NEUROMI: PROGRESSIVE EEG FOUNDATION MODELING FOR NEURAL DECODING OF MOTOR INTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Motor imagery (MI) provides a non-invasive paradigm for decoding motor intention-related neural states through brain-computer interfaces (BCIs). However, the practical deployment of MI-based BCIs remains constrained by data scarcity, substantial inter-subject variability, and poor cross-dataset generalization. Existing methods either train task-specific models from scratch on small MI datasets, which compromises robustness, or directly fine-tune generic electroencephalography (EEG) pre-trained models, resulting in a significant domain gap between generic representations and MI-specific neural signatures. This paper proposes a hierarchical three-stage foundation model framework, NeuroMI, to systematically address this gap. In the first stage, a general-purpose EEG encoder is designed and pre-trained on large-scale EEG corpora via self-supervised learning, yielding transferable spatio-temporal-spectral representations. In the second stage, we build a multi-dataset MI-specific EEG corpus (approximately 300,000 samples) and specialize the encoder on it with neurophysiologically motivated objectives, explicitly steering the model toward motor-relevant neural dynamics. In the third stage, the specialized encoder is adapted to downstream MI tasks using lightweight task-specific heads, enabling efficient transfer across subjects and datasets. Experiments on multiple public MI datasets demonstrate that the proposed framework consistently outperforms both generic EEG foundation models and conventional MI-specific architectures. Beyond achieving improved decoding accuracy, the model yields interpretable representations that align with established MI neurophysiology. These results provide a practical pathway toward robust MI decoding in real-world BCI applications.

1 INTRODUCTION

Motor intention decoding, which aims to infer a user's intended movement directly from neural activity, is a primary goal of non-invasive brain-computer interfaces (BCIs) and holds promise for assistive control, neurorehabilitation, and human-machine interaction, particularly for individuals with motor impairments. Because overt movement is often unavailable or impractical in these populations, motor imagery (MI) has become the most widely adopted paradigm for eliciting and studying motor intention-related neural states. MI, the mental rehearsal of a movement without actual execution, allows a user to voluntarily modulate neural activity in a manner closely resembling movement preparation and execution, and therefore serves as a practical surrogate for decoding motor intention.

Electroencephalography (EEG) is widely adopted for MI decoding because of its non-invasiveness, high temporal resolution, and practical acquisition cost. MI modulates activity in the μ and β bands of the sensorimotor cortex and is accompanied by neurophysiological phenomena such as event-related desynchronization (ERD), providing an important basis for inferring MI from EEG signals (Pfurtscheller & Lopes da Silva, 1999). Traditional MI decoding methods typically rely on handcrafted features such as common spatial pattern and band power, followed by classifiers such as linear discriminant analysis or support vector machines (Blankertz et al., 2008; Ang et al., 2008). These methods are relatively interpretable, but their performance depends on preset frequency ranges, spatial filters, and feature-extraction pipelines. In recent years, deep learning has become the dominant approach to MI-EEG decoding, with architectures designed to capture complementary spectral, spatial, and temporal patterns. Representative convolutional neural networks (CNNs) include ShallowCNN, DeepCNN, EEGNet, and FBCNet (Schirrmeister et al., 2017; Lawhern et al., 2018; Mane

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.

et al., 2021). CNN-sequence or attention hybrids, such as EEG-TCNet, EEG-ITNet, and ATCNet, further incorporate temporal convolution or attention to model long-range dynamics (Ingolfsson et al., 2020; Salami et al., 2022; Altaheri et al., 2023). Other approaches employ Transformers, exemplified by EEG Conformer (Song et al., 2023), or graph neural networks, such as GCNs-Net (Hou et al., 2024), to capture global dependencies or inter-electrode relationships. Despite these architectural advances, most conventional MI-specific decoders are trained with supervised classification objectives on individual, relatively small datasets. Their representations can therefore become entangled with subject-, session-, and dataset-specific variations. Consequently, although such models can achieve competitive accuracy when evaluated on the same dataset or within-subject protocols, their performance typically degrades substantially when transferred to new subjects, sessions, or datasets due to inter-subject variability, which raises a fundamental question of **how to learn MI representations that generalize beyond the training distribution.**

Large-scale self-supervised pretraining has recently emerged as a promising alternative for learning transferable EEG representations. BENDR pioneered contrastive pretraining on large unlabeled EEG corpora (Kostas et al., 2021), while BIOT addressed variations in channel configuration, signal length, and missing values during cross-dataset learning (Yang et al., 2023). Subsequent EEG foundation models, including LaBraM, EEGPT, CBraMod, and REVE, scaled masked modeling or representation alignment to increasingly large and diverse EEG corpora and demonstrated transfer across clinical and BCI tasks, including MI decoding (Jiang et al., 2024; Wang et al., 2024; 2025a; El Ouahidi et al., 2025). However, despite their broad success in general EEG tasks, directly applying or fine-tuning these generic foundation models for MI decoding often yields sub-optimal performance (Liu et al., 2026a), frequently failing to surpass conventional task-specific models (Figure 1(a)). We attribute this gap to two fundamental limitations. **First, MI-specific data scarcity.** Compared with the large-scale, multi-paradigm EEG corpora used to pretrain general foundation models, publicly available MI datasets are orders of magnitude smaller, typically on the order of tens of thousands of trials, far below the volume required to train a dedicated MI foundation model from scratch. Directly fine-tuning a generic EEG model on such limited data does not close the domain gap between general EEG representations and MI-specific neural signatures. **Second, lack of MI-oriented pretraining objectives.** Most existing pretraining objectives remain task-agnostic, such as reconstruction, contextual prediction, or generic representation alignment, and therefore do not explicitly require the learned representations to emphasize the oscillatory and event-related patterns underlying MI (e.g., μ/β rhythm modulation and ERD dynamics). Even paradigm-specific classification supervision may exploit dataset-dependent discriminative cues without physiologically grounding the representation. In other words, current foundation models acquire broad EEG knowledge but are never explicitly guided to attend to the motor-relevant neural dynamics that distinguish MI from other paradigms. These observations motivate a key question: *Can we combine the data scale of general EEG pretraining with the physiological specificity of MI-oriented objectives to build a foundation model that is both broadly transferable and MI-attuned?*

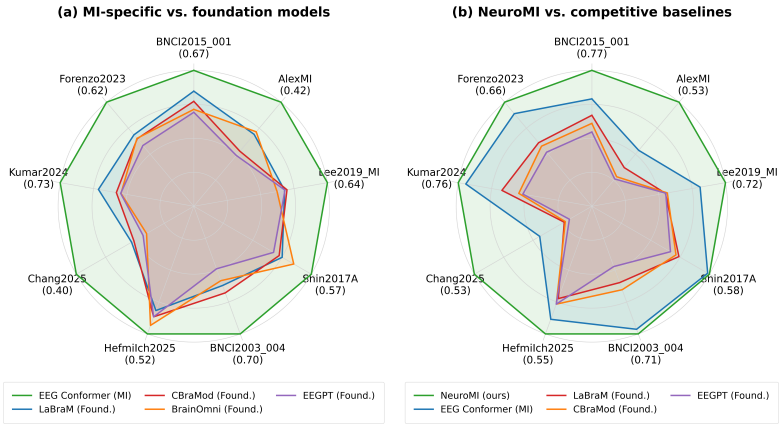


Figure 1: Visualization of results on nine MI downstream datasets. (a) Comparison of EEG Conformer with EEG foundation models; (b) comparison of NeuroMI (ours) with competitive MI-specific and foundation baselines.

To address this challenge, we propose NeuroMI, a foundation model designed for transferable MI EEG decoding. NeuroMI follows a two-stage hierarchical pretraining strategy that progressively transfers knowledge from large-scale EEG to the MI paradigm. **The first stage mitigates MI data scarcity** by pretraining on large-scale EEG recordings using signal reconstruction and frequency-band reconstruction objectives. This stage enables the encoder to learn general waveform structures and spectral regularities from substantially more diverse EEG data than are available in typical MI datasets. **The second stage closes the domain gap** by further pretraining on a large collection of MI recordings to align the general EEG representation with the target paradigm. We assign a lower learning rate to the lower layers of the pretrained Transformer to retain general EEG knowledge while adapting the remaining layers using a joint objective that includes μ and β rhythmic-energy reconstruction and ERD reconstruction. Unlike generic reconstruction or classification alone, these auxiliary objectives explicitly encourage the representation to retain frequency-specific sensorimotor activity and its event-related modulation, thereby **embedding MI neurophysiological priors directly into the pretraining process.** This progressive design separates general EEG representation learning from MI-specific physiological adaptation, reducing reliance on any single downstream dataset.

To support this progressive pretraining and effectively represent the multi-scale, multi-band, and spatially coupled nature of MI-EEG signals, NeuroMI further introduces a hierarchical encoding architecture tailored to the multi-scale and coupled nature of EEG signals. It integrates a multi-scale CNN patch embedding, frequency-aware positional encoding with injected spectral priors, and spatial-temporal fusion Transformers. The patch embedding represents EEG patterns at complementary temporal resolutions, while the frequency-aware encoding provides the model with an explicit reference to oscillatory structure. The fusion Transformer then models interactions between temporal dynamics and distributed spatial activity instead of treating these dimensions independently. Together, these components enable NeuroMI to capture multi-band oscillations and spatial-temporal brain coupling patterns while maintaining transferable representations across subjects and datasets. Importantly, **the architecture and physiological pretraining objectives serve complementary roles:** the former determines how heterogeneous EEG structures are represented, whereas the latter constrains which MI-relevant information should be preserved.

Our main contributions are summarized as follows:

1. We propose NeuroMI, an EEG foundation model for MI decoding that progressively transfers representations from large-scale EEG to the MI paradigm. By leveraging a curated large-scale MI corpus and a two-stage mechanism-guided pre-training strategy, it effectively overcomes the data scarcity bottleneck and the performance ceiling of generic foundation models in MI tasks.
2. We develop a hierarchical EEG encoder that combines multi-scale CNN patch embedding, frequency-aware positional encoding, and spatial-temporal fusion Transformers, enabling joint modeling of multi-band oscillatory activity and distributed brain coupling patterns.
3. We conduct a unified evaluation on nine downstream MI datasets against both task-specific models and general EEG foundation models. NeuroMI consistently achieves the best performance across the evaluated datasets and metrics, supporting both its decoding superiority and generalization capability.

2 METHOD

We propose **NeuroMI**, a progressive EEG foundation modeling framework for neural decoding of MI. The central motivation is that effective transferable MI representation requires ***bridging general-purpose EEG knowledge*** and ***MI-specific neurophysiological mechanisms.*** NeuroMI progressively specializes a general-purpose EEG encoder through three stages, as illustrated in Figure 2. In Stage 1, a generic EEG encoder is designed and pretrained on large-scale EEG corpora to learn general spatio-temporal-spectral representations. In Stage 2, the pretrained encoder is further adapted on a curated multi-dataset MI-specific EEG corpus with mechanism-guided objectives to inject sensorimotor neurophysiological priors into the backbone. In Stage 3, the adapted encoder is transferred to downstream MI decoding tasks with a lightweight head, enabling efficient adaptation across subjects and datasets.

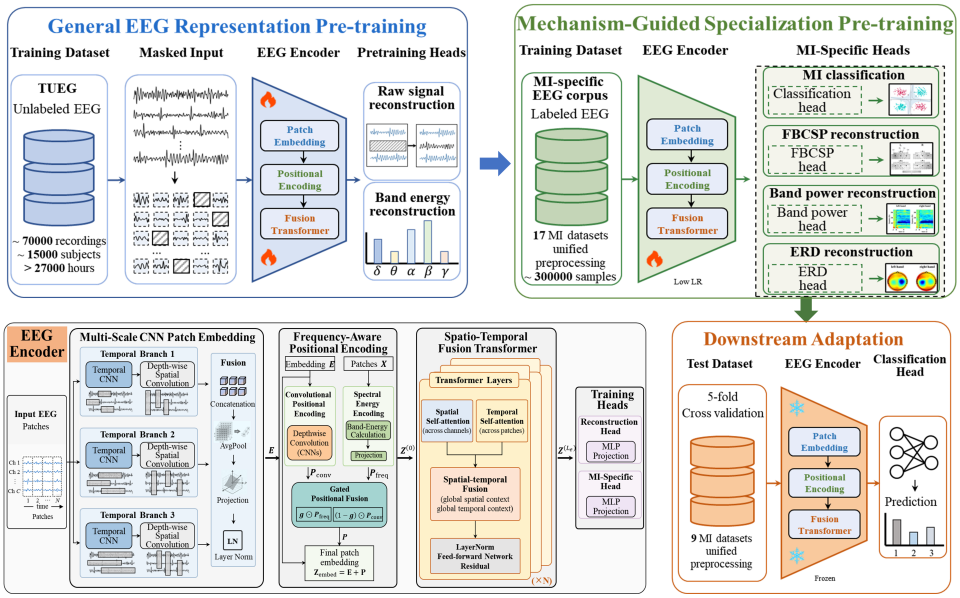


Figure 2: Architecture of the proposed NeuroMI.

