TAHLP: A TOPOLOGY-AWARE HIERARCHICAL LATENT PREDICTION FOUNDATION MODEL FOR CROSS-ELECTRODE MOTOR IMAGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Motor imagery (MI) is a core electroencephalography (EEG) brain-computer interface paradigm, yet developing a single MI foundation model that transfers across heterogeneous electrode montages remains challenging. Generic EEG foundation models pre-trained on multi-paradigm mixtures may learn representations that do not align well with MI's frequency- and region-specific sensorimotor rhythms. Existing electrode-position encodings often rely on hand-defined channel orders or learn geometry-to-feature mappings that may not fully capture the electrode-coordinate topology. To overcome these challenges, we propose TAHLP, a Topology-Aware Hierarchical Latent Prediction foundation model for cross-electrode MI, pre-trained via hierarchical latent-prediction self-distillation, where a student predicts the multi-level cumulative residuals of an exponential moving average (EMA) teacher's continuous latent features using a weighted cosine loss. This objective is self-supervised and optimized in a single stage, requiring no codebook, negative pairs, or labels, which simplifies optimization and reduces reliance on large labeled datasets for pre-training. To explicitly model electrode-coordinate topology, we introduce interleaved 3D electrode rotary position embedding, a parameter-free rotary transform on adjacent query/key feature-dimension pairs, with angles from each electrode's 3D scalp coordinates. By construction, attention scores depend on relative 3D electrode displacement, not absolute position or channel order/count. Extensive experiments across four public MI benchmarks demonstrate that TAHLP consistently achieves state-of-the-art performance while using less than 8.8% as much pre-training data as generic EEG foundation models do. Code and models will be publicly available.

1 INTRODUCTION

Brain-computer interfaces (BCIs) decode neural activity into control commands, bypassing peripheral nerves and muscles (Lotte et al., 2018). Electroencephalography (EEG) is the modality of choice for deployable BCIs because it is inexpensive, portable, and safe. Within EEG-based BCIs, motor imagery (MI)—the voluntary imagination of a movement without overt execution—is a central paradigm. Unlike evoked responses, MI modulates ongoing sensorimotor rhythms, producing event-related desynchronization and synchronization (ERD/ERS) localized over sensorimotor cortex and confined mainly to the α and β bands (Pfurtscheller & Lopes da Silva, 1999; Pfurtscheller & Neuper, 2001). The effect is low in amplitude, narrowband, non-phase-locked to the cue, and highly variable across subjects and sessions, which makes single-trial decoding difficult and has historically motivated per-subject and per-dataset decoders (Lotte et al., 2018). Yet MI supports stroke rehabilitation, assistive robotics, and embodied control (Lotte et al., 2018; Liu et al., 2026b), motivating one model that transfers across users and recording setups.

Self-supervised pretraining is a natural route to this goal. Generic EEG foundation models (Kostas et al., 2021; Yang et al., 2023; Wang et al., 2024; Jiang et al., 2024a; Wang et al., 2025) pretrain one encoder on unlabeled corpora spanning multiple acquisition protocols and paradigms (Liu et al., 2026a), then fine-tune it lightly. MI, SSVEP, and event-related potentials arise from different cortical generators and carry distinct spectral signatures (Lotte et al., 2018), yet the objective does not favor any paradigm: one set of weights serves all, so capacity follows the mixture's aggregate statistics,

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.

and a broadly predictive representation need not preserve the band-limited, region-specific structure essential for MI decoding (Zhang et al., 2025; Yang et al., 2026). This is not merely theoretical: a recent EEG foundation model benchmark reports that specialist models remain competitive and larger models do not necessarily generalize better (Liu et al., 2026a), and (Liu et al., 2026b) show that an MI-specific model outperforms generic ones on MI benchmarks while using far less pretraining data. MI therefore benefits from a foundation model whose data, objective, and inductive biases are chosen for MI.

A central question is then what such a model should be trained to predict. Any candidate objective has to cope with three properties of real-world MI recordings. First, montage heterogeneity: devices differ in electrode set, channel count, and reference scheme, and the same label may sit at slightly different positions across devices (Liu et al., 2026a). Unless trials from different corpora share one token space, they cannot be pretrained jointly, and every new dataset forces channel selection or interpolation—discarding the electrodes absent from the chosen template and capping the scale pretraining is meant to provide. Second, label scarcity: labeled MI data require a cued protocol, sustained attention, and many trials per participant, and inter-subject variability is large (Lotte et al., 2018). Their volume therefore grows far more slowly than that of unlabeled recordings, which penalizes any objective that consumes labels and makes instance-level discrimination fragile, since trials that should agree often come from different subjects. Third, contamination: the recorded signal is routinely corrupted by ocular, myogenic, and line noise (Lotte et al., 2018), so a target defined at waveform level obliges the model to reproduce nuisance variance alongside the neural content.

Existing self-supervised objectives fall into a few families, each of which runs into at least one of these properties. Reconstruction is the dominant paradigm (Liu et al., 2026a): whether models predict the raw waveform, embedded tokens, or a spectral target, the prediction target is always a transformed copy of the signal, and a sample-level regression term gives nuisance variance the same standing as the narrow band-limited structure MI decoding depends on (Pfurtscheller & Lopes da Silva, 1999; Lotte et al., 2018). Predicting latent representations rather than inputs removes that bias (Baevski et al., 2022; Assran et al., 2023). Discretizing the target into codebook indices (Jiang et al., 2024b; Cui et al., 2026) reduces waveform sensitivity, but buys this robustness at the cost of deciding the target space before the encoder sees any data: the codebook must first be fitted—its own size, commitment weight, and residual depth—and is then frozen, so the encoder is trained against a vocabulary optimized for signal fidelity rather than for the representation to be learned—and one that may collapse or be used unevenly when those choices are mis-set (Liu et al., 2026a). Pretraining thereby becomes multi-stage, and the target resolution is fixed by the tokenizer rather than by the downstream representation. Autoregressive objectives impose a causal token order that need not reflect neurophysiological structure (Liu et al., 2026a), and contrastive objectives (Kostas et al., 2021) depend on augmentation and on how positive and negative pairs are constructed, so trials that share a class but differ in subject or montage are liable to be pushed apart—fragile because labeled within-subject data are scarce. MI-specialized objectives (Liu et al., 2026b) instead pair reconstruction with a supervised classification term, spending the labels that are scarce. Latent prediction, although well established elsewhere (Baevski et al., 2022; Assran et al., 2023), has not been carried over to MI-EEG pre-training: every objective in use here recovers either a transformed copy of the signal or a codebook index, and the closest relative, teacher–student distillation in EEG-X (Foumani et al., 2025), still denoises the waveform. What is missing for MI is a target that is continuous rather than quantized, abstract rather than waveform-level, and trained without negative pairs or class labels.

