ± 0.16
Conformer	19.40 ± 0.17	0.1778 ± 0.0017	17.85 ± 0.15
DBConformer	14.48 ± 0.35	0.1294 ± 0.0032	12.56 ± 0.22
Deformer	23.60 ± 0.29	0.2202 ± 0.0032	21.70 ± 0.31
PyraNet	**26.86** ± 0.47	**0.2533** ± 0.0036	**25.43** ± 0.23

text-MEG. The text-MEG dataset (Jia et al., 2025) contains recordings from nine participants who read aloud 48 Mandarin words displayed on a screen. Each trial begins with a preparation period, followed by a word presentation and reading phase, then a rest period. Each participant completes 15 blocks containing all 48 words. We used 57–61 MEG channels per subject at 1 kHz, extracted the 70–170 Hz high-gamma envelope, and resampled it to 400 Hz. Each 1.6 s window starts at trial onset. The task is 48-class word classification under leave-one-session-out evaluation within each subject.

bouchard2019. In bouchard2019 (Bouchard & Chang, 2020), four participants (EC2, EC9, GP31, GP33) produce consonant–vowel syllables with annotated syllable identities and speech onsets. We retained 229–249 ECoG channels per subject (native rate 3,051.76 Hz), extracted the 70–170 Hz envelope, and resampled it to 400 Hz. Each 1.0 s window starts at speech onset. The task is 54-class syllable classification under leave-one-session-out evaluation within each subject, with 7/7/14/3 sessions per subject.

D.4 MA State Classification

Table 15: Full metrics for EEGMAT.

Method	EEGMAT BAcc	EEGMAT AUC–PR	EEGMAT AUROC
EEGNet	64.20 ± 1.17	74.01 ± 1.27	71.70 ± 1.61
FBCNet	71.08 ± 3.11	**88.34** ± 0.33	**88.28** ± 0.38
IFNet	**75.19** ± 1.10	<u>87.13</u> ± 0.79	<u>87.12</u> ± 0.76
Conformer	58.21 ± 0.54	72.97 ± 1.73	71.15 ± 1.55
DBConformer	73.83 ± 1.24	85.57 ± 0.67	84.95 ± 0.72
Deformer	61.27 ± 1.68	77.75 ± 0.42	77.06 ± 0.31
CBraMod	71.94 ± 1.37	81.99 ± 1.59	79.53 ± 1.57
LaBraM	68.98 ± 2.05	80.72 ± 0.78	79.26 ± 1.25
REVE	<u>73.83</u> ± 0.51	83.64 ± 0.30	83.82 ± 0.22
PyraNet	71.20 ± 0.80	79.99 ± 0.41	80.05 ± 0.20

EEGMAT. In EEGMAT (Zyma et al., 2019), thirty-six participants either rest with their eyes closed or perform serial mental subtraction. We retained 19 EEG channels at 500 Hz, took the final 60 s of the resting recording and the first 60 s of the arithmetic recording, and divided them into non-overlapping 4 s windows without additional filtering. The task is binary rest/arithmetic classification of each window.

E Parameters and FLOPs Comparison

We compared the nine baselines and PyraNet using BNCI2014_001 as an example (Table 3). Each input trial contains 22 channels and 1,001 samples at 250 Hz. LaBraM, CBraMod, and REVE received the same trial resampled to 200 Hz. LaBraM and CBraMod used 800 samples internally, while REVE retained the complete resampled trial. Parameter counts include weights and biases but exclude buffers. Floating-point operations (FLOPs) count convolution, linear projection, and attention matrix products for one trial, with one multiply–accumulate counted as two operations. Elementwise operations, pooling, and non-neural preprocessing or filtering are excluded. With grid 8, PyraNet has 5,188 parameters, the second-smallest total after EEGNet. It uses fewer parameters than IFNet and FBCNet while jointly processing grouped spatial features over multiple temporal receptive fields. Its approximately 7.95 million FLOPs are the lowest among the evaluated models under this counting convention, indicating the lowest theoretical computational cost during inference. Increasing the grid mainly enlarges the classification head. Its parameters grow from 1,028 to 16,388, whereas total FLOPs increase by less than 0.4%. The selector therefore has a substantial effect on retained parameter count even when its effect on forward arithmetic is small. Joint warm-up temporarily retains all three heads and totals 25,676 parameters, while these costs are excluded from the single-branch inference table.

F Interpretability Analysis

F.1 Channel Importance Across Paradigms

We analyzed the spatial filter weights of the seed 42 LOSO models on the six EEG datasets BNCI2014_001, BNCI2015_001, BNCI2014_008, BNCI2014_009, Nakanishi2015, and Wang2016. For each recording channel, we computed the root mean square across normalized filters, divided by the sum across channels, and averaged the resulting relative usage across LOSO folds. Figure 7 shows the channel importance across datasets. MI emphasized C3 and C4, consistent with sensorimotor recordings (Pfurtscheller & Neuper, 2001). On Nakanishi2015, SSVEP placed the largest weight on Oz, consistent with visual responses (Norcia et al., 2015). P300 emphasized posterior channels, particularly PO7 and PO8, consistent with the posterior distribution of

the P300 response (Polich, 2007). The weights therefore show task-dependent channel preferences, and PyraNet emphasizes channels associated with each paradigm.