2.1 SENSORIMOTOR-AWARE EEG ENCODER

Multi-Scale CNN Patch Embedding. MI EEG signals contain discriminative information across multiple temporal scales. To capture this multi-scale structure, we design a **multi-scale temporal CNN** that independently processes each patch before Transformer encoding. Given an EEG segment $X \in \mathbb{R}^{C\times L}$, we divide it into $P = \lfloor L/T \rfloor$ non-overlapping patches of length T, obtaining $X \in \mathbb{R}^{C\times P\times T}$. Each patch is processed by K parallel temporal convolution branches with different kernel sizes, followed by depth-wise spatial convolution across EEG channels. The branch outputs are concatenated and projected into a d-dimensional embedding:

$$E = LN\left(W_{proj}\, Concat_{i=1}^{K}\left[DWConv_{spatial}\left(Conv_{k_i}\left(X\right)\right)\right]\right). \tag{1}$$

This design captures temporal patterns at multiple scales.

Frequency-Aware Positional Encoding. EEG patches carry rich spectral information that varies dramatically with the neurophysiological context, especially for MI decoding. We therefore propose a frequency-aware positional encoding that injects band-specific spectral content into each patch embedding. For each patch, its **spectral energy** in N_f canonical frequency bands is computed and projected into the model space:

$$p_{freq} = W_{freq}\, s_{b,c,p}, \tag{2}$$

where $s_{b,c,p} \in \mathbb{R}^{N_f}$ denotes the band-power vector. In parallel, a **convolutional positional encoding** p_{conv} captures local spatial-temporal structure. The two components are combined via a learned gating mechanism:

$$g = \sigma\left(W_{gate}\left[p_{conv}; p_{freq}\right]\right), \qquad P = g \odot p_{freq} + (1-g)\odot p_{conv}, \tag{3}$$

where the g denotes the gate weight to ensure both encodings contribute meaningfully. The final patch embedding is $Z_{embed} = E + P$. This design ensures that each token's positional information reflects its actual spectral content, a critical property for MI, where the same spatial location can convey markedly different information depending on its frequency content.

Spatio-Temporal Fusion Transformer. EEG signals exhibit two fundamentally different types of dependencies: **spatial** (across electrode channels) and **temporal** (across time windows). This is important for MI decoding because sensorimotor patterns are not restricted to an isolated channel or a single time point, but are reflected in coordinated changes across the sensorimotor network. We therefore adopt a **spatial-temporal attention** design that decouples spatial and temporal attention, followed by a cross fusion that re-integrates global information. The d-dimensional features are split into two halves: Z_s and Z_t. Then spatial self-attention is computed across the C channels: $Z_{s'} =$

$MHSA_s(Z_s)$. For each channel c, temporal self-attention is computed across the P patches: $Z_{t'} = MHSA_t(Z_t)$. Both attention outputs are concatenated back: $Z_{attn} = [Z_{s'}; Z_{t'}]$, and a residual connection is applied: $Z_{st} = Z + Dropout(Z_{attn})$. While spatial-temporal attention processes spatial and temporal axes independently, MI decoding requires joint spatio-temporal reasoning. We introduce a lightweight cross fusion that exchanges global information between the two branches: $z_s^{global} = Mean_C(Z_{s'})$ and $z_t^{global} = Mean_P(Z_{t'})$. The spatial branch is enhanced with global temporal context (and vice versa):

$$Z_s^{enh} = Z_{s'} + \widetilde{z_t^{global}}, \qquad Z_t^{enh} = Z_{t'} + \widetilde{z_s^{global}}, \tag{4}$$

where $\widetilde{\cdot}$ denotes broadcasting to the appropriate shape. The enhanced branches are concatenated and combined with the input via a learnable residual gate:

$$Z_{fus} = Z_{st} + \alpha \odot \left[Z_s^{enh}; Z_t^{enh}\right], \tag{5}$$

and followed by a pre-norm feed-forward network: $Z_{out} = Z_{fus} + FFN(LN(Z_{fus}))$. The encoder output serves as a transferable representation.

2.2 General EEG Representation Pre-training

MI datasets are typically limited in size and exhibit substantial inter-subject and inter-dataset variability. Training a high-capacity model directly on MI data can easily result in overfitting to subject-specific or dataset-specific characteristics. We therefore first pre-train the EEG encoder on the large-scale and heterogeneous EEG corpus. The aim of this stage is not to model MI-specific activity, but to learn general EEG representations that capture transferable spatio-temporal-spectral structures.

This stage (Stage 1) employs two complementary self-supervised objectives. The **raw signal reconstruction objective** $\mathcal{L}_{raw}$ encourages the model to reconstruct the original EEG signal at masked patches, and forces the encoder to learn a representation sufficient to recover the temporal waveform. Beyond raw signal recovery, we supervise the model to reconstruct the band energy across canonical frequency bands. This **band energy objective** $\mathcal{L}_{energy}$ is important for MI. The spectral content in the μ and β bands carries the discriminative information for MI, and explicitly supervising band energy ensures the encoder develops spectral awareness that transfers to downstream MI tasks. The **total Stage 1 loss is**:

$$\mathcal{L}_{Stage1} = \mathcal{L}_{raw} + \lambda_{energy}\,\mathcal{L}_{energy}, \tag{6}$$

where λ_{energy} is a weighting coefficient.

2.3 Mechanism-Guided Specialization Pre-training

While Stage 1 provides a strong general-purpose EEG backbone, it does not specifically encode the neurophysiological mechanisms underlying MI. To bridge this gap, we build a **dedicated multi-dataset MI-specific EEG corpus** for Stage 2 specialization. MI decoding relies on several well-established phenomena, including the μ **and** β **rhythm power modulation**, the **ERD**, and the **spatial filtering patterns**. We design Stage 2 to explicitly inject these mechanisms as auxiliary supervision.

The Stage 1 pre-trained weights are loaded into the backbone, and new MI-specific heads are initialized. The backbone is adapted with **differential learning rates** to balance knowledge retention and MI specialization. The smaller learning rates for the backbone prevent forgetting of the general EEG representations learned in Stage 1, while the full learning rate for new heads allows rapid MI specialization. First, the **log band power in the** μ **and** β **bands** serves as a regression target. This **band power loss** $\mathcal{L}_{\mu\beta}$ ensures the encoder's representations encode the absolute spectral power in the two most MI-relevant frequency bands. Second, ERD measures the relative power change from a baseline period to the active imagination period. Therefore, we design the ERD as a regression target. The **ERD loss** $\mathcal{L}_{ERD}$ forces the model to not only encode absolute band power but also its dynamics relative to baseline, the core neurophysiological signature of MI. Third, filter bank common spatial pattern (FBCSP) captures discriminative spatial filtering patterns across multiple frequency bands. We train the model to reconstruct FBCSP features from the representation. This **FBCSP objective** $\mathcal{L}_{fbcsp}$ distills the knowledge of classical spatial filtering into the Transformer's representations, ensuring that the learned features align with established MI discriminative patterns. Also, the primary task objective $\mathcal{L}_{cls}$ is standard cross-entropy classification. The **total Stage 2 loss**

is:

$$\mathcal{L}_{Stage2} = \mathcal{L}_{cls} + \lambda_{\mu_\beta}\mathcal{L}_{\mu_\beta} + \lambda_{ERD}\mathcal{L}_{ERD} + \lambda_{fbcsp}\mathcal{L}_{fbcsp}. \tag{7}$$

The weight coefficients $\{\lambda\}$ control the relative influence of each mechanism-guided objective.

After Stage 2, the model has been specialized for MI through both data-driven classification and mechanism-guided auxiliary objectives. In Stage 3, the adapted encoder is transferred to downstream MI decoding tasks. During downstream adaptation, only the classification head is updated, while the backbone remains frozen, preserving the general and MI-specific knowledge encoded in Stages 1-2.

3 Experimental Setup

3.1 Pre-training

Dataset and Preprocessing. The **first stage** pretraining uses the Temple University Hospital EEG Corpus (TUEG) (Obeid & Picone, 2016). TUEG is a large-scale EEG database containing diverse EEG signals from different subjects, acquisition devices, and recording settings, providing highly heterogeneous EEG samples for learning general-purpose EEG representations. We selected 15 channels commonly used in MI studies (F7, F3, Fz, F4, F8, T7, C3, Cz, C4, T8, T5, P3, Pz, P4, and T6) and resampled the signals to 200 Hz. A 1-40 Hz band-pass filter and a 60 Hz notch filter were applied, the signal amplitudes were normalized to [-1, 1], and the data were then segmented into 30-s non-overlapping samples. For the **second stage**, we curate a mixed MI training corpus by aggregating multiple public MI datasets. This corpus currently contains 290,673 four-second training samples. Since electrode layouts and label definitions vary across datasets, we apply a unified preprocessing pipeline to standardize the 15-channel template, sampling rate, and duration. A detailed description of the datasets is provided in Appendix C.

Training Settings. The EEG encoder mainly consists of a multi-scale CNN patch embedding module and a fusion Transformer. The patch embedding module uses temporal convolution kernels sizes of 15, 31, and 63. The fusion Transformer is composed of 12 stacked layers, with a hidden dimension of 200, a feed-forward dimension of 800, and 8 attention heads. The first stage uses a batch size of 64 and an initial learning rate of 1×10^{-4}. Masking is independently applied with probability 0.5. The weight λ_{energy} is empirically set to 0.1. The second stage is initialized from the first-stage checkpoint and uses a batch size of 128. The learning rates for the patch embedding module and fusion Transformer are 5×10^{-5} and 2×10^{-5}, respectively, while the task head uses a learning rate of 1×10^{-4}. The weights $\lambda_{\mu\beta}$, λ_{ERD}, and λ_{fbcsp} are empirically set to 0.1, 0.1, and 0.1, respectively. Both stages are trained for 50 epochs using the AdamW optimizer. Details are provided in the Appendix D.

3.2 Downstream Tasks

Datasets. We conduct downstream evaluation on nine public MI EEG datasets, including **Chang2025** (Chang et al., 2025), **Forenzo2023** (Forenzo et al., 2024), **Lee2019_MI** (Lee et al., 2019), **Shin2017A** (Shin et al., 2017), **BNCI2003_004** (Dornhege et al., 2004), **HefmiIch2025** (Shi et al., 2025), **BNCI2015_001** (Faller et al., 2012), **Kumar2024** (Kumar et al., 2024), and **AlexMI** (Barachant, 2012). The downstream data are uniformly mapped to the 15-channel template, resampled to 200 Hz, and band-pass filtered from 1 to 40 Hz. Each trial is standardized to 4 s. The channel template is kept consistent with Stage 2 pretraining MI data. Detailed dataset descriptions are provided in the Appendix C.

Baselines. Task-specific baselines include **ADFCNN** (Tao et al., 2024), **FBCNet** (Mane et al., 2021), **EEGNet** (Lawhern et al., 2018), **EEG Conformer** (Song et al., 2023), **DeepCNN** (Schirrmeister et al., 2017), **ShallowCNN** (Schirrmeister et al., 2017), **IFNet** (Wang et al., 2023), and **EDPNet** (Han et al., 2024). Foundation-model baselines include **BIOT** (Yang et al., 2023), **BrainOmni** (Xiao et al., 2025), **LaBraM** (Jiang et al., 2024), **EEGPT** (Wang et al., 2024), **CBraMod** (Wang et al., 2025a), and **MIRepNet** (Liu et al., 2026b). All foundation models in our comparison are evaluated via **linear probing**, meaning that the pretrained encoders are frozen and only a lightweight classifier is fine-tuned on the downstream data.

Evaluation Settings and Metrics. All experiments use five-fold cross-validation. For each subject, data are split into five folds. One fold is held out as the test set in turn, and the remaining four folds are divided into training (80%) and validation (20%). The linear probe is trained with cross-entropy loss using the AdamW optimizer for 50 epochs, with a batch size of 64, a learning rate of 5×10^{-5}. The classifier is selected according to accuracy on the validation set and then evaluated on the test fold. For metrics, we report Balanced Accuracy (BA), Cohen's Kappa, and Weighted F1. Details are provided in the Appendix D.

Table 1: Performance comparison on the MI decoding (mean ± standard deviation).