To address this gap, we propose **TAHLP** (**T**opology-**A**ware **H**ierarchical **L**atent **P**rediction), a fully self-supervised framework coupling a topology-aware cross-electrode encoder with a new pretraining objective, *hierarchical latent-prediction self-distillation*. Following data2vec (Baevski et al., 2022; 2023) and I-JEPA (Assran et al., 2023), TAHLP predicts continuous latents rather than raw signals. An EMA teacher (He et al., 2020; Grill et al., 2020; Caron et al., 2021) encodes the unmasked trial and assigns each token a continuous target: the detached, instance-normalized average of its last K block outputs from the full recording context, supplied by the model itself. The student sees the trial twice under complementary masks, so each token is visible in one view, masked in the other, and predicted once per step. At masked positions, the student refines predictions over several levels: each level adds its head output to the stop-gradient cumulative prediction of the previous level; all levels regress toward the same teacher target under a weighted cosine loss. The objective is single-stage, requiring no fitted codebook, negative pairs, or class labels.

To handle electrode heterogeneity, TAHLP uses a shared temporal encoder to map each (electrode, window) pair to a token, and then applies a Transformer encoder over these tokens, padding different montages onto a common grid within each batch and masking padded slots from attention. Montage information enters the model through two complementary components: a learnable embedding indexed by electrode name, and a parameter-free rotary transform applied to adjacent query/key feature-dimension pairs, whose angles are computed by combining each electrode's 3-D scalp coordinate with the rotation frequency assigned to the corresponding axis. We call this transform *Interleaved 3D Electrode Rotary Position Embedding* (I3D-ERoPE). Inspired by MRoPE-Interleave (Huang et al., 2026), which distributes feature-dimension pairs among positional axes in a fine-grained round-robin manner so that every axis spans the full frequency spectrum, we adapt this interleaved frequency allocation to the three axes of an electrode's 3-D scalp coordinate. This adaptation enables the model to exploit the geometric topology of electrodes without relying on a fixed montage template or channel order.

Our main contributions are as follows:

- **Hierarchical Latent-Prediction Self-Distillation for MI.** We introduce a self-supervised pretraining paradigm in which the model predicts its own continuous latent representations across hierarchical levels, rather than reconstructing raw waveforms or contrasting augmented views. The paradigm is single-stage and requires no separately fitted codebook, negative pairs, or class labels.
- **Topology-Aware Cross-Electrode Modeling.** We propose I3D-ERoPE, a parameter-free rotary transform that injects electrode geometry from 3D scalp coordinates into the Transformer attention mechanism via query/key rotations. This enables topology-aware modeling across heterogeneous montages without template alignment or spatial interpolation.
- **A Generalist Foundation Model for Motor Imagery Decoding.** To the best of our knowledge, we present the first MI foundation model built upon latent-prediction pretraining. It achieves state-of-the-art decoding on all four held-out benchmarks, outperforming both specialized single-dataset decoders and all generalist EEG foundation models, while using less than 8.8% as much pre-training data as the smallest-pretraining-corpus generalist baseline among those we compare against (221 h vs. 2,534 h).

2 METHOD

This section details the overall framework of TAHLP. First, temporal tokenization and electrode-identity encoding (Section 2.1) construct the input token representations. Then, a topology-aware Transformer (Section 2.2) uses I3D-ERoPE to inject each electrode's scalp geometry into the rotation of queries and keys, so that attention depends on the relative position of two electrodes rather than on absolute channel indices. Finally, hierarchical latent-prediction-based self-distillation (Section 2.3), as illustrated in Figure 1, employs an EMA teacher to provide stable targets, while the student predicts them under complementary masks via cumulative residual refinement.

2.1 TEMPORAL TOKENIZATION AND ELECTRODE IDENTITY ENCODING

Let $X \in \mathbb{R}^{C\times T}$ be a preprocessed trial (see Appendix E, F.3 for preprocessing details) with ordered channel names $\mathcal{C} = (\mathrm{ch}_1, \ldots, \mathrm{ch}_C)$, where C and T denote the numbers of channels and time samples, respectively, and both vary across corpora.

Patch Segmentation and Padding. We segment X into non-overlapping windows of length t (in our setting $t = 200$ samples, corresponding to 1 second at 200 Hz) to obtain a set of patches

$$\mathcal{X} = \{x_{i,j} \mid i \in [1, \ldots, C],\ j \in [1, \ldots, A]\}, \qquad x_{i,j} \in \mathbb{R}^t, \qquad A = \left\lfloor \frac{T}{t} \right\rfloor, \tag{1}$$

where A is the number of patches per channel, *i.e.*, the temporal length of that trial. Because different corpora yield different numbers of channels C and patches A, each trial is further padded with zero patches along both axes to the maximum of its batch; in the remainder of the paper we reuse C and A for the dimensions of this common (padded) grid. Binary validity masks $\mathbf{b}^{\mathrm{ch}}$ and $\mathbf{b}^{\mathrm{time}}$ mark the real channels and time patches, so that a token (i, j) is valid if and only if $\mathbf{b}^{\mathrm{ch}}_i = \mathbf{b}^{\mathrm{time}}_j = 1$; padded positions are excluded from the attention of valid tokens (Section 2.2).

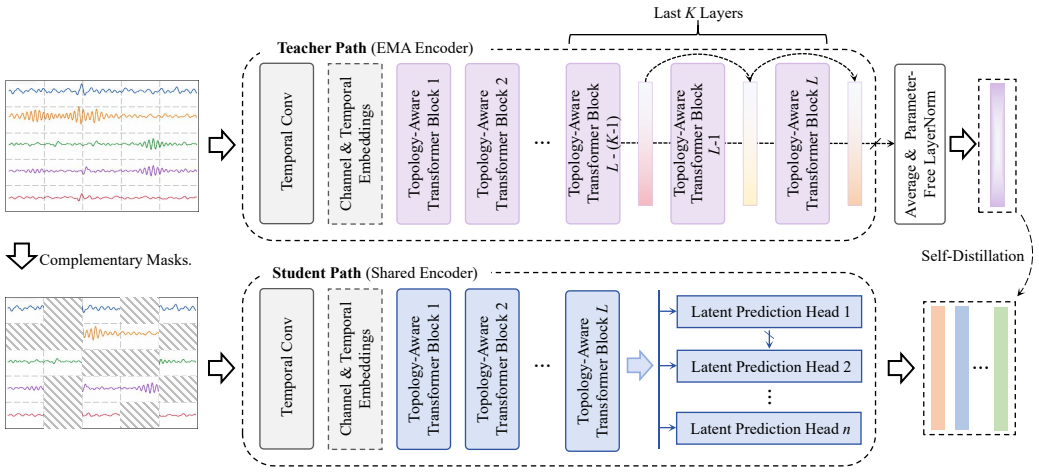

Figure 1: Overview of TAHLP pre-training. An EMA teacher encodes the unmasked EEG and produces a full-context latent target for each token, detached from the graph. The student (topology-aware Transformer blocks with I3D-ERoPE) is applied twice to the *same* EEG under two complementary masks. At the masked positions of each pass, it emits cumulative latent predictions across refinement levels; each level adds a residual to the stop-gradient prediction of the previous level and is regressed to the corresponding teacher targets.