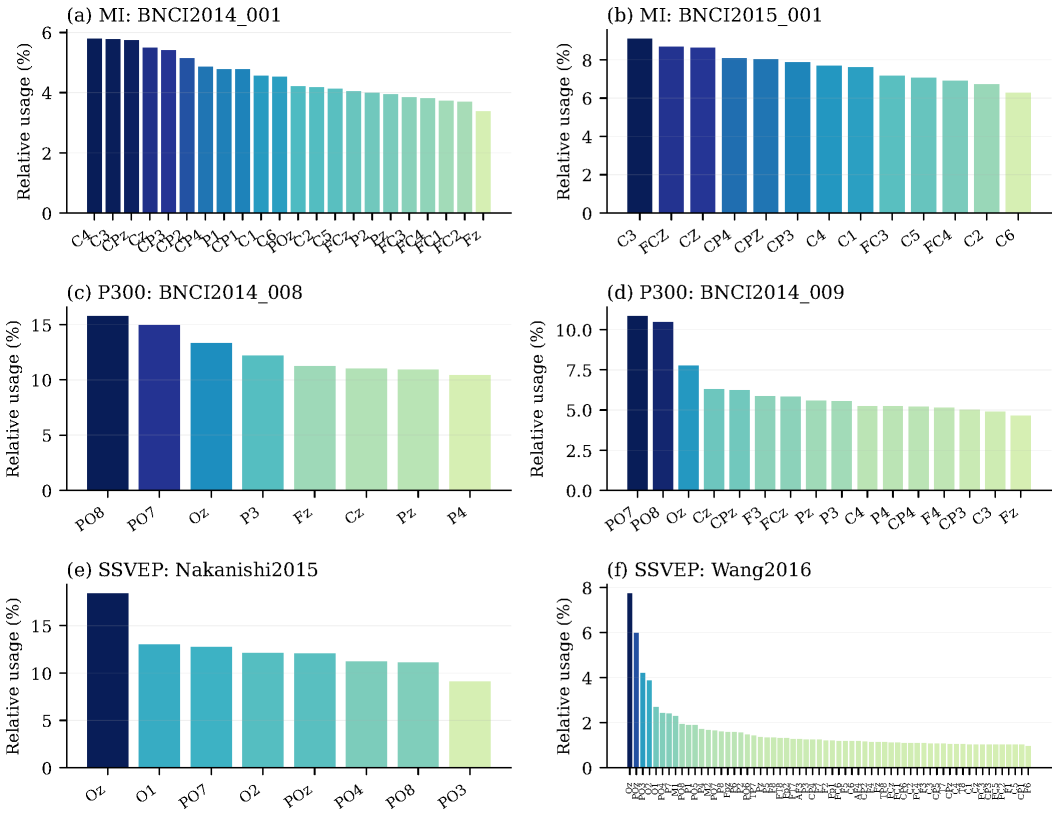

Figure 7: Channel importance of electrodes on different paradigms.

F.2 Spectral Analysis of Multi-Scale Convolutions

We analyzed the temporal branch activations of the seed 42 LOSO models on the six EEG datasets, with subject 2 held out for each dataset, to examine the frequency characteristics captured by the multi-scale convolutions. For each held-out trial, we recorded each temporal branch's output after convolution and BN but before GELU. Each activation time series was demeaned, transformed with a real-valued discrete Fourier transform, and converted to squared-magnitude power. We normalized each series by its total spectral power, then averaged over trials and feature channels. Figure 8 shows the averaged branch spectra for each dataset. The branches retained similar broad spectral shapes but emphasized different frequency ranges. MI differences appeared near the dominant rhythm and at higher frequencies, P300 differences mainly at low frequencies, and SSVEP differences around stimulus-related and higher-frequency components. This supports partial spectral diversity across scales, with substantial overlap.