Method	Chang2025			Forenzo2023			Lee2019_MI		
	BA	Kappa	F1	BA	Kappa	F1	BA	Kappa	F1
ADFCNN	0.343 ± 0.029	0.015 ± 0.044	0.227 ± 0.040	0.520 ± 0.032	0.040 ± 0.064	0.434 ± 0.047	0.523 ± 0.040	0.045 ± 0.081	0.427 ± 0.052
FBCNet	0.362 ± 0.065	0.044 ± 0.097	0.292 ± 0.075	0.587 ± 0.082	0.174 ± 0.164	0.554 ± 0.099	0.587 ± 0.102	0.173 ± 0.204	0.538 ± 0.123
EEGNet	0.332 ± 0.035	-0.002 ± 0.052	0.305 ± 0.042	0.516 ± 0.029	0.033 ± 0.059	0.493 ± 0.034	0.514 ± 0.052	0.028 ± 0.103	0.486 ± 0.053
EEG Conformer	0.402 ± 0.093	0.103 ± 0.139	0.356 ± 0.110	0.622 ± 0.104	0.245 ± 0.209	0.605 ± 0.118	0.637 ± 0.128	0.273 ± 0.257	0.612 ± 0.143
DeepCNN	0.346 ± 0.036	0.019 ± 0.054	0.261 ± 0.037	0.507 ± 0.038	0.015 ± 0.077	0.437 ± 0.059	0.513 ± 0.044	0.027 ± 0.088	0.449 ± 0.058
ShallowCNN	0.391 ± 0.096	0.086 ± 0.144	0.377 ± 0.096	0.592 ± 0.095	0.184 ± 0.189	0.583 ± 0.099	0.600 ± 0.120	0.200 ± 0.240	0.587 ± 0.126
IFNet	0.409 ± 0.113	0.114 ± 0.170	0.370 ± 0.130	0.626 ± 0.095	0.251 ± 0.191	0.609 ± 0.106	0.602 ± 0.117	0.204 ± 0.234	0.555 ± 0.138
EDPNet	0.337 ± 0.045	0.005 ± 0.067	0.315 ± 0.045	0.539 ± 0.048	0.078 ± 0.096	0.522 ± 0.053	0.555 ± 0.120	0.111 ± 0.241	0.540 ± 0.127
BIOT	0.353 ± 0.051	0.029 ± 0.076	0.310 ± 0.062	0.525 ± 0.047	0.050 ± 0.093	0.504 ± 0.054	0.530 ± 0.066	0.059 ± 0.132	0.508 ± 0.073
BrainOmni	0.341 ± 0.045	0.012 ± 0.068	0.266 ± 0.055	0.510 ± 0.042	0.020 ± 0.083	0.473 ± 0.054	0.510 ± 0.058	0.019 ± 0.117	0.480 ± 0.068
LaBraM	0.354 ± 0.045	0.031 ± 0.067	0.309 ± 0.048	0.521 ± 0.047	0.042 ± 0.094	0.474 ± 0.063	0.529 ± 0.090	0.059 ± 0.180	0.488 ± 0.106
EEGPT	0.344 ± 0.044	0.017 ± 0.067	0.301 ± 0.055	0.488 ± 0.037	-0.024 ± 0.074	0.451 ± 0.051	0.530 ± 0.025	0.060 ± 0.049	0.508 ± 0.026
CBraMod	0.352 ± 0.053	0.028 ± 0.079	0.288 ± 0.069	0.509 ± 0.052	0.018 ± 0.104	0.458 ± 0.073	0.535 ± 0.052	0.069 ± 0.104	0.486 ± 0.069
MIRepNet	0.353 ± 0.045	0.029 ± 0.068	0.306 ± 0.063	0.532 ± 0.061	0.065 ± 0.122	0.501 ± 0.077	0.572 ± 0.102	0.144 ± 0.204	0.538 ± 0.120
NeuroMI (ours)	**0.530 ± 0.124**	**0.296 ± 0.186**	**0.514 ± 0.132**	**0.662 ± 0.110**	**0.324 ± 0.221**	**0.654 ± 0.119**	**0.716 ± 0.144**	**0.431 ± 0.288**	**0.708 ± 0.150**
Method	**Shin2017A**			**BNCI2003_004**			**HefmiIch2025**		
	BA	**Kappa**	**F1**	**BA**	**Kappa**	**F1**	**BA**	**Kappa**	**F1**
ADFCNN	0.508 ± 0.037	0.015 ± 0.075	0.409 ± 0.051	0.535 ± 0.102	0.071 ± 0.204	0.457 ± 0.118	0.497 ± 0.037	-0.006 ± 0.075	0.395 ± 0.054
FBCNet	0.551 ± 0.080	0.101 ± 0.161	0.494 ± 0.097	0.613 ± 0.144	0.225 ± 0.288	0.597 ± 0.150	0.506 ± 0.052	0.012 ± 0.104	0.428 ± 0.070
EEGNet	0.488 ± 0.063	-0.024 ± 0.127	0.457 ± 0.068	0.510 ± 0.069	0.019 ± 0.137	0.490 ± 0.063	0.501 ± 0.052	0.001 ± 0.104	0.466 ± 0.057
EEG Conformer	0.574 ± 0.083	0.148 ± 0.166	0.536 ± 0.100	0.698 ± 0.137	0.393 ± 0.275	0.692 ± 0.142	0.524 ± 0.071	0.047 ± 0.142	0.460 ± 0.098
DeepCNN	0.523 ± 0.071	0.046 ± 0.143	0.449 ± 0.091	0.509 ± 0.038	0.019 ± 0.073	0.419 ± 0.082	0.505 ± 0.040	0.011 ± 0.081	0.439 ± 0.049
ShallowCNN	0.552 ± 0.075	0.103 ± 0.149	0.533 ± 0.079	0.652 ± 0.151	0.307 ± 0.300	0.646 ± 0.155	0.517 ± 0.056	0.034 ± 0.112	0.492 ± 0.066
IFNet	0.566 ± 0.094	0.132 ± 0.187	0.517 ± 0.112	0.647 ± 0.160	0.296 ± 0.319	0.627 ± 0.173	0.517 ± 0.057	0.033 ± 0.114	0.443 ± 0.068
EDPNet	0.515 ± 0.071	0.030 ± 0.142	0.494 ± 0.073	0.636 ± 0.132	0.271 ± 0.264	0.631 ± 0.134	0.474 ± 0.067	-0.051 ± 0.134	0.448 ± 0.074
BIOT	0.494 ± 0.067	-0.012 ± 0.135	0.470 ± 0.074	0.574 ± 0.080	0.160 ± 0.153	0.560 ± 0.084	0.505 ± 0.055	0.010 ± 0.110	0.465 ± 0.066
BrainOmni	0.533 ± 0.079	0.066 ± 0.157	0.493 ± 0.093	0.532 ± 0.050	0.063 ± 0.099	0.505 ± 0.066	0.509 ± 0.066	0.018 ± 0.131	0.456 ± 0.078
LaBraM	0.506 ± 0.053	0.012 ± 0.106	0.452 ± 0.070	0.547 ± 0.136	0.091 ± 0.273	0.528 ± 0.146	0.483 ± 0.068	-0.034 ± 0.135	0.432 ± 0.081
EEGPT	0.486 ± 0.065	-0.029 ± 0.130	0.452 ± 0.068	0.495 ± 0.067	-0.006 ± 0.132	0.483 ± 0.071	0.494 ± 0.073	-0.011 ± 0.145	0.462 ± 0.085
CBraMod	0.499 ± 0.060	-0.001 ± 0.119	0.433 ± 0.080	0.570 ± 0.063	0.123 ± 0.161	0.528 ± 0.097	0.494 ± 0.048	-0.012 ± 0.095	0.420 ± 0.062
MIRepNet	0.531 ± 0.071	0.062 ± 0.141	0.477 ± 0.086	0.602 ± 0.112	0.204 ± 0.222	0.575 ± 0.139	0.516 ± 0.061	0.032 ± 0.122	0.464 ± 0.075
NeuroMI (ours)	**0.578 ± 0.099**	**0.155 ± 0.198**	**0.558 ± 0.107**	**0.713 ± 0.111**	**0.426 ± 0.220**	**0.717 ± 0.104**	**0.553 ± 0.107**	**0.105 ± 0.215**	**0.529 ± 0.116**
Method	**BNCI2015_001**			**Kumar2024**			**AlexMI**		
	BA	**Kappa**	**F1**	**BA**	**Kappa**	**F1**	**BA**	**Kappa**	**F1**
ADFCNN	0.544 ± 0.073	0.087 ± 0.145	0.482 ± 0.094	0.599 ± 0.098	0.198 ± 0.197	0.551 ± 0.129	0.331 ± 0.045	-0.003 ± 0.068	0.235 ± 0.051
FBCNet	0.637 ± 0.126	0.274 ± 0.251	0.628 ± 0.131	0.709 ± 0.106	0.417 ± 0.212	0.704 ± 0.110	0.421 ± 0.075	0.131 ± 0.112	0.325 ± 0.079
EEGNet	0.522 ± 0.059	0.045 ± 0.119	0.504 ± 0.069	0.552 ± 0.081	0.105 ± 0.162	0.540 ± 0.087	0.331 ± 0.057	-0.003 ± 0.086	0.297 ± 0.063
EEG Conformer	0.673 ± 0.125	0.345 ± 0.251	0.663 ± 0.133	0.733 ± 0.099	0.466 ± 0.197	0.729 ± 0.101	0.423 ± 0.087	0.134 ± 0.131	0.373 ± 0.097
DeepCNN	0.550 ± 0.097	0.099 ± 0.194	0.493 ± 0.115	0.580 ± 0.100	0.161 ± 0.200	0.529 ± 0.124	0.367 ± 0.070	0.050 ± 0.105	0.283 ± 0.059
ShallowCNN	0.652 ± 0.134	0.304 ± 0.268	0.649 ± 0.135	0.724 ± 0.114	0.448 ± 0.227	0.722 ± 0.115	0.448 ± 0.122	0.172 ± 0.183	0.418 ± 0.127
IFNet	0.657 ± 0.144	0.314 ± 0.288	0.649 ± 0.150	0.751 ± 0.107	0.503 ± 0.213	0.748 ± 0.109	0.421 ± 0.116	0.131 ± 0.174	0.348 ± 0.131
EDPNet	0.592 ± 0.119	0.184 ± 0.237	0.581 ± 0.127	0.651 ± 0.103	0.303 ± 0.206	0.646 ± 0.106	0.333 ± 0.035	-0.000 ± 0.053	0.295 ± 0.035
BIOT	0.611 ± 0.111	0.222 ± 0.222	0.599 ± 0.118	0.567 ± 0.073	0.123 ± 0.146	0.545 ± 0.083	0.373 ± 0.095	0.059 ± 0.142	0.338 ± 0.094
BrainOmni	0.566 ± 0.070	0.132 ± 0.140	0.557 ± 0.074	0.536 ± 0.064	0.072 ± 0.128	0.515 ± 0.075	0.388 ± 0.067	0.081 ± 0.100	0.323 ± 0.081
LaBraM	0.616 ± 0.148	0.232 ± 0.296	0.600 ± 0.159	0.609 ± 0.103	0.218 ± 0.207	0.591 ± 0.116	0.385 ± 0.072	0.078 ± 0.109	0.345 ± 0.083
EEGPT	0.558 ± 0.078	0.116 ± 0.157	0.545 ± 0.081	0.537 ± 0.062	0.063 ± 0.123	0.510 ± 0.069	0.360 ± 0.050	0.041 ± 0.075	0.318 ± 0.055
CBraMod	0.588 ± 0.062	0.176 ± 0.123	0.564 ± 0.082	0.551 ± 0.060	0.102 ± 0.120	0.533 ± 0.069	0.365 ± 0.059	0.047 ± 0.088	0.278 ± 0.074
MIRepNet	0.663 ± 0.125	0.326 ± 0.251	0.654 ± 0.132	0.626 ± 0.091	0.251 ± 0.182	0.619 ± 0.095	0.338 ± 0.046	0.006 ± 0.069	0.270 ± 0.075
NeuroMI (ours)	**0.772 ± 0.126**	**0.544 ± 0.252**	**0.769 ± 0.128**	**0.759 ± 0.094**	**0.519 ± 0.188**	**0.757 ± 0.096**	**0.529 ± 0.121**	**0.294 ± 0.181**	**0.515 ± 0.126**

4 RESULTS

4.1 PERFORMANCE COMPARISON WITH BASELINES

Table 1 summarizes the performance comparison across all nine datasets. **NeuroMI achieves the best performance on all 27 metric-dataset combinations (9 datasets × 3 metrics)**, demonstrating consistent and robust superiority over all baselines. Averaged across all nine datasets, NeuroMI attains a mean BA of 0.646, Kappa of 0.344, and F1 of 0.636. Compared to the strongest baseline (EEG Conformer), NeuroMI yields absolute improvements of **+5.9% BA, +10.5% Kappa**, and **+7.8% F1**. These gains are achieved consistently across datasets with widely varying difficulty levels and class distributions, underscoring the broad generalization capability of our proposed foundation model. Traditional task-specific baselines perform competitively only when their inductive biases align with the dataset, but they generalize poorly and struggle on harder benchmarks; on Chang2025, the best such model (IFNet) reaches only 0.409 BA, whereas NeuroMI achieves 0.530. EEG foundation models also consistently underperform NeuroMI, indicating that general EEG pre-

training alone is insufficient for MI decoding. This gap is mainly due to our Stage 2 specialization pre-training, which injects sensorimotor neurophysiological priors as explicit supervision; without such domain-informed objectives, foundation models may fail to capture the discriminative MI representations.