Intra-patch Temporal Conv. A raw patch of t samples is too high-dimensional to serve directly as a token, and tokenization should not depend on the montage. We therefore map every patch to one token with a temporal feature encoder (TFE) shared by all electrodes and all patches: it stacks three blocks, each a temporal convolution followed by group normalization and a GELU activation, and flattens the resulting 8 feature maps of 25 time positions into $D = 200$ dimensions:

$$e_{i,j} = \mathrm{TFE}(x_{i,j}) \in \mathbb{R}^D. \tag{2}$$

All convolutions have unit kernel extent along the electrode–patch axis, so the encoder acts only along time within each patch — no mixing across electrodes or patches, one patch per token — and, being shared over all (i, j), it depends on neither C nor A and applies unchanged to any montage and trial length.

Complementary Masks. From the token embeddings $e_{i,j}$ we draw a binary mask $\mathcal{M} \in \{0,1\}^{C\times A}$ with entries $\mathcal{M}_{i,j} \sim \mathrm{Bernoulli}(r)$, $r = 0.7$ throughout, and replace the selected tokens by a learnable `mask_token`. The substitution occurs only in the student path (Section 2.3) and replaces the token embedding $e_{i,j}$ in Equation 3 below, *before* the channel-identity and temporal embeddings are added; a masked token therefore loses its content but keeps its electrode identity and temporal position. Only valid (non-padded) positions are eligible for masking. To let every token serve as a target, we use two complementary masks $\mathcal{M}$ and $\bar{\mathcal{M}}$ across two passes, so each token is predicted once and, in the other pass, serves as context.

Channel Identity and Temporal Embedding. Let $\mathcal{V}$ be a global electrode vocabulary, and let $v(i) \in \{1, \ldots, |\mathcal{V}|\}$ denote the vocabulary index of the name of channel i. A learned channel-name embedding $e^{\mathrm{ch}}_{v(i)}$ is added to every token belonging to channel i. Then a sinusoidal temporal embedding e^{time}_{j} is added to every token at patch index j. A learnable `[CLS]` token is prepended, and the Transformer input is

$$h^{(0)}_{i,j} = e_{i,j} + e^{\mathrm{ch}}_{v(i)} + e^{\mathrm{time}}_{j}, \qquad h^{(0)}_{\mathrm{cls}} = e^{\mathrm{cls}}. \tag{3}$$

2.2 Topology-Aware Transformer

Tokens form the sequence

$$\{h^{(0)}_{\mathrm{cls}}\} \cup \{h^{(0)}_{i,j}\}_{i\le C,\, j\le A}, \tag{4}$$

whose length $1 + CA$ is trial-dependent (within a batch all trials are zero-padded to a common grid; Section 2.1). We pass the binary validity masks $\mathbf{b}^{\mathrm{ch}}$ and $\mathbf{b}^{\mathrm{time}}$ to the attention layer: a token

(i, j) is valid if and only if $\mathbf{b}_i^{\text{ch}} = \mathbf{b}_j^{\text{time}} = 1$ (the `[CLS]` token is always valid), and attention is admitted only between two valid tokens, so valid tokens aggregate exclusively over valid tokens while padded rows remain self-consistent. Batches may therefore mix corpora with heterogeneous C and A without contaminating attention. The encoder is a pre-norm Transformer of L blocks ($L = 12$ in pretraining) with QK normalization and stochastic depth.

I3D-ERoPE. Absolute channel embeddings identify each electrode by name but do not encode its geometric arrangement. We therefore inject this geometric inductive bias through I3D-ERoPE(Interleaved 3D Electrode Rotary Position Embedding): every electrode is assigned a 3D scalp coordinate $\mathbf{u}_i$ read from the native electrode positions stored with each dataset and normalized to the unit head sphere; the coordinate is purely spatial, depending on the electrode but not on the time patch (the `[CLS]` token has zero coordinate and is not rotated). The query and key of every token are rotated by an angle that depends on this coordinate, while the values are left unrotated. For a head of dimension d_h, the $d_h/2$ frequency planes are assigned to the coordinate axes round-robin, so that the three axes receive nearly equal shares of planes and each spans the full frequency range; the query (and likewise the key) of token (i, j) is rotated in plane c ($c = 0, \ldots, d_h/2 - 1$) by

$$\begin{pmatrix} q'_{2c} \\ q'_{2c+1} \end{pmatrix} = R_c(\mathbf{u}_i) \begin{pmatrix} q_{2c} \\ q_{2c+1} \end{pmatrix}, \qquad R_c(\mathbf{u}) = \begin{pmatrix} \cos(\gamma\, \mathbf{u}[a]\, \omega_c) & -\sin(\gamma\, \mathbf{u}[a]\, \omega_c) \\ \sin(\gamma\, \mathbf{u}[a]\, \omega_c) & \cos(\gamma\, \mathbf{u}[a]\, \omega_c) \end{pmatrix}, \tag{5}$$

with $a = c \bmod d_u$, $d_u = 3$, $\omega_c = \text{base}^{-2c/d_h}$, base $= 10^4$, and γ a coordinate scale. Because RoPE acts multiplicatively, attention scores depend on the relative coordinate offset between tokens:

$$\langle q', k' \rangle = \sum_{c=0}^{d_h/2-1} \begin{pmatrix} q_{2c} \\ q_{2c+1} \end{pmatrix}^{\top} R_c(\mathbf{u}_{i'} - \mathbf{u}_i) \begin{pmatrix} k_{2c} \\ k_{2c+1} \end{pmatrix}, \tag{6}$$

so spatial relationships are expressed directly in the attention rather than baked into absolute indices. Coordinate normalization and the full derivation are in Appendix F.4.

2.3 Hierarchical Latent-Prediction-Based Self-Distillation

We pretrain the encoder in a self-supervised manner, so that it learns general spatiotemporal representations that are not bound to the label space of any single dataset and that transfer across the heterogeneous MI corpora. We adopt a masked-prediction objective with continuous latent targets.

Masked Views and Teacher–Student Targets. Let $I = \{1, \ldots, C\} \times \{1, \ldots, A\}$ be the set of token indices. Each valid token $(i, j) \in I$ is independently masked with probability r as described in Section 2.1, and $\Omega = \{(i, j) : \mathcal{M}_{i,j} = 1\} \subseteq I$ denotes the set of masked tokens. The teacher encodes the *unmasked* signal while the student encodes the masked one, with masked positions replaced by a learnable `mask_token`, and the same construction is applied a second time with the complementary set $I \setminus \Omega$ as the masked positions (Section 2.1). Crucially, the two views are not two independent augmentations: the prediction target is the full-context latent of *the same signal*. When augmentation is used, it is applied once to the shared input *before* masking, so the teacher and student views remain positionally aligned. Masking details are in Appendix B.