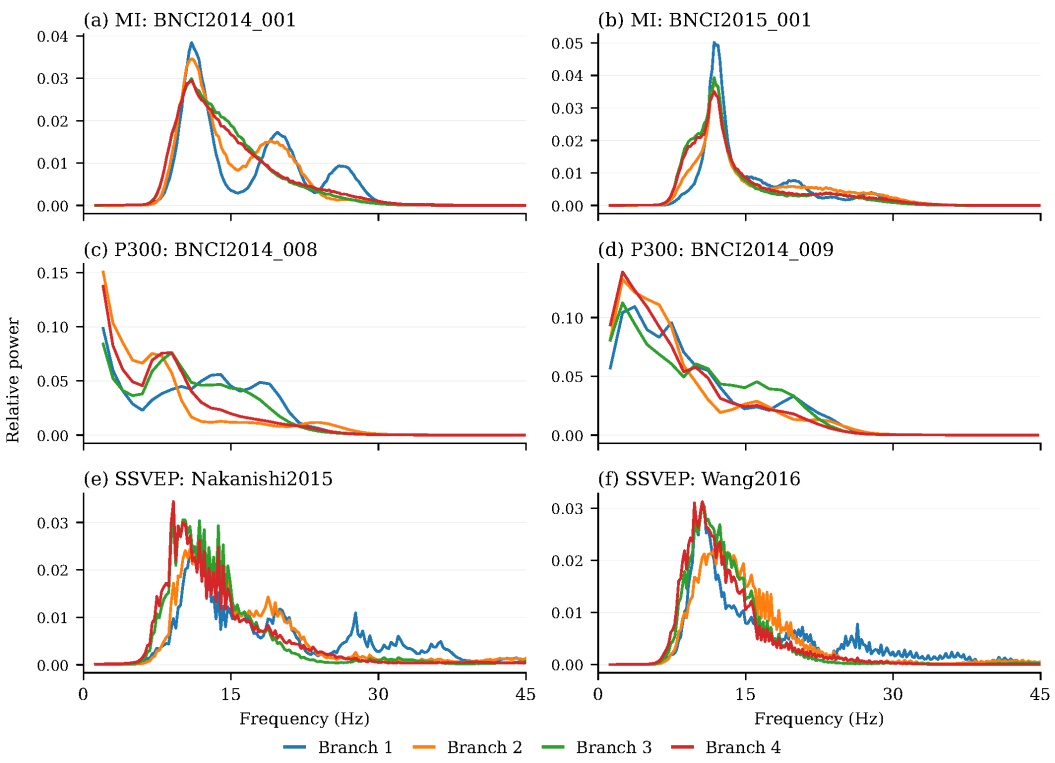

Figure 8: Relative activation power spectra of the four temporal branches across datasets.

F.3 Representation Analysis of the Grid Selector

We analyzed the grid selector on Nakanishi2015 using the seed 42 LOSO model with subject 2 held out. From the same encoder and the same 180 held-out trials, we obtained the pre-classifier representations by average pooling to grids 8 and 128 and flattening. Their dimensions were 256 and 4,096, respectively. For visualization only, each feature dimension was standardized, principal component analysis reduced the representation to 50 dimensions, and t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten & Hinton, 2008) projected it to two dimensions with perplexity 30 and random seed 42. Test labels were used only to color points and describe class separation. As shown in Figure 4, grid 128 showed clearer class separation, indicating that finer temporal granularity benefits SSVEP classification.

G Warm-up Epoch Parameter Sensitivity Analysis

We examine how the warm-up epoch parameter e_w affects decoding accuracy. Starting from the default $e_w = 10$, we vary $e_w \in \{1, 4, 20\}$ and evaluate on four representative datasets—BNCI2014_001 (MI), BNCI2014_002 (MI), BNCI2014_009 (P300), and Nakanishi2015 (SSVEP). Figure 9 reports mean BAcc. On the three MI and P300 datasets, BAcc changes by less than half a percentage point across all tested warm-up lengths. On Nakanishi2015, however, accuracy is more sensitive. Reducing e_w to 1 or 4 lowers BAcc to 49.01% and 59.65%, respectively. These results indicate that the default ten warm-up epochs are sufficient for stable performance across paradigms, and that the risk of under-warming is concentrated on datasets with rapid temporal dynamics such as SSVEP.

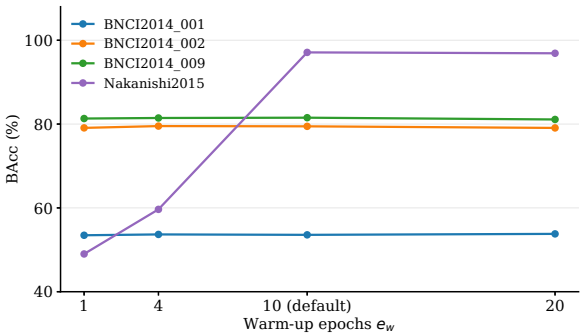

Figure 9: Warm-up epoch sensitivity on four datasets. Lines connect mean BAcc values across three random seeds.

H Discussion

H.1 Implications

Our proposed PyraNet shows that strong generalizability does not necessarily require a large model. PyraNet achieves competitive performance across paradigms, datasets, and recording modalities with a lightweight architecture informed by neurophysiological principles. These findings suggest that model development should focus on how effectively parameters are used, alongside how many parameters a model contains.

Future brain foundation models should be grounded in the physiology and structure of brain signals. Their architectures and pretraining objectives should reflect the spatial, spectral, and temporal characteristics of neural activity. Pretraining objectives should be chosen to help models learn useful features of brain signals, rather than simply borrowed from fields where they have already proved successful. The aim should be to use model capacity effectively to learn shared neural features that support a broad range of decoding tasks.

H.2 Limitations

Our work still has some limitations. PyraNet effectively extracts discriminative features, but it does not yet support transferring these features across datasets or tasks. This limits further performance gains and its potential for zero-shot decoding.

H.3 Future Work

Given these limitations, future work will combine small and large models. PyraNet will serve as the small model to extract general spatial, temporal, and spectral features. A larger model will then learn from these features using continuous neural recordings and pretraining objectives guided by neurophysiological principles. This approach aims to capture how neural activity evolves across paradigms, support transfer learning across datasets and tasks, and enable zero-shot decoding.