4.2 Ablation Study on Specialization Pre-training

Table 2 shows the ablation results of loss terms on three datasets. Classification loss ($\mathcal{L}_{cls}$) alone is a strong baseline but consistently suboptimal. Adding FBCSP loss ($\mathcal{L}_{fbcsp}$) gives a stable single-loss gain, e.g., on Chang2025, BA improves from 0.476 to 0.512. Band power loss ($\mathcal{L}_{\mu\beta}$) also helps on several datasets, while ERD loss ($\mathcal{L}_{ERD}$) is less stable. Combining losses generally outperforms single auxiliary losses. The full NeuroMI loss achieves the best result on all three datasets and yields the highest average BA, Kappa, and F1. These results indicate that the loss terms provide complementary supervision, and their combination gives the most consistent cross-dataset performance. Results on the remaining six datasets are provided in Appendix A.

Table 2: Ablation results of loss terms on three MI datasets (mean $\pm$ standard deviation).

Method	Chang2025			Forenzo2023			Lee2019_MI		
	BA	Kappa	F1	BA	Kappa	F1	BA	Kappa	F1
$\mathcal{L}_{cls}$	0.476 $\pm$ 0.133	0.214 $\pm$ 0.200	0.448 $\pm$ 0.146	0.629 $\pm$ 0.113	0.259 $\pm$ 0.225	0.615 $\pm$ 0.123	0.712 $\pm$ 0.140	0.423 $\pm$ 0.280	0.694 $\pm$ 0.155
$\mathcal{L}_{cls} + \mathcal{L}_{fbcsp}$	0.512 $\pm$ 0.134	0.268 $\pm$ 0.201	0.483 $\pm$ 0.151	0.644 $\pm$ 0.102	0.289 $\pm$ 0.203	0.629 $\pm$ 0.112	0.705 $\pm$ 0.140	0.411 $\pm$ 0.280	0.691 $\pm$ 0.152
$\mathcal{L}_{cls} + \mathcal{L}_{ERD}$	0.482 $\pm$ 0.118	0.224 $\pm$ 0.178	0.455 $\pm$ 0.136	0.622 $\pm$ 0.105	0.243 $\pm$ 0.210	0.606 $\pm$ 0.114	0.697 $\pm$ 0.140	0.395 $\pm$ 0.280	0.680 $\pm$ 0.153
$\mathcal{L}_{cls} + \mathcal{L}_{\mu\beta}$	0.463 $\pm$ 0.110	0.195 $\pm$ 0.165	0.430 $\pm$ 0.123	0.629 $\pm$ 0.085	0.258 $\pm$ 0.170	0.611 $\pm$ 0.096	0.677 $\pm$ 0.139	0.353 $\pm$ 0.277	0.660 $\pm$ 0.149
$\mathcal{L}_{cls} + \mathcal{L}_{fbcsp} + \mathcal{L}_{ERD}$	0.503 $\pm$ 0.132	0.255 $\pm$ 0.198	0.480 $\pm$ 0.145	0.643 $\pm$ 0.102	0.286 $\pm$ 0.204	0.632 $\pm$ 0.109	0.699 $\pm$ 0.145	0.397 $\pm$ 0.289	0.684 $\pm$ 0.156
$\mathcal{L}_{cls} + \mathcal{L}_{fbcsp} + \mathcal{L}_{\mu\beta}$	0.517 $\pm$ 0.116	0.276 $\pm$ 0.174	0.494 $\pm$ 0.129	0.661 $\pm$ 0.110	0.322 $\pm$ 0.220	0.650 $\pm$ 0.119	0.709 $\pm$ 0.136	0.418 $\pm$ 0.272	0.700 $\pm$ 0.143
$\mathcal{L}_{cls} + \mathcal{L}_{ERD} + \mathcal{L}_{\mu\beta}$	0.495 $\pm$ 0.137	0.242 $\pm$ 0.205	0.466 $\pm$ 0.152	0.644 $\pm$ 0.101	0.289 $\pm$ 0.203	0.629 $\pm$ 0.110	0.712 $\pm$ 0.140	0.424 $\pm$ 0.279	0.701 $\pm$ 0.149
NeuroMI (all losses)	**0.530 $\pm$ 0.124**	**0.296 $\pm$ 0.186**	**0.514 $\pm$ 0.132**	**0.662 $\pm$ 0.110**	**0.324 $\pm$ 0.221**	**0.654 $\pm$ 0.119**	**0.716 $\pm$ 0.144**	**0.431 $\pm$ 0.288**	**0.708 $\pm$ 0.150**

4.3 Visualization Analysis

Analysis of Frequency-Aware Positional Encoding. Figure 3 shows the deviation maps of the gate on the positional encoding for the left- and right-hand classes. A positive deviation indicates a stronger contribution of the spectral prior at that position, whereas a negative deviation indicates a greater reliance on convolutional positional encoding. Although the magnitude of the deviations is modest, a clear double dissociation emerges at the sensorimotor channels C3 and C4. For the left-hand class, C3 exhibits a consistently positive deviation across all patches, while C4 shows a negative deviation of comparable magnitude. For the right-hand class, this pattern is reversed. This result is consistent with the contralateral lateralization of MI, in which left- and right-hand movements (imagery) predominantly engage the contralateral sensorimotor cortex. The presence of this pattern indicates that the frequency-aware positional encoding is not a static module, but adjusts its reliance on spectral priors at task-relevant channels.

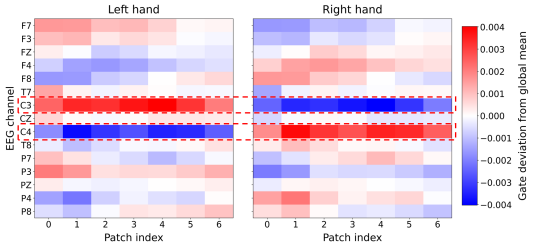


Figure 3: Class-specific deviation of the frequency-aware positional encoding gate.

Analysis of Model's Attention. To assess whether the model attends to physiologically meaningful regions, Figure 4 compares the class-discriminative topography derived from raw EEG band energy and from Grad-CAM channel importance. For each subject, the left panel shows the difference in raw 8-30 Hz log-power between the right-hand and foot classes, and the right panel shows the corresponding difference in Grad-CAM importance. The raw energy difference maps show that the subjects display clear lateralized patterns, with distinct positive and negative regions over the

sensorimotor areas. The Grad-CAM difference maps consistently exhibit focal patterns of high importance concentrated over the sensorimotor regions, particularly near C3, C4, and Cz. Also, the model's attention remains localized to the same spatial regions across both subjects. This alignment between Grad-CAM peaks and the sensorimotor cortex indicates that the model has learned to rely on physiologically meaningful channels for discrimination, rather than distributing attention uniformly across the scalp. The consistency of this spatial focus across subjects further supports the interpretability of the learned representations.

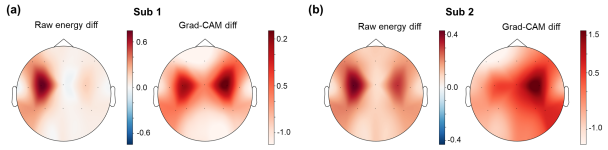


Figure 4: Comparison of raw energy topography and Grad-CAM channel importance.

Representation Visualizations across Pretraining Stages. To examine how the learned representations evolve across pretraining stages, we visualized the feature spaces of one subject using t-SNE (van der Maaten & Hinton, 2008), comparing raw EEG, the Stage 1 encoder output, and the Stage 2 encoder output (Figure 5). The raw EEG features exhibit heavily overlapping distributions between the two classes (Figure 5(a)), indicating that the original signal space provides little class separability (silhouette score: 0.016). After Stage 1 self-supervised pretraining (Figure 5(b)), the two classes begin to separate (silhouette score: 0.380), suggesting that the encoder has captured meaningful neurophysiological structure even without MI-specific supervision. After Stage 2 specialization pre-training (Figure 5(c)), the encoder output forms clearly distinct clusters for the two classes (silhouette score: 0.651). This progressive improvement demonstrates that the proposed pretraining pipeline learns increasingly discriminative and transferable representations, and also indicates that the model gradually transforms the raw EEG space into a well-structured latent space in which different classes are separable.

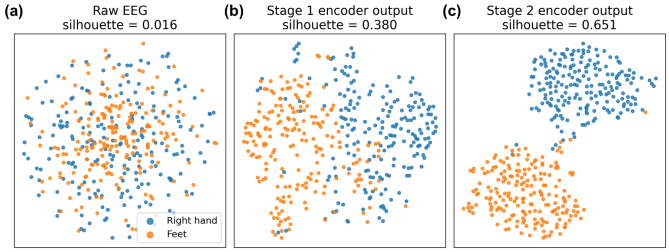


Figure 5: t-SNE visualization of feature representations.

5 Conclusion

In this paper, we proposed NeuroMI, a progressive EEG foundation model for MI decoding that bridges general EEG knowledge with MI-specific neurophysiology through three stages. The model integrates multi-scale CNN patch embedding, frequency-aware positional encoding, and spatial-temporal fusion Transformers to jointly capture multi-band oscillations and distributed brain coupling. Experimental results showed that NeuroMI outperforms both task-specific architectures and EEG foundation models, while ablation and visualization analyses confirmed the contribution of the mechanism-guided objectives and the physiological validity of the learned representations. These results indicate that explicitly grounding foundation models in domain-specific neurophysiology offers a practical pathway toward robust MI decoding in real-world BCIs.

AI Use Statement

In this work, the authors used generative AI tools to assist with language polishing and improving the clarity of the writing. The tools were used to refine grammar, sentence structure, and overall

readability of the text. All technical content, experimental design, data analysis, interpretations, and conclusions are the original work of the human authors. The authors have reviewed and verified all AI-assisted output and take full responsibility for the accuracy, originality, and integrity of the entire manuscript.

REPRODUCIBILITY STATEMENT

We have taken several measures to ensure the reproducibility of our results. All datasets used in this work are publicly available and were downloaded from TUEG and MOABB. The preprocessing pipeline is described in detail in Appendix C.3, including channel standardization, resampling, filtering, trial alignment, patch extraction, and label processing. The model architecture and training configurations for all three stages are provided in Section 3 and Appendix D, covering hyperparameters, optimizers, learning rates, schedulers, masking ratios, and loss weights. All experiments were implemented in PyTorch and trained on NVIDIA GPUs. The code and pretrained weights will be released upon publication. For the baseline methods, we use their official implementations and evaluate them under the same linear-probing protocol to ensure a fair comparison. Any remaining details required to reproduce the results are available from the authors upon reasonable request.

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-informed attention temporal convolutional network for EEG-based motor imagery classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, 2023. doi: 10.1109/TII.2022.3197419.

Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain–computer interface. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 2390–2397, 2008. doi: 10.1109/IJCNN.2008.4634130.

Alexandre Barachant. EEG motor imagery dataset from the PhD thesis "commande robuste d'un effecteur par une interface cerveau machine EEG asynchrone". Zenodo, 2012. URL https://doi.org/10.5281/zenodo.806023.

Benjamin Blankertz, Florian Losch, Matthias Krauledat, Guido Dornhege, Gabriel Curio, and Klaus-Robert Müller. The Berlin Brain–Computer Interface: Accurate performance from first-session in BCI-naive subjects. *IEEE Transactions on Biomedical Engineering*, 55(10):2452–2462, 2008. doi: 10.1109/TBME.2008.923152.

Stephanie Brandl and Benjamin Blankertz. Motor imagery under distraction—an open access BCI dataset. *Frontiers in Neuroscience*, 14, 2020. doi: 10.3389/fnins.2020.566147.

Wenwen Chang, Weixuan Kong, Guanghui Yan, Renjie Lv, Kaiyue Du, Muhammad Tariq Sadiq, Bin Guo, Rong Yin, and Xuan Liu. A multi-paradigm EEG dataset for studying upper limb rehabilitation exercises. *Scientific Data*, 12(1):1877, 2025. doi: 10.1038/s41597-025-06147-6.

Hohyun Cho, Minkyu Ahn, Sangtae Ahn, Moonyoung Kwon, and Sung Chan Jun. EEG datasets for motor imagery brain–computer interface. *GigaScience*, 6(7):gix034, 2017. doi: 10.1093/gigascience/gix034.

Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-GPT: Towards a foundation model for EEG. arXiv preprint arXiv:2311.03764, 2024. URL https://arxiv.org/abs/2311.03764.

G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller. Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms. *IEEE Transactions on Biomedical Engineering*, 51(6):993–1002, 2004. doi: 10.1109/TBME.2004.827088.

Pauline Dreyer, Aline Roc, Léa Pillette, Sébastien Rimbert, and Fabien Lotte. A large EEG database with users' profile information for motor imagery brain–computer interface research. *Scientific Data*, 10(1):580, 2023. doi: 10.1038/s41597-023-02445-z.

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, pp. 22541–22577, 2025. doi: 10.52202/085713-0760.

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.