Teacher Targets. The teacher is an exponential moving average (EMA) of the student,

$$\theta_{\text{t}} \leftarrow m\, \theta_{\text{t}} + (1 - m)\, \theta_{\text{s}}, \tag{7}$$

where m follows a cosine schedule from 0.996 to 1.0 over the course of pretraining (updated per optimization step). Let $\tilde{h}_{\text{t}}^{(l)}$ denote the teacher's block-l output after removing the `[CLS]` token. The teacher target is the instance-normalized average of the last K block outputs,

$$T_{\text{ar}} = \text{LN}_{\text{inst}}\left(\frac{1}{K} \sum_{l=L-K+1}^{L} \tilde{h}_{\text{t}}^{(l)} \right) \in \mathbb{R}^{N \times D}, \qquad N = CA, \tag{8}$$

where C and A are the numbers of electrodes and time patches, so $N = CA$ is the number of tokens (excluding `[CLS]`), and "ar" stands for averaged representation. Here $\text{LN}_{\text{inst}}(\cdot)$ denotes a parameter-free LayerNorm over the feature dimension applied to each token independently (*i.e.*, instance normalization: zero mean and unit variance per token); the same operator is reused in the conditioning path below. Averaging the top K blocks and applying this normalization yields a stable

target that can be predicted from partial context *without a discrete codebook*. The teacher target T_{ar} is detached from the teacher computation graph. In pretraining we use $K = 6$.

Hierarchical Cumulative Residual Refinement. Rather than predicting T_{ar} in a single step, the student refines it over n levels. For a branch whose prediction positions are $P \subseteq I$, let $z_P \in \mathbb{R}^{|P|\times D}$ be the output of the student's final block at those positions. With $S_{-1} = \mathbf{0}$ and $c_0 = \mathbf{0}$, the student computes, for $\kappa = 0, \ldots, n-1$,

$$z^{(\kappa)} = \mathrm{Norm}^{(\kappa)}(z_P + c_\kappa)\,, \qquad \Delta_\kappa = g_\kappa\Big(z^{(\kappa)}\Big)\,, \tag{9}$$

$$S_\kappa = \mathrm{sg}(S_{\kappa-1}) + \Delta_\kappa, \qquad c_{\kappa+1} = c_\kappa + \phi_\kappa(\mathrm{LN}_{\mathrm{inst}}(\mathrm{sg}(S_\kappa)))\,, \tag{10}$$

where the update of $c_{\kappa+1}$ in Equation 10 is applied for $\kappa = 0, \ldots, n-2$. Here $c_\kappa \in \mathbb{R}^{|P|\times D}$ is the level-κ conditioning state, $S_\kappa \in \mathbb{R}^{|P|\times D}$ the cumulative prediction after level κ, g_κ the level-κ predictor head, $\phi_\kappa : \mathbb{R}^D \to \mathbb{R}^D$ a conditioning projection applied to each prediction position independently, $\mathrm{Norm}^{(\kappa)}$ a level-specific LayerNorm, with learnable parameters for $\kappa > 0$ and $\mathrm{Norm}^{(0)}$ the identity, and $\mathrm{sg}(\cdot)$ stop-gradient; $g_\kappa, \phi_\kappa, \mathrm{Norm}^{(\kappa)}$ have separate parameters for each level. Conditioning is therefore per token: c_κ and z_P share the same shape and are combined element-wise. Each g_κ is a two-layer MLP bottleneck $D \to D/2 \to D$ (BYOL (Grill et al., 2020)-style, to prevent collapse); we set $n = 3$. Level κ is conditioned on c_κ, which is built from the cumulative predictions up to $S_{\kappa-1}$, each with stop-gradient, so the conditioning path cannot leak target information; deeper levels must therefore infer the residual $T_{\mathrm{ar}}[P] - S_{\kappa-1}$ rather than recover it through ϕ_κ.

Symmetric Branch and Objective. We define the two branches by $P_1 = \Omega$ and $P_2 = I \setminus \Omega$. In branch $b \in \{1, 2\}$, tokens in P_b are replaced by `mask_token` and predicted, while tokens outside P_b remain visible. Both branches share the same student parameters and regress to the same detached teacher target T_{ar}. For a set of prediction positions $P \subseteq I$ and matrices $U, V \in \mathbb{R}^{|P|\times D}$ whose rows are indexed by $p \in P$, we define the position-averaged cosine loss

$$\ell(U, V) = \frac{1}{|P|}\sum_{p\in P}(1 - \cos(U_p, V_p))\,, \qquad \cos(\alpha, \beta) = \frac{\alpha^\top \beta}{\|\alpha\|_2\|\beta\|_2}. \tag{11}$$

The overall loss sums over levels with unnormalized weights w_κ and over both branches:

$$\mathcal{L} = \sum_{\kappa=0}^{n-1} w_\kappa\Big[\ell\big(S_\kappa^{(1)},\, T_{\mathrm{ar}}[\Omega]\big) + \ell\big(S_\kappa^{(2)},\, T_{\mathrm{ar}}[I \setminus \Omega]\big)\Big], \tag{12}$$

where $S_\kappa^{(b)} \in \mathbb{R}^{|P_b|\times D}$ is the branch-b cumulative prediction and $[\cdot]$ selects rows by token indices under a fixed canonical order of I, so that the rows of $S_\kappa^{(b)}$ and of $T_{\mathrm{ar}}[\cdot]$ are aligned position by position; we use unnormalized level weights $w = [1, 0.5, 0.25]$. The symmetric pass increases the number of predicted tokens from $|\Omega|$ to all N tokens per step and, because every token serves as context in one branch and as a target in the other, prevents the degenerate solution in which the encoder ignores masked positions.

3 Experiments

3.1 Pre-training

Data. We pre-train on a multi-source MI-EEG repository aggregating **22** publicly released datasets (Table 4, Appendix D): **559** subjects—healthy participants and clinical cohorts with stroke, intracerebral hemorrhage and knee pathology—**207 887** trials and ≈**221 h** of recording. The corpus is heterogeneous along every axis a deployed system faces: sampling rates **200–2 500 Hz**, retained channel counts **14–62** over 10–20, 10–05, Biosemi-32 and g.Nautilus montages, and paradigms from binary hand-laterality to **11-class** whole-body imagery, in single-session and longitudinal designs (**6** datasets, up to 11 sessions with online neurofeedback).

Setup. The backbone is a 12-layer Transformer encoder (hidden dimension 200, 10 heads, 6.2 M parameters) fed by a shared temporal convolution stem that maps each 1 s window of an electrode to a single token of dimension 200; heterogeneous montages are handled by an electrode-name

embedding table, I3D-ERoPE over normalized 3D scalp coordinates and a dynamic-padding attention mask, so no recording is interpolated onto a reference montage. Each dataset is standardized independently to 200 Hz and μV with a common average reference, a 0.5–40 Hz band-pass and 50/60 Hz notches (Appendix E). We pre-train for 50 epochs at batch size 64 with AdamW (peak learning rate 5×10^{-4}, cosine decay with 2% warm-up, weight decay 0.05) on 8×A800 for about 19 h wall-clock. Monitoring runs, masking, augmentation and the full hyper-parameter list are in Appendices B and F.

3.2 Downstream Setup

Datasets. We fine-tune the pretrained encoder on four held-out benchmarks under a strictly subject-independent protocol: **PhysioNet-MI** (Schalk et al., 2004) (4-class, 109 subjects, 64 channels), **SHU-MI** (Ma et al., 2022) (2-class, 25 subjects, 32 channels), **BCICIV-2a** (Tangermann et al., 2012) (4-class, 9 subjects, 22 channels) and **BNCI2015001** (Faller et al., 2012) (2-class, 12 subjects, 13 channels); details in Table 6 (Appendix F.5).