Dylan Forenzo, Yixuan Liu, Jeehyun Kim, Yidan Ding, Taehyung Yoon, and Bin He. Integrating simultaneous motor imagery and spatial attention for EEG-BCI control. *IEEE Transactions on Biomedical Engineering*, 71(1):282–294, 2024. doi: 10.1109/TBME.2023.3298957.

Moritz Grosse-Wentrup, Christian Liefhold, Klaus Gramann, and Martin Buss. Beamforming in noninvasive brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 56(4):1209–1219, 2009. doi: 10.1109/TBME.2008.2009768.

Eva Guttmann-Flury, Xinjun Sheng, and Xiangyang Zhu. Dataset combining EEG, eye-tracking, and high-speed video for ocular activity analysis across BCI paradigms. *Scientific Data*, 12(1):587, 2025. doi: 10.1038/s41597-025-04861-9.

Can Han, Chen Liu, Yaqi Wang, Crystal Cai, Jun Wang, and Dahong Qian. A spatial-spectral and temporal dual prototype network for motor imagery brain–computer interface. arXiv preprint arXiv:2407.03177, 2024. URL https://arxiv.org/abs/2407.03177.

Yimin Hou, Shuyue Jia, Xiangmin Lun, Ziqian Hao, Yan Shi, Yang Li, Rui Zeng, and Jinglei Lv. GCNs-Net: A graph convolutional neural network approach for decoding time-resolved EEG motor imagery signals. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7312–7323, 2024. doi: 10.1109/TNNLS.2022.3202569.

Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2958–2965, 2020. doi: 10.1109/SMC42975.2020.9283028.

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

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf.

Wei-Bang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations*, pp. 55436–55457, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8b4add8b0aa8749d80a34ca5d941c355-Paper-Conference.pdf.

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

Satyam Kumar, Hussein Alawieh, Frigyes Samuel Racz, Rawan Fakhreddine, and José del R. Millán. Transfer learning promotes acquisition of individual BCI skills. *PNAS Nexus*, 3(2):pgae076, 2024. doi: 10.1093/pnasnexus/pgae076.

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, July 2018. doi: 10.1088/1741-2552/aace8c.

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

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. arXiv preprint arXiv:2601.17883, 2026a. URL https://arxiv.org/abs/2601.17883.

Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, Yuheng Chen, Shaojie Hou, Xiaolian Zhu, and Dongrui Wu. MIRepNet: A pipeline and pre-trained model for EEG-based motor imagery classification. *Knowledge-Based Systems*, 343:115966, 2026b. doi: 10.1016/j.knosys.2026.115966.

Haijie Liu, Penghu Wei, Haochong Wang, Xiaodong Lv, Wei Duan, Meijie Li, Yan Zhao, Qingmei Wang, Xinyuan Chen, Gaige Shi, Bo Han, and Junwei Hao. An EEG motor imagery dataset for brain computer interface in acute stroke patients. *Scientific Data*, 11(1):131, 2024. doi: 10.1038/s41597-023-02787-8.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. arXiv preprint arXiv:2506.09110, 2026. URL https://arxiv.org/abs/2506.09110.

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.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10, 2016. doi: 10.3389/fnins.2016.00196.

G. Pfurtscheller and F. 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.

Abbas Salami, Javier Andreu-Perez, and Helge Gillmeister. EEG-ITNet: An explainable inception temporal convolutional network for motor imagery classification. *IEEE Access*, 10:36672–36685, 2022. doi: 10.1109/ACCESS.2022.3161489.

G. Schalk, D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. BCI2000: A general-purpose brain–computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. doi: 10.1002/hbm.23730.

Jian Shi, Danyang Chen, Xingwei Zhao, Zhixian Zhao, Shengjie Li, Yeguang Xu, Tao Ding, Zheng Zhu, Peng Zhang, Qing Ye, Yingxin Tang, Ping Zhang, Bo Tao, and Zhouping Tang. HEFMI-ICH: a hybrid EEG-fnirs motor imagery dataset for brain-computer interface in intracerebral hemorrhage. *Scientific Data*, 12(1):1816, 2025. doi: 10.1038/s41597-025-06100-7.

Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Do-Won Kim, Jichai Jeong, Han-Jeong Hwang, and Klaus-Robert Müller. Open access dataset for EEG + NIRS single-trial classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1735–1745, 2017. doi: 10.1109/TNSRE.2016.2628057.

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.

James R. Stieger, Stephen A. Engel, and Bin He. Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Scientific Data*, 8(1):98, 2021. doi: 10.1038/s41597-021-00883-1.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot Mueller-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, 2012. doi: 10.3389/fnins.2012.00055.

Wei Tao, Ze Wang, Chi Man Wong, Ziyu Jia, Chang Li, Xun Chen, C. L. Philip Chen, and Feng Wan. ADFCNN: Attention-based dual-scale fusion convolutional neural network for motor imagery brain–computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:154–165, 2024. doi: 10.1109/TNSRE.2023.3342331.

Mojgan Tavakolan, Zack Frehlick, Xinyi Yong, and Carlo Menon. Classifying three imaginary states of the same upper extremity using time-domain features. *PLOS ONE*, 12(3):1–18, 2017. doi: 10.1371/journal.pone.0174161.

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

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024. doi: 10.52202/079017-1239.

J. Wang, L. Yao, and Y. 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 *International Conference on Learning Representations*, pp. 75310–75346, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bbbd6d915cb90be21c1254a82d45cedd-Paper-Conference.pdf.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. EEGMamba: An EEG foundation model with Mamba. *Neural Networks*, 192:107816, 2025b. doi: 10.1016/j.neunet.2025.107816.

Xiaoli Wu, Wenhui Zhang, Zhi-Bo Fu, Roy Tsz-Hei Cheung, and Rosa H. M. Chan. An investigation of in-ear sensing for motor task classification. *Journal of Neural Engineering*, 17, 2020. URL https://api.semanticscholar.org/CorpusID:222830818.

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41179–41212, 2025. doi: 10.52202/085713-1375.

Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. A multi-day and high-quality EEG dataset for motor imagery brain-computer interface. *Scientific Data*, 12 (1):488, 2025. doi: 10.1038/s41597-025-04826-y.

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, pp. 78240–78260, 2023. doi: 10.52202/075280-3420.

Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLOS ONE*, 9(12):1–19, 2014. doi: 10.1371/journal.pone.0114853.

Qing Zhou, Jiafan Lin, Lin Yao, Yueming Wang, Yan Han, and Kedi Xu. Relative power correlates with the decoding performance of motor imagery both across time and subjects. *Frontiers in Human Neuroscience*, 15, 2021. doi: 10.3389/fnhum.2021.701091.

A Additional Results

A.1 Ablation Results on Remaining Six Test Datasets

Table A.1 reports the ablation results on the remaining six datasets. The overall trend is consistent with that observed on the three representative datasets in the main text. Classification loss ($\mathcal{L}_{cls}$) alone provides a reasonable baseline, and adding FBCSP loss ($\mathcal{L}_{fbcsp}$) gives a stable improvement on most datasets, e.g., BA increases from 0.549 to 0.566 on Shin2017A and from 0.512 to 0.542 on AlexMI. Band power loss ($\mathcal{L}_{\mu\beta}$) and ERD loss ($\mathcal{L}_{ERD}$) are individually less stable, but combining them with FBCSP loss generally yields further gains. The three-term variant '$\mathcal{L}_{cls}$+$\mathcal{L}_{fbcsp}$+$\mathcal{L}_{\mu\beta}$' achieves the best BA on three datasets, while the full NeuroMI loss achieves the best or tied-best result on the other three. These results confirm that the mechanism-guided objectives provide complementary supervision, and their combination yields the most consistent cross-dataset performance. No single auxiliary loss dominates across all datasets, which further supports the design choice of jointly optimizing all loss terms.

Table A.1: Ablation results on the remaining six MI datasets (mean ± standard deviation).

Method	Shin2017A			BNCI2003_004			HefmiIch2025		
	BA	Kappa	F1	BA	Kappa	F1	BA	Kappa	F1
$\mathcal{L}_{cls}$	0.549 ± 0.097	0.098 ± 0.194	0.512 ± 0.111	0.690 ± 0.129	0.377 ± 0.261	0.680 ± 0.136	0.540 ± 0.123	0.080 ± 0.246	0.500 ± 0.142
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$	0.566 ± 0.107	0.132 ± 0.214	0.532 ± 0.122	0.681 ± 0.149	0.357 ± 0.304	0.646 ± 0.184	0.538 ± 0.118	0.075 ± 0.237	0.498 ± 0.134
$\mathcal{L}_{cls}$ + $\mathcal{L}_{ERD}$	0.537 ± 0.100	0.075 ± 0.200	0.505 ± 0.108	0.687 ± 0.119	0.376 ± 0.238	0.673 ± 0.128	**0.553 ± 0.111**	**0.105 ± 0.222**	0.515 ± 0.129
$\mathcal{L}_{cls}$ + $\mathcal{L}_{\mu\beta}$	0.553 ± 0.090	0.106 ± 0.180	0.515 ± 0.104	0.689 ± 0.168	0.381 ± 0.331	0.667 ± 0.186	0.536 ± 0.105	0.071 ± 0.211	0.496 ± 0.124
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$ + $\mathcal{L}_{ERD}$	0.580 ± 0.089	0.161 ± 0.179	0.548 ± 0.107	0.712 ± 0.091	**0.427 ± 0.180**	0.700 ± 0.097	0.535 ± 0.105	0.070 ± 0.210	0.500 ± 0.121
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$ + $\mathcal{L}_{\mu\beta}$	**0.590 ± 0.103**	**0.179 ± 0.206**	**0.563 ± 0.118**	0.689 ± 0.115	0.385 ± 0.227	0.677 ± 0.132	0.537 ± 0.117	0.075 ± 0.233	0.502 ± 0.134
$\mathcal{L}_{cls}$ + $\mathcal{L}_{ERD}$ + $\mathcal{L}_{\mu\beta}$	0.568 ± 0.088	0.137 ± 0.176	0.540 ± 0.100	0.701 ± 0.112	0.399 ± 0.225	0.674 ± 0.127	0.535 ± 0.107	0.071 ± 0.213	0.494 ± 0.127
NeuroMI (all losses)	0.578 ± 0.099	0.155 ± 0.198	0.558 ± 0.107	**0.713 ± 0.111**	0.426 ± 0.220	**0.717 ± 0.104**	**0.553 ± 0.107**	**0.105 ± 0.215**	**0.529 ± 0.116**
Method	**BNCI2015_001**			**Kumar2024**			**AlexMI**		
	BA	Kappa	F1	BA	Kappa	F1	BA	Kappa	F1
$\mathcal{L}_{cls}$	0.762 ± 0.134	0.523 ± 0.267	0.758 ± 0.137	0.727 ± 0.102	0.455 ± 0.203	0.722 ± 0.106	0.512 ± 0.095	0.269 ± 0.142	0.474 ± 0.111
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$	0.771 ± 0.128	0.541 ± 0.255	0.765 ± 0.133	0.743 ± 0.103	0.485 ± 0.205	0.738 ± 0.106	0.542 ± 0.083	0.312 ± 0.124	0.514 ± 0.091
$\mathcal{L}_{cls}$ + $\mathcal{L}_{ERD}$	0.765 ± 0.131	0.531 ± 0.261	0.761 ± 0.135	0.738 ± 0.095	0.476 ± 0.191	0.734 ± 0.100	0.502 ± 0.093	0.253 ± 0.140	0.466 ± 0.109
$\mathcal{L}_{cls}$ + $\mathcal{L}_{\mu\beta}$	0.752 ± 0.142	0.504 ± 0.284	0.746 ± 0.148	0.729 ± 0.095	0.459 ± 0.189	0.725 ± 0.099	0.533 ± 0.087	0.300 ± 0.131	0.503 ± 0.092
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$ + $\mathcal{L}_{ERD}$	0.765 ± 0.133	0.529 ± 0.266	0.759 ± 0.141	0.737 ± 0.106	0.474 ± 0.211	0.733 ± 0.109	0.469 ± 0.117	0.203 ± 0.176	0.434 ± 0.141
$\mathcal{L}_{cls}$ + $\mathcal{L}_{fbcsp}$ + $\mathcal{L}_{\mu\beta}$	**0.777 ± 0.124**	**0.555 ± 0.248**	**0.774 ± 0.128**	0.754 ± 0.098	0.508 ± 0.196	0.751 ± 0.100	**0.560 ± 0.065**	**0.341 ± 0.097**	**0.544 ± 0.066**
$\mathcal{L}_{cls}$ + $\mathcal{L}_{ERD}$ + $\mathcal{L}_{\mu\beta}$	0.758 ± 0.130	0.516 ± 0.260	0.755 ± 0.133	0.749 ± 0.099	0.497 ± 0.198	0.745 ± 0.103	0.502 ± 0.129	0.253 ± 0.194	0.483 ± 0.129
NeuroMI (all losses)	0.772 ± 0.126	0.544 ± 0.252	0.769 ± 0.128	**0.759 ± 0.094**	**0.519 ± 0.188**	**0.757 ± 0.096**	0.529 ± 0.121	0.294 ± 0.181	0.515 ± 0.126

A.2 Analysis of Multi-Scale CNN Patch Embedding

To validate the multi-scale CNN design, we analyzed the branch responses for three kernel sizes on Kumar2024. Figure A.1(a) shows that larger kernels produce stronger response magnitudes, which is expected due to the longer temporal receptive field. Figure A.1(b) shows the class-discriminative response (left hand - right hand) for each branch, with white contours marking statistically significant units (Welch t-test, $p < 0.05$). The number of significant units is 501, 620, and 559 for kernel size = 15, 31, and 63, respectively. The non-monotonic distribution of discriminability across scales suggests that the three kernels contribute differently to class discrimination, rather than being fully redundant, which supports the use of a multi-scale design over a single-scale one.

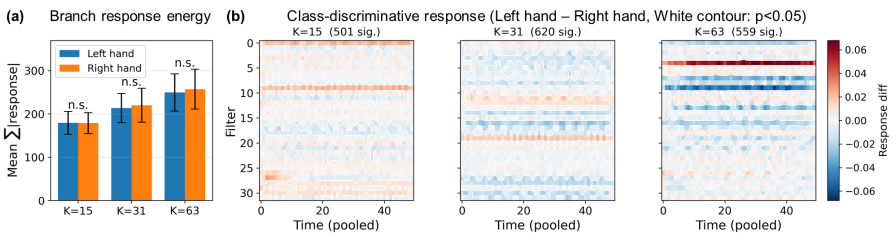


Figure A.1: The branch responses of multi-scale CNN patch embedding. (a) Branch response magnitude: summed absolute response per trial for each temporal kernel, averaged over trials and subjects. (b) Class-discriminative units per branch: difference of branch responses between Left hand and Right hand (all subjects pooled, all patches averaged).

A.3 Analysis of Spatial Representation

To further characterize the spatial representations learned by the model, we visualized the directed inter-channel attention and the channel-wise class correlation for left- and right-hand MI decoding, as shown in Figure A.2. The circular panel shows the strongest spatial-attention connections among the 15 EEG channels, where ribbon width encodes relative attention strength and color denotes the source region. The connections span frontal, central, temporal, and parietal regions, suggesting that the model relies on distributed inter-channel interactions rather than a small set of isolated electrodes. The two scalp maps show the Pearson correlation between each channel's activity and the class label for left- and right-hand trials, respectively. The two classes exhibit partially complementary correlation patterns, with the strongest positive and negative correlations concentrated over sensorimotor regions. These observations are consistent with the distributed and class-dependent spatial organization of hand MI, and provide qualitative evidence that the learned spatial attention is not concentrated on a single electrode.

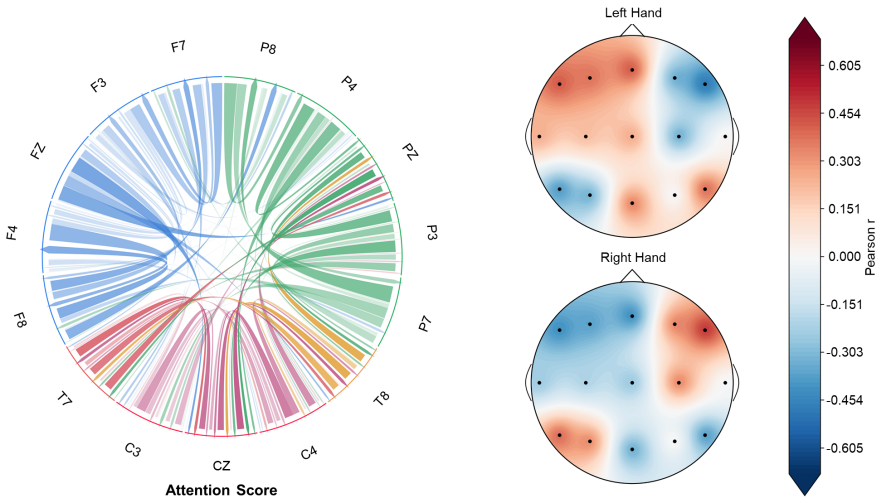


Figure A.2: Directed inter-channel spatial-attention relationships and channel-wise Pearson correlation with the class label.

A.4 Effect of Scaling Data Size and Model Size

We first investigated how decoding performance scales with the amount of Stage 2 training data, as shown in Figure A.3. For each test dataset, we varied the proportion of training data from 20% to 100% (denoted as data ratio) while keeping the model architecture fixed at the base configuration, and evaluated BA across all subjects. As shown in Figure A.3, BA improves monotonically with increasing data ratio on nearly all nine datasets, confirming a clear positive scaling effect. The gains are most pronounced on datasets with moderate baseline performance, such as Lee2019_MI (from 0.660 to 0.716) and Chang2025 (from 0.466 to 0.530). When averaging across datasets, the BA increases from 0.612 at data ratio 0.2 to 0.646 at data ratio 1.0, corresponding to an average absolute gain of approximately 0.034. These results suggest that further expanding the pretraining data is likely to yield continued gains, motivating the data-efficient design of the proposed framework.

We then examined the complementary dimension of model capacity by training four variants (tiny, small, base, and large) on all datasets under the full data ratio, as shown in Figure A.4. In contrast to the monotonic trend observed for data scaling, model-capacity scaling exhibits a clear non-monotonic pattern. Across eight of the nine datasets, balanced accuracy peaks at the base configuration, with tiny-to-base gains typically between 0.01 and 0.03, whereas the large variant degrades consistently. When averaged across datasets, the base model achieves the best mean BA, followed by small, tiny, and large models. These results indicate that increasing model capacity alone, without a corresponding increase in training data, does not necessarily lead to better performance.

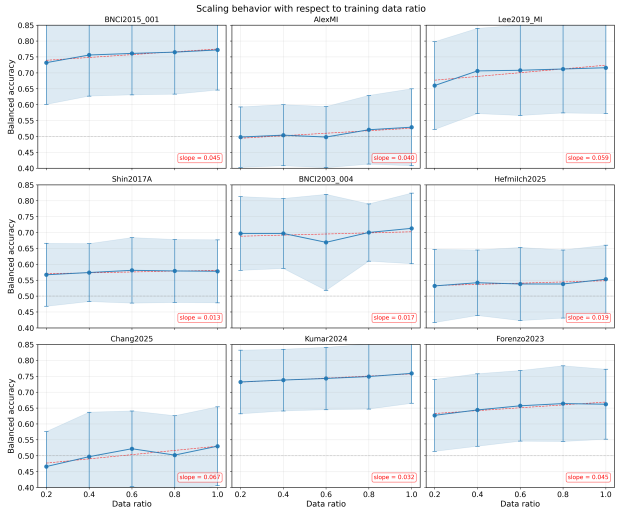


Figure A.3: BA variation under different Stage 2 training data volumes.

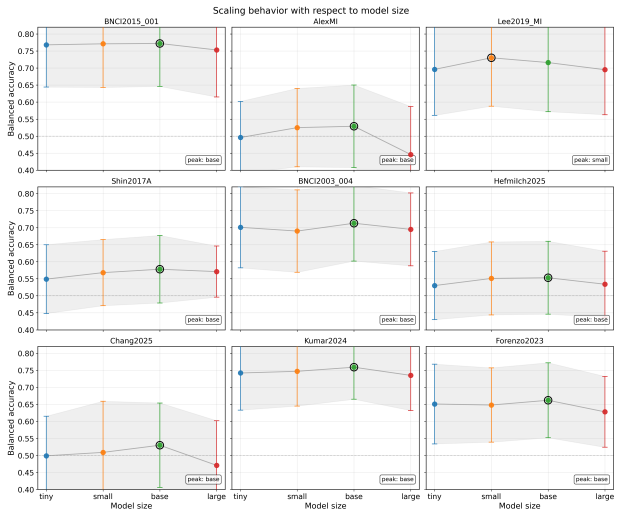


Figure A.4: BA variation under different model capacities.

A.5 Analysis of Model-Predicted Neurophysiological Features

To assess whether the model learns physiologically meaningful representations, we compared the spatial distribution of class differences between raw features and model-predicted features, as shown in Figure A.5. For the β band (Figure A.5(a)), the raw difference map (Left hand - Right hand) shows only weak and diffuse patterns with no clear spatial structure. In contrast, the model-predicted β difference map exhibits a pronounced contralateral pattern, with strong positive differences (red) concentrated over the left sensorimotor area (C3) and negative differences (blue) over the right sensorimotor area (C4). A similar trend is observed for the μ band (Figure A.5(b)). The raw μ difference map displays only mild lateralization, whereas the model-predicted μ difference map exhibits a clear double dissociation, with prominent positive differences at C3 and C4. These patterns are consistent with the well-documented contralateral ERD lateralization of MI, indicating that the model has internalized the underlying neurophysiological mechanisms.

To further quantify these observations, we computed three class-discriminability metrics (Fisher score, Cohen's d, and L2 distance) for both raw ERD and model-predicted ERD across the C3, CZ, and C4 channels (Figure A.5(c)). Across all three metrics, the model-predicted features consistently

outperform the raw features. At C3 and C4, the model-predicted features achieve substantially higher scores than the raw features, whereas at CZ, both remain relatively low. This indicates that the model's predictions not only reproduce the expected spatial patterns but also enhance the separability of the two classes at the most task-relevant channels. Taken together, these results demonstrate that the mechanism-guided specialization pre-training enables the model to learn representations that are both physiologically plausible and highly discriminative for MI decoding.

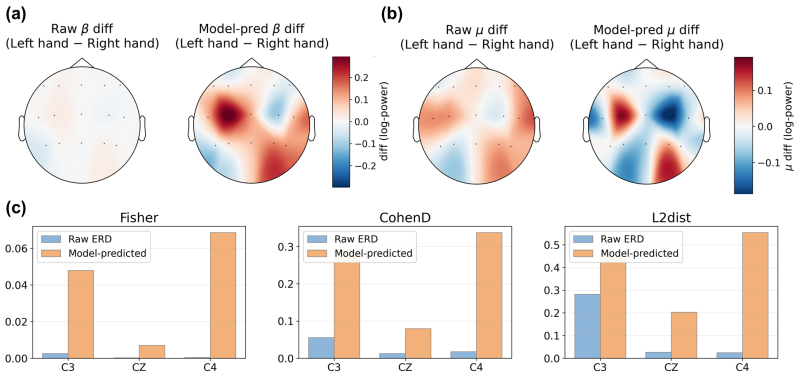


Figure A.5: The spatial distribution of class differences between raw features and model-predicted features.

Figure A.6(a) compares the model-predicted values against the ground-truth values for four neurophysiological features: μ band power, β band power, μ ERD, and β ERD. Across all four features, the predictions exhibit a strong linear relationship with the ground truth, with Pearson correlation coefficients of r = 0.95 for μ power, r = 0.97 for β power, r = 0.86 for μ ERD, and r = 0.72 for β ERD (all $p < 0.001$). The density plots show that the majority of samples lie closely along the diagonal, indicating that the model can accurately infer these features even without task-specific supervision. The slightly lower correlations for ERD compared to absolute band power suggest that ERD, which reflects relative changes with respect to a baseline, is more challenging to predict, likely due to its dependence on baseline stability and trial-to-trial variability.

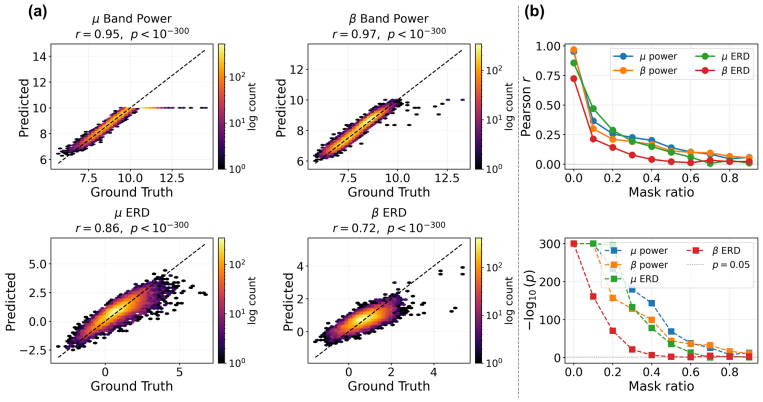


Figure A.6: Prediction of EEG features and robustness to masking. (a) Predicted vs. true values for four EEG features. (b) Correlation drops as more input patches are masked (mask ratio from 0 to 0.9).

Figure A.6(b) evaluates how the prediction quality degrades as the proportion of masked patches increases from 0% to 90%. As expected, the Pearson correlation decreases monotonically for all four features as the mask ratio increases. However, the degradation rates differ markedly across features. The μ and β band power maintain relatively high correlations ($r > 0.2$) up to a mask ratio of 0.5, whereas the ERD features drop more rapidly, with r falling below 0.1 beyond a mask ratio of 0.3. This indicates that absolute band power is more robust to missing information, whereas ERD

relies more heavily on the temporal context of the full signal. The lower panel reports the statistical significance of these correlations. All four features remain statistically significant ($p < 0.05$) up to a mask ratio of approximately 0.6, beyond which the ERD features lose significance while the band power features remain marginally significant. Overall, these results demonstrate that the model learns robust internal representations that support accurate prediction of key neurophysiological features, with graceful degradation under increasing input masking.