Metrics. We report Balanced Accuracy, Cohen's κ, and Weighted F1 for the 4-class tasks, and Balanced Accuracy, AUC-PR, and AUROC for the binary task. All numbers are means over five seeds, in percent, with the standard deviation in subscript. Model selection uses κ for the 4-class tasks and Balanced Accuracy for the binary task.

Baselines. We compare against specialist decoders trained from scratch on each benchmark—EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2021), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022) and ST-Transformer (Song et al., 2021)—and against generalist EEG foundation models fine-tuned from their released checkpoints: BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024b), CBraMod (Wang et al., 2025), CSBrain (Zhou et al., 2025) and BrainRVQ (Cui et al., 2026). Baselines that require a fixed montage receive the interpolated input prescribed by their own implementation, which is precisely the constraint our method removes.

To compare the pre-training objective and the electrode encoding under a matched data budget, we additionally re-pre-train every generalist baseline from its own published recipe *on our 221 h MI corpus* and fine-tune it under the identical downstream protocol († rows; Appendix G, Tables 7 and 8). This holds the pre-training corpus fixed across methods and isolates the two design choices we care about.

3.3 Main Results

Table 1 compares TAHLP with all baselines on the three benchmarks with published results under our protocol. All baseline numbers come from their original reports, which share our preprocessing, subject-level splits and five-seed protocol; where sources disagree, we use the value from the paper that reproduced the baseline itself (*e.g.*, CBraMod's FFCL AUROC 60.14 rather than CSBrain's 63.26, which exceeds FFCL's own AUC-PR by 3.8 points). Three observations follow.

(i) TAHLP sets a new state of the art on all three benchmarks of Table 1 and leads every column. It gains 2.10 Balanced Accuracy points on PhysioNet-MI (66.30 vs. 64.20), 0.86 on SHU-MI (65.03 vs. 64.17) and 0.58 on BCICIV-2a (57.15 vs. 56.57), measured against the strongest baseline on each benchmark, using only 221 h of pre-training data—under 8.8% of the 2,534 h used by the generalist baseline with the smallest pre-training corpus, and under 0.4% of the 58,021 h used by the one with the largest (Table 3). On the fourth benchmark, BNCI2015001, where only generalist baselines report numbers under our protocol, TAHLP likewise leads every listed method (74.94 vs. 73.88; Table 7).

(ii) Holding the corpus fixed preserves the ordering and widens the margin. Under the matched 221 h budget the six re-pre-trained models no longer move in one direction: all six lose accuracy on the two binary benchmarks, while on BCICIV-2a—whose released checkpoints saw little motor imagery—five of six improve and only CSBrain gives back 2.67 Balanced Accuracy points. Their ordering is broadly unchanged, and measured against the strongest generalist within each block TAHLP's margin *widens* on all four benchmarks: +1.42 vs. +0.86 on SHU-MI, +2.56 vs. +0.58 on BCICIV-2a, +2.75 vs. +2.10 on PhysioNet-MI and +1.52 vs. +1.06 on BNCI2015001. The

Table 1: Main results on three held-out motor-imagery benchmarks. Best value per column in **bold**, second best underlined.

Method	PhysioNet-MI (4-class)			SHU-MI (2-class)			BCICIV-2a (4-class)		
	Balanced Acc.	Cohen's κ	Weighted F1	Balanced Acc.	AUC-PR	AUROC	Balanced Acc.	Cohen's κ	Weighted F1
Specialist models (trained from scratch on the target dataset)									
EEGNet	$58.14_{\pm 1.25}$	$44.68_{\pm 1.99}$	$57.96_{\pm 1.15}$	$58.89_{\pm 1.77}$	$63.11_{\pm 1.42}$	$62.83_{\pm 1.52}$	$44.82_{\pm 0.94}$	$26.93_{\pm 1.21}$	$42.26_{\pm 1.08}$
EEGConformer	$60.49_{\pm 1.04}$	$47.36_{\pm 1.71}$	$60.62_{\pm 0.95}$	$59.00_{\pm 1.07}$	$63.70_{\pm 0.93}$	$63.51_{\pm 1.01}$	$46.96_{\pm 1.06}$	$29.24_{\pm 1.41}$	$45.33_{\pm 1.28}$
SPaRCNet	$59.32_{\pm 1.52}$	$45.64_{\pm 2.34}$	$59.37_{\pm 1.47}$	$59.78_{\pm 0.97}$	$65.10_{\pm 0.62}$	$64.31_{\pm 0.82}$	$46.35_{\pm 1.17}$	$28.47_{\pm 1.47}$	$44.32_{\pm 1.26}$
ContraWR	$58.92_{\pm 1.33}$	$45.27_{\pm 2.48}$	$59.18_{\pm 1.16}$	$58.73_{\pm 1.28}$	$63.15_{\pm 1.05}$	$62.73_{\pm 1.13}$	$46.78_{\pm 1.25}$	$29.05_{\pm 1.60}$	$44.13_{\pm 1.42}$
CNN-Transformer	$60.53_{\pm 1.18}$	$47.25_{\pm 2.23}$	$60.41_{\pm 1.05}$	$59.75_{\pm 1.69}$	$64.12_{\pm 0.76}$	$63.23_{\pm 0.82}$	$46.00_{\pm 1.08}$	$28.00_{\pm 1.48}$	$44.60_{\pm 1.14}$
FFCL	$57.26_{\pm 0.92}$	$43.23_{\pm 1.82}$	$57.01_{\pm 0.79}$	$56.92_{\pm 2.52}$	$59.43_{\pm 1.72}$	$60.14_{\pm 1.68}$	$44.70_{\pm 1.43}$	$26.27_{\pm 1.76}$	$42.38_{\pm 1.39}$
ST-Transformer	$60.35_{\pm 0.81}$	$47.12_{\pm 1.99}$	$60.53_{\pm 0.75}$	$59.92_{\pm 2.06}$	$63.94_{\pm 1.22}$	$64.31_{\pm 1.11}$	$45.75_{\pm 1.45}$	$27.33_{\pm 1.98}$	$44.71_{\pm 1.42}$
Generalist EEG foundation models (released checkpoints)									
BENDR	$42.20_{\pm 2.20}$	$40.10_{\pm 2.50}$	$40.80_{\pm 2.40}$	$55.70_{\pm 2.30}$	$58.50_{\pm 2.70}$	$58.60_{\pm 2.80}$	$49.00_{\pm 0.70}$	$32.00_{\pm 0.90}$	$48.40_{\pm 0.70}$
BIOT	$61.53_{\pm 1.54}$	$48.75_{\pm 2.72}$	$61.58_{\pm 1.97}$	$61.79_{\pm 1.83}$	$67.70_{\pm 1.19}$	$66.09_{\pm 1.27}$	$47.48_{\pm 0.93}$	$29.97_{\pm 1.39}$	$46.07_{\pm 1.25}$
LaBraM	$61.73_{\pm 1.22}$	$49.12_{\pm 1.92}$	$61.77_{\pm 1.41}$	$61.66_{\pm 1.92}$	$67.61_{\pm 0.83}$	$66.04_{\pm 0.91}$	$48.69_{\pm 0.85}$	$31.59_{\pm 1.54}$	$47.58_{\pm 1.03}$
CSBrain	$63.04_{\pm 0.90}$	$50.71_{\pm 1.20}$	$63.08_{\pm 0.95}$	$64.17_{\pm 0.37}$	$72.30_{\pm 0.79}$	$72.00_{\pm 0.58}$	$56.57_{\pm 0.71}$	$42.09_{\pm 0.93}$	$56.37_{\pm 0.87}$
CBraMod	$64.17_{\pm 0.91}$	$52.22_{\pm 1.69}$	$64.27_{\pm 1.00}$	$63.70_{\pm 1.51}$	$71.39_{\pm 0.88}$	$69.88_{\pm 0.68}$	$51.38_{\pm 0.66}$	$35.18_{\pm 0.94}$	$49.84_{\pm 0.85}$
BrainRVQ	$64.20_{\pm 0.80}$	$52.30_{\pm 1.50}$	$64.40_{\pm 0.90}$	$63.70_{\pm 1.80}$	$72.30_{\pm 1.10}$	$71.70_{\pm 0.70}$	$54.10_{\pm 0.80}$	$38.80_{\pm 0.80}$	$53.30_{\pm 1.20}$
TAHLP (ours)	$\mathbf{66.30}_{\pm 0.86}$	$\mathbf{55.07}_{\pm 1.42}$	$\mathbf{66.50}_{\pm 0.94}$	$\mathbf{65.03}_{\pm 1.92}$	$\mathbf{73.60}_{\pm 1.28}$	$\mathbf{73.24}_{\pm 0.92}$	$\mathbf{57.15}_{\pm 0.71}$	$\mathbf{42.87}_{\pm 0.76}$	$\mathbf{56.76}_{\pm 1.05}$