B Related Work

B.1 Traditional MI Decoding Methods

MI EEG decoding has long relied on processing pipelines composed of feature engineering and shallow classifiers. Common Spatial Pattern (CSP) (Blankertz et al., 2008) extracts discriminative features related to motor intention from multichannel EEG by finding spatial projections that maximize inter-class variance differences, and it is typically combined with linear discriminant analysis or support vector machines. FBCSP (Ang et al., 2008) extends this process to multiple frequency bands, allowing the model to use ERD and ERS patterns across different neural oscillations. These methods offer good physiological interpretability and remain competitive when training samples are limited, but their performance depends heavily on frequency-band partitioning, time-window selection, and covariance estimation. Optimizing feature extraction and the classifier independently also makes it difficult for the model to adapt low-level representations to the final recognition objective.

In turn, deep learning transforms MI decoding into an end-to-end representation-learning problem. DeepCNN (Schirrmeister et al., 2017) progressively builds hierarchical features through multiple temporal and spatial convolutional layers, learning local spatiotemporal patterns directly from raw EEG. ShallowCNN (Schirrmeister et al., 2017) retains the basic idea of FBCSP and highlights band-power-related discriminative information through squaring, pooling, and logarithmic transformation. EEGNet (Lawhern et al., 2018) further uses depth-wise and separable convolutions to perform temporal filtering, cross-channel spatial modeling, and feature fusion with few parameters, making it a widely used compact baseline for EEG decoding. FBCNet (Mane et al., 2021) explicitly combines filter banks, multi-view spatial convolutions, and variance statistics, enabling the network to extract features related to sensorimotor rhythms from different frequency ranges. Compared with traditional machine-learning methods, these networks reduce reliance on handcrafted feature design, but their temporal receptive fields and frequency selection are generally still determined by fixed convolutional kernels or predefined filter banks.

Recent methods have begun to introduce more detailed frequency-, scale-, and discriminability-aware modeling into network architectures. ADFCNN (Tao et al., 2024) uses dual-scale convolutional branches to extract EEG patterns over different temporal ranges and fuses complementary features through attention. IFNet (Wang et al., 2023) models relationships among different frequency components through interactive frequency convolution rather than treating each band as a completely independent information source. EDPNet (Han et al., 2024) builds a lightweight decoding framework from spatial-spectral fusion, long-range temporal features, and few-shot classification. It uses multiscale variance pooling to describe energy changes over different durations and constrains within-class compactness and between-class separation through dual-prototype learning. These designs show that incorporating neurophysiological priors into network architectures can effectively improve the separability of MI features, while also indicating that different datasets may depend on different temporal scales and frequency patterns.

Self-attention provides another technical route for modeling long-range dependencies. EEG Conformer (Song et al., 2023) first converts continuous EEG into a sequence of local features through convolutional modules and then uses a Transformer to capture global relationships across time segments, thereby combining local pattern extraction with long-range context modeling. These methods broaden the scope of EEG representation learning, but existing models generally still use fixed-length time segments and generic positional encodings to describe sequence order. Neural oscillations at different scales, frequency structures with clear physiological meaning, and the coupling formed by spatial activation changing over time have not yet been sufficiently modeled in a unified representation space. Accordingly, how to learn general representations from large-scale EEG data that retain multiscale, frequency-aware, and spatiotemporal interaction information remains an important problem for further improving MI decoding.

B.2 EEG Foundation Models

EEG foundation models seek to learn transferable representations from large-scale unlabeled recordings, thereby reducing the reliance of downstream applications on task-specific annotations. BENDR pioneered large-scale self-supervised EEG pre-training by applying a wav2vec-style masked contrastive objective to the TUEG corpus and demonstrated transfer across subjects, acquisition systems, and EEG tasks (Kostas et al., 2021). BIOT subsequently represented channel-wise signal segments as a unified sequence of biosignal tokens, enabling joint learning from datasets with different channel configurations and recording lengths (Yang et al., 2023). Building upon masked modeling, LaBraM trains a neural tokenizer through vector-quantized spectrum prediction and learns contextual representations by recovering the discrete codes of masked channel patches (Jiang et al., 2024). EEGPT instead employs teacher-student learning with spatio-temporal representation alignment to derive robust representations from noisy EEG signals (Wang et al., 2024). NeuroLM further integrates discretized EEG tokens with a large language model through multi-channel autoregressive pre-training and multi-task instruction tuning (Jiang et al., 2025). Together, these studies have established a broad family of EEG pre-training objectives, ranging from contrastive learning and masked token prediction to representation alignment and generative sequence modeling.

Beyond pre-training objectives, recent research has increasingly considered the structural heterogeneity and computational characteristics of EEG. CBraMod introduces a criss-cross Transformer that models spatial and temporal dependencies through separate attention pathways and learns transferable features using masked waveform reconstruction (Wang et al., 2025a). EEGMamba adopts bidirectional state-space modeling to capture long-range dependencies with favorable computational efficiency on large EEG corpora (Wang et al., 2025b). To accommodate heterogeneous acquisition configurations, REVE combines masked autoencoding with four-dimensional positional encodings that describe the spatial locations of electrodes and the temporal positions of signal patches (El Ouahidi et al., 2025). Its large-scale pre-training across diverse datasets and electrode layouts supports transfer to a wide range of clinical and BCI tasks. CodeBrain further employs decoupled temporal-frequency tokenization and multi-scale sequence modeling to preserve spectral information while capturing both local neural events and long-range dependencies (Ma et al., 2026). These developments extend EEG foundation modeling from general-purpose representation learning toward explicit consideration of channel topology, spatio-temporal organization, spectral content, and computational scalability.

An emerging direction applies foundation-model pre-training to MI decoding and explores representations specialized for the MI paradigm. Neuro-GPT combines an EEG encoder with a GPT-style contextual model, using masked signal reconstruction for pre-training before adaptation to low-data MI classification (Cui et al., 2024). General-purpose foundation models have also demonstrated transferability to MI benchmarks: CBraMod evaluates cross-subject decoding on PhysioNet-MI and SHU-MI (Wang et al., 2025a), while REVE reports results on both PhysioNet-MI and BCI Competition IV-2a under a unified evaluation framework (El Ouahidi et al., 2025). More recently, MIRepNet develops a paradigm-specific foundation model using a channel template, cross-dataset signal alignment, and a hybrid objective that combines masked token reconstruction with supervised MI classification (Liu et al., 2026b). This design supports rapid subject-specific adaptation across multiple MI datasets with limited calibration data. Overall, this line of research reflects a gradual progression from transferring general EEG representations to developing paradigm-aware foundation models that jointly consider cross-dataset generalization and the distinctive structure of sensorimotor EEG.

C More Details for Experiment Settings

C.1 Second-Stage Pretraining Dataset

For the second-stage pretraining, we integrated **17 public MI EEG datasets** to build a mixed training corpus covering diverse subjects, acquisition devices, channel configurations, sampling rates, and MI paradigms. The datasets are listed as follows:

Dreyer2023 (Dreyer et al., 2023): This dataset contains multiple MI experiment subsets covering different subjects and experimental conditions. The data mainly include left- and right-hand MI tasks, along with baseline recordings such as eyes-open and eyes-closed conditions.

Cho2017 (Cho et al., 2017): This dataset contains left- and right-hand MI EEG recordings from 52 subjects. The data were acquired with a 64-channel system, with electrodes arranged according to the 10-10 system and an original sampling rate of 512 Hz.

PhysionetMI (Schalk et al., 2004): This dataset comes from the PhysioNet EEG Motor Movement/Imagery Database and contains large-scale EEG recordings from 109 subjects. The data were acquired with 64 channels and cover multiple tasks, including actual movements and MI of the fists and feet.

Weibo2014 (Yi et al., 2014): This dataset contains EEG signals recorded from 10 subjects performing multiple motor and MI tasks. A multichannel EEG system was used, with tasks covering left-hand, right-hand, both-feet, and combined MI classes, together with corresponding rest periods.

Wu2020 (Wu et al., 2020): This dataset contains EEG recordings from multiple subjects performing MI tasks, mainly involving different limb-imagery classes. It uses a multichannel EEG system and includes multiple sessions or experimental runs, providing MI samples from different subjects and recording conditions.

Tavakolan2017 (Tavakolan et al., 2017): This is an MI EEG dataset for BCI research, containing multichannel EEG signals recorded while subjects performed different limb-imagery tasks.

Zhou2020 (Zhou et al., 2021): This dataset contains EEG data recorded from multiple subjects performing MI tasks, covering different motor classes and corresponding experimental runs.

GuttmannFlury2025_ME (Guttmann-Flury et al., 2025): This is the motor execution (ME) subset of a multimodal dataset collected from 31 subjects, with EEG, eye-tracking, and high-speed video recordings. Subjects performed actual left- and right-hand grasping, following the same task structure as the corresponding MI subset. EEG was sampled at 1000 Hz.

Schirrmeister2017 (Schirrmeister et al., 2017): This dataset contains MI EEG recordings from multiple subjects and MI classes.

Stieger2021 (Stieger et al., 2021): This dataset contains multichannel EEG signals recorded while subjects performed MI tasks, covering different motor classes and experimental sessions.

GuttmannFlury2025_MI (Guttmann-Flury et al., 2025): This is a relatively recent MI EEG dataset containing EEG signals recorded while subjects performed different hand-imagery tasks. It uses a multichannel EEG acquisition system and includes repeated experiments and multiple MI trials.

Yang2025 (Yang et al., 2025): This multi-day dataset contains EEG recordings from 62 healthy subjects across three sessions on different days. The two-class subset includes 51 subjects performing left- and right-hand grasping imagery, while the three-class subset includes 11 subjects and adds foot-hooking imagery. EEG was recorded from 59 channels at 1000 Hz. Each session contains 200 or 300 trials, respectively, with a 4-s MI period in each trial.

Liu2024 (Liu et al., 2024): This dataset contains EEG recordings from 50 acute stroke patients performing left- and right-hand grasping imagery. Signals were acquired from 29 EEG channels and two EOG channels at 500 Hz. Each subject completed one session with 40 trials, comprising 20 trials for each hand. Each trial lasts 8 s and includes instruction, a 4-s MI period, and rest, providing clinical MI recordings from patients within 30 days after stroke.

Brandl2020 (Brandl & Blankertz, 2020): This dataset contains left- and right-hand MI recordings from 16 healthy subjects, acquired from 63 EEG channels at 1000 Hz. Each subject completed one calibration run and six runs with auditory feedback, totaling 504 trials.

BNCI2014_001 (Tangermann et al., 2012): Also known as BCI Competition IV Dataset 2a, this dataset contains recordings from nine subjects performing four MI tasks: left hand, right hand, both feet, and tongue. EEG was acquired from 22 channels at 250 Hz, with three additional EOG channels. Each subject completed two sessions on different days, with six runs and 288 trials per session, yielding 72 trials for each class in each session.

BNCI2014_004 (Tangermann et al., 2012): Also known as BCI Competition IV Dataset 2b, this dataset contains left- and right-hand MI recordings from nine subjects. Signals were acquired at 250 Hz using three bipolar EEG channels centered on C3, Cz, and C4, together with three EOG channels. Each subject completed five sessions: two without feedback and three with visual feedback.

GrosseWentrup2009 (Grosse-Wentrup et al., 2009): This dataset contains EEG recordings of subjects performing left- and right-hand MI tasks. It focuses on EEG changes in different sensorimotor regions during MI.

C.2 Test Datasets

Chang2025 (Chang et al., 2025): This dataset contains multiparadigm upper-limb rehabilitation EEG data from 28 healthy subjects, acquired with a 59-channel system at 1000 Hz. Its MI subset includes left-hand, right-hand, and both-hand grasp-and-release imagery; the task phase lasts 4 s, with 40 trials per class in this paradigm.

Forenzo2023 (Forenzo et al., 2024): This dataset contains EEG recordings from 25 subjects across five sessions, acquired with a 64-channel system at 1000 Hz. Its standard MI subset consists of left- and right-hand MI using a continuous cursor-tracking paradigm, with five 60-s trials per run.

Lee2019_MI (Lee et al., 2019): This is the MI subset of OpenBMI and contains left- and right-hand grasping-imagery recordings from 54 subjects across two sessions. EEG was acquired with 62 electrodes and an amplifier at 1000 Hz. Each session's training and test phases contain 100 trials balanced between the two hands, and each MI trial lasts 4 s.

Shin2017A (Shin et al., 2017): This dataset contains left- and right-hand MI recordings from 29 subjects and is a hybrid BCI dataset with simultaneous EEG and near-infrared signal acquisition. EEG was recorded with 30 active electrodes at an original sampling rate of 1000 Hz. Each subject completed three MI sessions, each containing 20 trials, and each MI trial lasts 10 s.

BNCI2003_004 (Dornhege et al., 2004): This corresponds to the data from BCI Competition III and contains high-density EEG recordings from 5 healthy subjects. The public classification data provide right-hand and foot MI labels and contain 118 EEG channels; the distributed downsampled version uses 100 Hz, and the task cue lasts 3.5 s.

HefmiIch2025 (Shi et al., 2025): This dataset contains mixed EEG-fNIRS recordings from 17 healthy controls and 20 patients with intracerebral hemorrhage. EEG was acquired from 32 channels at 256 Hz, with a left- and right-hand grasping-imagery task. Each session contains 30 trials, 15 for each hand; a trial consists of a 2-s cue, 10 s of MI, and 15 s of rest.

BNCI2015_001 (Faller et al., 2012): This dataset contains right-hand and both-feet MI recordings from 12 subjects, acquired with 13 EEG channels at 512 Hz. It includes multiple recording sessions.

Kumar2024 (Kumar et al., 2024): This dataset contains left- and right-hand MI recordings from 18 healthy BCI-naive subjects across six sessions held on different dates. EEG was acquired from 22 channels at 512 Hz, with 3 additional EOG channels. The experiment includes offline calibration and online feedback training.

AlexMI (Barachant, 2012): This dataset contains EEG recordings from 8 subjects, with original labels for right-hand MI, both-feet MI, and rest. The signals were acquired using 16 wet electrodes and an amplifier at 512 Hz. Each trial lasts 3 s, and each subject has 20 trials per class.

It should be noted that all MI datasets, including both the training and test datasets, were downloaded from MOABB (Jayaram & Barachant, 2018).

C.3 MI Data Preprocessing

All MI datasets used in this work, including both the Stage-2 pretraining corpus and the downstream test sets, are processed through an identical pipeline to ensure consistent data formatting. The pipeline consists of the following steps.

Channel standardization and interpolation. We adopt a fixed 15-channel template covering frontal, central, temporal, and parietal regions: F7, F3, FZ, F4, F8, T7, C3, CZ, C4, T8, P7, P3, PZ, P4, P8. For each dataset, the original channel names are first normalized by removing common suffixes, converting to uppercase, and mapping legacy names to the standard 10-20 system (e.g., T3→T7, T4→T8, T5→P7, T6→P8). Non-standard names such as 'FCC5h', 'CCP3h', and 'AFF5h' are mapped to their closest standard equivalents. For each target channel that cannot be directly matched, we estimate its signal by inverse-distance-weighted interpolation over the available

channels, using a 2D projection of the 10-20 montage coordinates. This yields a uniform 15-channel representation for every dataset while preserving the spatial topology of the original recordings.

Resampling and filtering. All signals are resampled to 200 Hz. A 4th-order Butterworth band-pass filter (1-40 Hz) is then applied using zero-phase forward-backward filtering ('filtfilt') to avoid phase distortion. No notch filter is applied in this pipeline; power-line interference is expected to be attenuated by the 1-40 Hz band-pass.

Trial alignment. Each trial is aligned to a fixed duration of 4 seconds (800 points). Trials shorter than 4 seconds are padded by mirror reflection of the existing signal, which avoids boundary discontinuities and repeated edge values. Trials longer than 4 seconds are truncated to the first 4 seconds. This ensures that all trials share the same temporal length before patch extraction.

Patch extraction. Each 4-second trial is divided into 1-second patches with a 0.5-second stride, producing 7 overlapping patches per channel. The resulting tensor for one trial has shape $C \times P \times T = 15 \times 7 \times 200$, where C is the number of channels, P is the number of patches, and T is the number of time points per patch. The trial dimension is preserved so that downstream models can operate on either patch-level or trial-level representations.

Label processing. Event labels are unified across datasets using a global label space. Events referring to the same class (e.g., 'both_hand' and 'hands') are merged, and compound events such as 'left_hand_right_foot' and 'right_hand_left_foot' are discarded to avoid ambiguous supervision. After merging and filtering, only five classes are retained: left hand, right hand, both hands, feet, and rest. For each dataset, a local label mapping is additionally generated based on the classes actually present, allowing both global (cross-dataset) and local (dataset-specific) evaluation. The global label space is stored alongside the preprocessed data for reference.

The same pipeline is applied to both the Stage-2 pretraining corpus and the nine downstream MI datasets. This guarantees that the only source of distribution shift between pretraining and evaluation is the underlying data itself, rather than differences in channel layout, sampling rate, or trial length.

C.4 BASELINES

We summarize the baseline models used to evaluate MI EEG decoding performance.

ADFCNN (Tao et al., 2024) is an attention-based dual-scale convolutional network that extracts spectral and spatial EEG features at different temporal scales and combines complementary features through attention for MI classification.

FBCNet (Mane et al., 2021) is a filter-bank convolutional network that represents EEG in multiple frequency bands, learns band-specific spatial filters, and aggregates temporal information through a variance layer to extract discriminative sensorimotor features.

EEGNet (Lawhern et al., 2018) is a compact convolutional network that combines temporal filtering with depth-wise spatial convolutions and separable convolutions to learn EEG features with a small number of parameters.

EEG Conformer (Song et al., 2023) combines convolutional layers for extracting local temporal and spatial EEG features with self-attention for modeling global dependencies among the resulting temporal feature segments.

DeepCNN (Schirrmeister et al., 2017) is a deep convolutional network that learns hierarchical features directly from raw EEG through initial temporal and spatial convolutions followed by additional convolutional and pooling layers.

ShallowCNN (Schirrmeister et al., 2017) is a shallow convolutional network inspired by FBCSP. It uses temporal and spatial filtering, squaring, mean pooling, and logarithmic transformation to extract band-power-related EEG features.

IFNet (Wang et al., 2023) is a lightweight convolutional network that models interactions between different frequency components, combining spectral, spatial, and temporal information to learn discriminative representations for MI decoding.

EDPNet (Han et al., 2024) is a lightweight network that combines adaptive spatial-spectral fusion with multiscale variance pooling. Dual-prototype learning further shapes the feature distribution to improve class discrimination with limited training data.

BIOT (Yang et al., 2023) is a biosignal Transformer that represents channel-wise signal segments as a unified token sequence with channel and positional embeddings, enabling representation learning across datasets with heterogeneous recording configurations.

BrainOmni (Xiao et al., 2025) is a foundation model for both EEG and MEG. It uses a sensor-aware tokenizer to encode sensor properties and quantize brain signals into discrete representations for subsequent self-supervised pretraining.

LaBraM (Jiang et al., 2024) is an EEG foundation model pretrained to predict discrete neural codes for masked channel patches. Its neural tokenizer is trained through vector-quantized neural spectrum prediction to produce the pretraining targets.

EEGPT (Wang et al., 2024) is a pretrained Transformer that learns transferable EEG features through masked self-supervised learning with spatiotemporal representation alignment and signal reconstruction. Its hierarchical architecture processes spatial and temporal information separately.

CBraMod (Wang et al., 2025a) is an EEG foundation model that uses parallel spatial and temporal attention in a criss-cross Transformer. It combines asymmetric conditional positional encoding with masked EEG reconstruction to learn transferable representations.

MIRepNet (Liu et al., 2026b) is a pretrained model tailored to MI. It combines preprocessing based on a unified EEG channel template with masked-token reconstruction and supervised MI classification to learn representations across MI datasets.

C.5 Evaluation Metrics

We report Balanced Accuracy (BA), Cohen's Kappa, and Weighted F1.

BA averages the recall of each class with equal weight to reflect recognition performance across classes.

Kappa measures agreement between predictions and labels after correcting for chance agreement.

F1 aggregates class-level F1 scores according to the proportion of test samples in each class.

D Training Details

In this section, we provide a detailed description of the training configuration for the three stages of NeuroMI: (1) large-scale general EEG pre-training, (2) MI-specific mechanism-guided specialization pre-training, and (3) downstream linear-probe fine-tuning on MI datasets. All experiments were implemented in PyTorch and trained on NVIDIA GPUs.

D.1 Stage 1: General EEG Pre-training

Stage 1 is trained on the TUEG corpus to learn transferable spatio-temporal-spectral representations. We use a patch-level masking strategy with a mask ratio of 0.5, and the reconstruction losses (raw waveform and band energy) are computed only at masked positions. The optimizer is AdamW, and a cosine annealing scheduler with a one-epoch linear warm-up is used.

Table D.1: Hyperparameters for Stage 1 general EEG pre-training.

Hyperparameters	**Settings**
EEG sample	
Channels	15
Sampling rate	200 Hz
Trial duration	30 s
Patch length	1 s

Hyperparameters	Settings
Stride	1 s
Number of patches per trial	30
Mask ratio	0.5
Patch embedding	
Temporal kernel sizes	{15, 31, 63}
Temporal filters per branch	16
Pool size	4
Dropout	0.1
Frequency-aware positional encoding	
Frequency bands	{(4, 8), (8, 13), (13, 30), (30, 50)} Hz
Conv pos encoder kernel	(7, 7), depthwise
Transformer encoder	
Layers	12
Hidden dimension	200
Feed-forward dimension	800
Attention heads	8 (4 spatial, 4 temporal)
Cross-fusion	Enabled
Pre-training	
Epochs	50
Batch size	64
Optimizer	AdamW
Learning rate	1e-4
Weight decay	5e-2
Scheduler	Cosine annealing with 1-epoch warm-up
Minimum learning rate	1e-7
Gradient clipping	0.5
λ_{energy}	0.1

D.2 Stage 2: Mechanism-Guided Specialization Pre-training

Stage 2 loads the Stage 1 checkpoint and specializes the encoder on a curated multi-dataset MI corpus (approximately 300,000 4-second trials aggregated from 17 public MI datasets). We adopt differential learning rates: the patch embedding, encoder, and task heads use scaling factors of 0.5, 0.2, and 1.0 relative to the base learning rate, respectively. The mechanism-guided losses ($\mathcal{L}_{\mu\beta}$, $\mathcal{L}_{ERD}$, $\mathcal{L}_{fbcsp}$) are computed directly from the raw signal on-the-fly. The ERD baseline is extracted from the first 0.5 s of each trial.

Table D.2: Hyperparameters for Stage 2 mechanism-guided specialization pre-training.

Hyperparameters	Settings
EEG sample	
Channels	15
Sampling rate	200 Hz
Trial duration	4 s
Patch length	1 s
Stride	0.5 s
Number of patches per trial	7
Model	
Layers	12
Hidden dimension	200
Feed-forward dimension	800
Attention heads	8 (4 spatial, 4 temporal)
Number of classes	5
FBCSP feature dimension	280
Pre-training	
Epochs	50
Batch size	128

Hyperparameters	Settings
Optimizer	AdamW
Base learning rate	1e-4
LR scale (patch embedding)	0.5
LR scale (encoder)	0.2
LR scale (heads)	1.0
Weight decay	5e-2
Scheduler	Cosine annealing
Minimum learning rate	1e-7
Gradient clipping	0.5
$\lambda_{\mu\beta}, \lambda_{ERD}, \lambda_{fbcsp}$	0.1, 0.1, 0.1
ERD baseline length	0.5 s
μ band / β band	8-13 Hz / 13-30 Hz

D.3 Stage 3: Downstream Fine-tuning

For downstream evaluation, we freeze the Stage-2 encoder and train only a lightweight classifier on the extracted features. The classifier consists of a flatten-pooling layer (mean over patches) followed by a two-layer MLP with ELU activation and dropout. All experiments use within-subject 5-fold cross-validation with an internal 80/20 train-validation split for model selection.

Table D.3: Hyperparameters for downstream linear-probe fine-tuning.

Hyperparameters	Settings
Classifier	
Pooling	Mean over patches
Hidden layers	100
Activation	ELU
Dropout	0.1
Fine-tuning	
Epochs	50
Batch size	64
Optimizer	AdamW
Learning rate (head)	5e-5
Weight decay	1e-4
Scheduler	None
Gradient clipping	1.0
Loss	Cross-entropy
Evaluation	
Cross-validation	Stratified 5-fold (within-subject)
Internal validation split	80% / 20%
Model selection	Best validation balanced accuracy
Metrics	BA, Kappa, F1
Aggregation	Mean over folds, then mean over subjects

All downstream foundation-model baselines are evaluated under the identical linear-probing protocol to ensure a fair comparison.

D.4 Implementation Notes

Masking: A binary mask $M \in \{0,1\}^{B\times C\times P}$ is independently sampled for each sample on the (channel, patch) grid with probability $r = 0.5$. Masked patches are zeroed out before encoding.

ERD targets: Computed as $log(P_{active}) - log(P_{baseline})$ for both μ and β bands, then clamped to $[-10, 10]$ for numerical stability.

FBCSP features: Pre-computed offline with 280-dimensional output and used as regression targets during Stage 2.

Checkpoint selection: In both pre-training stages, the checkpoint with the lowest mean training loss is selected. In downstream fine-tuning, the classifier with the highest validation balanced accuracy is selected.