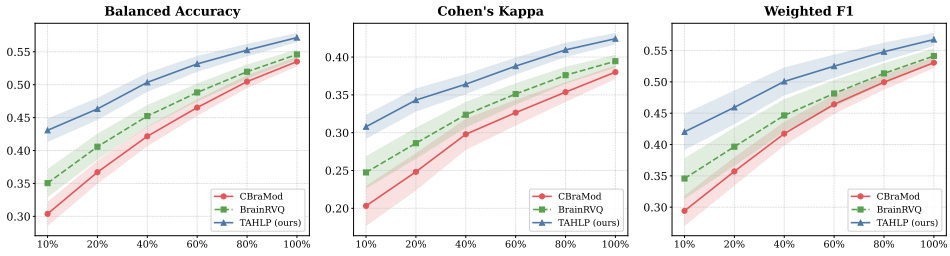

Figure 2: Downstream performance on BCICIV-2a versus the pre-training data budget: TAHLP scales best of the three and matches or surpasses CBraMod and BrainRVQ at every budget.

gain thus comes from the objective and electrode encoding, not the corpus. Why baselines move differently and why this control likely *under*-states them: see Appendix G.

(iii) Electrode identity tracks transfer; the margin co-varies with subject pool and channel count. The three strongest generalists (CBraMod, BrainRVQ, CSBrain) are exactly those that preserve electrode identity or an anatomical grouping—CBraMod's convolution over the channel axis, BrainRVQ's channel-identity tokens, CSBrain's region partition—whereas LaBraM, BIOT and BENDR bind channels to a fixed montage, a channel index or a learned token order (Table 3). Probing confirms that I3D-ERoPE yields electrode-topology awareness (Appendix C). The margin is largest on the largest subject pool ($+2.10$ at 109 subjects, against $+1.06$ at 12, $+0.86$ at 25 and $+0.58$ at 9) but is not monotone in it, and subject pool and channel count are confounded in our benchmarks (64, 32, 22 and 13 channels respectively), so we report this as a correlation only and do not attribute it to either factor. Data efficiency is quantified in Section 3.4.

3.4 Pre-training Data Efficiency

We vary the pre-training budget of TAHLP, CBraMod and BrainRVQ from 10% to 100% of our corpus and fine-tune each checkpoint on BCICIV-2a under the subject-independent split with five seeds; Figure 2 plots Balanced Accuracy, Cohen's κ and Weighted F1 against the budget. TAHLP leads both baselines at every budget and on every metric, and the gap is widest in the low-data regime: already at a 10% budget it exceeds both baselines trained at the same budget by a clear margin. Encoding electrode geometry with a zero-parameter rotary transform therefore buys data efficiency, not only final accuracy.

Table 2: Ablation path from a masked-reconstruction baseline to TAHLP. **SD**, latent-prediction self-distillation; **ECE**, electrode-coordinate encoding; **HC**, hierarchical cumulative prediction; **Mask**, single or complementary dual-mask pass.

Variant	Pre-training	ECE	HC	Mask	PhysioNet-MI	SHU-MI	BCICIV-2a
Baseline	Reconstruction	NONE	✗	Single	$59.63_{\pm 1.12}$	$58.36_{\pm 1.27}$	$46.40_{\pm 1.34}$
SD	Self-distillation	NONE	✗	Single	$62.63_{\pm 0.98}$	$61.36_{\pm 1.40}$	$50.70_{\pm 1.52}$
SD + REVE	Self-distillation	REVE	✗	Single	$63.13_{\pm 1.24}$	$61.86_{\pm 1.45}$	$51.55_{\pm 1.18}$
SD + I3D-ERoPE	Self-distillation	I3D-ERoPE	✗	Single	$64.33_{\pm 1.05}$	$63.06_{\pm 1.50}$	$53.10_{\pm 1.67}$
SD + I3D-ERoPE + HC	Self-distillation	I3D-ERoPE	✓	Single	$65.53_{\pm 0.93}$	$64.26_{\pm 1.60}$	$55.75_{\pm 1.43}$
Full (TAHLP)	Self-distillation	I3D-ERoPE	✓	Dual	$\mathbf{66.30}_{\pm \mathbf{0.86}}$	$\mathbf{65.03}_{\pm \mathbf{1.92}}$	$\mathbf{57.15}_{\pm \mathbf{0.71}}$

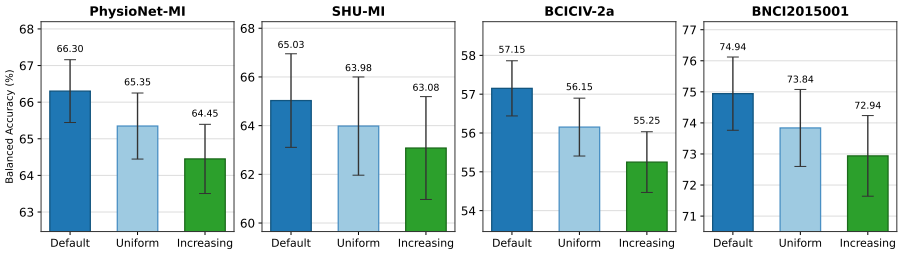

Figure 3: Ablation of the level weights w_κ in Equation 12: Balanced Accuracy over five seeds on each of the four held-out benchmarks.

3.5 Ablation Study

Component ablation. Table 2 adds one component at a time to a masked-reconstruction baseline; its last row is the TAHLP row of Table 1. Throughout, deltas are given in the column order of Table 2, PhysioNet-MI / SHU-MI / BCICIV-2a. *(i) The latent-prediction objective is the largest single change*: replacing masked waveform reconstruction with teacher–student self-distillation over continuous latents gains 3.00 / 3.00 / 4.30 Balanced Accuracy points. *(ii) Geometry helps beyond identity*: every row already carries the electrode-name embedding, and on top of it a learnable coordinate embedding in the style of REVE (El Ouahidi et al., 2025) adds 0.50 / 0.50 / 0.85, whereas the zero-parameter I3D-ERoPE adds 1.70 / 1.70 / 2.40—a strictly larger gain on every benchmark from a transform with no learned parameters, and more than double. *(iii) Hierarchy and the dual-mask pass are cumulative*: multi-level prediction adds 1.20 / 1.20 / 2.65 and the complementary dual-mask pass adds a further 0.77 / 0.77 / 1.40; on BCICIV-2a the hierarchical levels are the largest contribution after the objective itself.

Level weights in the hierarchical loss. Equation 12 leaves the level schedule open. Figure 3 compares TAHLP's default decreasing schedule $w = [1, 0.5, 0.25]$ with uniform $w = [0.5, 0.5, 0.5]$ and increasing $w = [0.25, 0.5, 1]$ under fixed hyper-parameters, reporting Balanced Accuracy. All four benchmarks give the same ordering—decreasing best, uniform about one point lower, increasing about two—because level 1 must predict from context alone, so weighting it highest anchors transferable features, while heavier deeper levels only reward residual corrections that are easy in pre-training but transfer worse.

4 Conclusion

We presented TAHLP, a self-supervised foundation model for motor imagery built on two choices: predict the model's own continuous latents instead of the waveform, and let electrode geometry enter attention as a relative displacement instead of a channel index. Hierarchical latent-prediction self-distillation gives a single-stage objective with no codebook, no negative pairs and no labels, and I3D-ERoPE is a parameter-free rotary transform that applies unchanged to an electrode never seen in pre-training. Pre-trained on 221 h of motor imagery from 22 heterogeneous public datasets, TAHLP sets a new state of the art on all four benchmarks, whose strongest competitors use one to two orders of magnitude more pre-training data. For motor imagery, what a model is trained to predict matters at least as much as how much data it predicts it from.

AI USE STATEMENT

We used generative AI tools solely for language polishing of author-written text. All AI-assisted text was reviewed and verified by the authors to preserve the original meaning and technical content. We take responsibility for the final content of this work, including text and claims produced with the aid of generative AI.

ETHICS STATEMENT

The experiments in this paper rely exclusively on pre-existing EEG datasets that are publicly available and de-identified. No new human-subject data were collected, and no personally identifiable or sensitive information was accessed. Because the study is limited to secondary analyses of widely used public benchmarks, additional ethics approval and informed consent were not required.

REFERENCES

Kai Keng Ang, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 2390–2397, 2008. 15

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 15

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: a general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, pp. 1298–1312, 2022. 2, 15

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 1416–1429, 2023. 2

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. 2, 15

Zhige Chen, Chengxuan Qin, Wenlong You, Rui Liu, Congying Chu, Rui Yang, et al. HEAR: an EEG foundation model with heterogeneous electrode adaptive representation. *arXiv preprint arXiv:2510.12515*, 2025. 15

Mingzhe Cui, Tao Chen, Yang Jiao, Yiqin Wang, Lei Xie, Yi Pan, and Luca Mainardi. BrainRVQ: a high-fidelity EEG foundation model via dual-domain residual quantization and hierarchical autoregression. *arXiv preprint arXiv:2602.16951*, 2026. 2, 7, 15, 16, 23

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: efficient and topology-agnostic foundation model for EEG signal analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 15

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 (NeurIPS)*, 2025. 9, 15

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. 7, 23

D. Forenzo, B. Zhu, and B. He. Forenzo2023 motor-imagery dataset. Distributed in MOABB, 2023. https://moabb.neurotechx.com/docs/datasets.html. 21

Navid Mohammadi Foumani, Soheila Ghane, Nam Nguyen, Mahsa Salehi, Geoffrey I. Webb, and Geoffrey Mackellar. EEG-X: device-agnostic and noise-robust foundation model for EEG. *arXiv preprint arXiv:2511.08861*, 2025. 2, 15

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: a new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 2, 6, 15

Haokun Gui, Xiucheng Li, and Xinyang Chen. Vector quantization pretraining for EEG time series with random projection and phase alignment. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 16731–16750, 2024. 15

Guttmann-Flury et al. GuttmannFlury2025 motor-imagery dataset. Distributed in MOABB, 2025. https://moabb.neurotechx.com/docs/datasets.html. 21

Danny Dongyeop Han, Ahhyun Lucy Lee, Taeyang Lee, Yonghyeon Gwon, Sebin Lee, Seong Jin Lee, David Keetae Park, Shinjae Yoo, Chun Kee Chung, and Jiook Cha. DIVER-0: a fully channel equivariant EEG foundation model. In *ICML 2025 Workshop on Generative AI and Biology (GenBio)*, 2025. 15

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: a Euclidean space data alignment approach. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28 (2):399–410, 2020. 20

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020. 2

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022. 15

Hefmi Ich study group. HefmiIch2025 motor-imagery dataset (hybrid EEG–fNIRS). Distributed in MOABB, 2025. https://moabb.neurotechx.com/docs/datasets.html. 21

Jie Huang, Xuejing Liu, Sibo Song, RuiBing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision–language models. In *The Fourteenth International Conference on Learning Representations*, 2026. 3, 15

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. 21

Jeong et al. Jeong2020-MI 11-class motor-imagery dataset. Distributed in MOABB, 2020. https://moabb.neurotechx.com/docs/datasets.html. 21

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: a universal multitask foundation model for bridging the gap between language and EEG signals. In *The Thirteenth International Conference on Learning Representations*, 2025. 15

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024a. 1

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024b. 2, 7, 15, 16, 25

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100 (17):e1750–e1762, 2023. 7

Mustafa Kaya, Mustafa K. Binli, Erdem Ozbay, Hilmi Yanar, and Yaroslav Mishchenko. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific Data*, 5:180211, 2018. 21

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. 1, 2, 7, 15

Kumar et al. Kumar2024 motor-imagery dataset. Distributed in MOABB, 2024. https://moabb.neurotechx.com/docs/datasets.html. 21

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. 7

Min-Ho Lee, O-Yeon Kwon, Yong-Jeong Kim, Hong-Kyung Kim, Young-Eun Lee, John Williamson, and Seong-Whan Lee. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5):giz002, 2019. 21

Hongli Li, Man Ding, Rong-Hua Zhang, and Chunbo Xiu. Motor imagery EEG classification algorithm based on CNN-LSTM feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022. 7

Dingkun Liu, Zhu Chen, and Dongrui Wu. CLEAN-MI: a scalable and efficient pipeline for constructing high-quality neurodata in motor imagery paradigm. *arXiv preprint arXiv:2506.11830*, 2025. 15

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. EEG-FM-Compass: progress, benchmarking, and future directions for EEG foundation models. *National Science Review*, 2026a. 1, 2, 15, 16

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. 1, 2, 15

Liu et al. Liu2024 stroke motor-imagery dataset. Distributed in MOABB, 2024. https://moabb.neurotechx.com/docs/datasets.html. 21

Liu et al. Liu2025 gait-imagery dataset. Distributed in MOABB, 2025. https://moabb.neurotechx.com/docs/datasets.html. 21

Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018. 1, 2, 15

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022. 7, 23

Patrick Ofner, Andreas Schwarz, Joana Pereira, and Gernot R. Müller-Putz. Upper limb movements can be decoded from the time-domain of low-frequency EEG. *PLoS ONE*, 12(8):e0182578, 2017. 21

Wei Yan Peh, Yu Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp EEG. *44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602, 2022. 7

Gert Pfurtscheller and Fernando H. Lopes da Silva. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857, 1999. 1, 2, 15

Gert Pfurtscheller and Christa Neuper. Motor imagery and direct brain-computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001. doi: 10.1109/5.939829. 1

Rozado et al. Rozado2015 motor-imagery dataset. Distributed in MOABB, 2015. https://moabb.neurotechx.com/docs/datasets.html. 21

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. 7, 23

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. 15

Jaeyoung Shin, Alexander von Lühmann, Benjamin Blankertz, Dong-Wook Kim, Jeonghun 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): 1695–1705, 2017. 21

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170*, 2021. 7

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022. 7

Jonathan R. Stieger, Scott A. Engel, and Bin He. Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Science Advances*, 7(45):eabi6264, 2021. 21

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 15

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, et al. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012. 7, 23

Tavakolan et al. Tavakolan2017 motor-imagery dataset. Distributed in MOABB, 2017. https://moabb.neurotechx.com/docs/datasets.html. 21

Triana-Guzman et al. TrianaGuzman2024 sit-to-stand imagery dataset. Distributed in MOABB, 2024. https://moabb.neurotechx.com/docs/datasets.html. 21

unknown. Motor-imagery EEG dataset (Zenodo record 5882500). Zenodo; distributed in MOABB, 2022. https://zenodo.org/records/5882500. 21

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 21

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 (NeurIPS)*, 2024. 1, 15

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: a criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 7, 15

Dongrui Wu. Revisiting Euclidean alignment for transfer learning in EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 22:031005, 2025. 20

Chaoqi Yang, Danica Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised EEG representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021. 7

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 (NeurIPS)*, 2023. 1, 7, 15

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5Xwm8e6vbh. 2

Wenchao Yang, Weidong Yan, Wenkang Liu, Yulan Ma, and Yang Li. THD-BAR: topology hierarchical derived brain autoregressive modeling for EEG generic representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 15

Florian Yger, Maxime Berar, and Fabien Lotte. Riemannian approaches in brain-computer interfaces: a review. *IEEE Transactions on Biomedical Engineering*, 64(6):1753–1762, 2017. 15

Weibo Yi, Shuang Qiu, Hongzhi Qi, Lixin Zhang, Baikun Wan, and Dong Ming. EEG feature comparison and classification of simple and compound limb motor imagery. *Journal of NeuroEngineering and Rehabilitation*, 11:106, 2014. 21

Yi et al. Yi2025 upper-limb imagery dataset. Distributed in MOABB, 2025. https://moabb.neurotechx.com/docs/datasets.html. 21

Yifan Zhang, Yang Yu, Hao Li, Anqi Wu, Xin Chen, Jin-Fu Liu, Ling-Li Zeng, and Dewen Hu. DMAE-EEG: a pretraining framework for EEG spatiotemporal representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36:17664–17678, 2025. 2

Baozeng Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojun Guo. A fully automated trial selection method for optimization of motor imagery based brain-computer interface. *PLoS ONE*, 11(8): e0162657, 2016. 21

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. CSBrain: a cross-scale spatiotemporal brain foundation model for EEG decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 7, 15, 23

Zhou et al. Zhou2020 motor-imagery dataset. Distributed in MOABB, 2020. https://moabb.neurotechx.com/docs/datasets.html. 21

Zuo et al. Zuo2025 leg-imagery dataset. Distributed in MOABB, 2025. https://moabb.neurotechx.com/docs/datasets.html. 21

A Related Work

Electrode-position encoding for heterogeneous montages. A survey of EEG foundation models (Liu et al., 2026a) organizes the architecture axis by how a model unifies heterogeneous electrode sets; the distinction that matters for transfer is whether the encoding is *absolute*—a channel index (Kostas et al., 2021), a template coordinate (Jiang et al., 2024b; Liu et al., 2026b), a convolution over the channel-index axis (Wang et al., 2025) or an anatomical region partition (Zhou et al., 2025)—or whether it expresses the *displacement* between two electrodes. Absolute encodings presuppose a canonical ordering: anatomically adjacent electrodes can be far apart in the index, an unseen electrode has no position until it is remapped, and every new dataset pays a remapping cost. Channel-wise tokenization (Yang et al., 2023) and projection into a fixed latent space (Döner et al., 2025) sidestep the ordering but discard geometry altogether. Coordinate-aware models take 3D positions as input—Fourier features (El Ouahidi et al., 2025), learned coordinate embeddings used as attention biases (Chen et al., 2025), or a binary same-channel bias on top of temporal RoPE (Han et al., 2025)—but these remain absolute: the attention score depends on each electrode's own coordinate, not on the displacement between two electrodes, and a learned embedding spends parameters on a coordinate-to-feature map that can only be trusted where coordinates were seen during pre-training. TAHLP instead applies a *zero-parameter* rotary transform, so the attention score is a function of the relative 3D displacement (Equation 6) and generalises by construction to any electrode with a known scalp coordinate; we adapt the interleaved frequency allocation of multimodal RoPE (Su et al., 2024; Huang et al., 2026) to three spatial axes.

Pre-training objectives. Objectives differ mainly in the prediction target. *Raw-signal reconstruction* regresses masked waveform patches (He et al., 2022; Wang et al., 2024; 2025); it is a strong baseline but makes nuisance variance as important as the narrowband structure MI decoding depends on (Pfurtscheller & Lopes da Silva, 1999; Lotte et al., 2018). *Embedded-token reconstruction* moves the target into a tokenizer's latent space (Kostas et al., 2021; Yang et al., 2023; Döner et al., 2025), which reduces amplitude sensitivity but keeps the target a transformed copy of the input. *Codebook-based* objectives discretize it into neural tokens (Jiang et al., 2024b; Gui et al., 2024; Cui et al., 2026; Chen et al., 2025); discretization suppresses low-level noise but makes pre-training multi-stage—the codebook is fitted first and then frozen as the target space—and it can collapse or be used unevenly when its size and commitment weight are mis-set (Liu et al., 2026a). *Autoregressive* objectives impose a causal token order (Jiang et al., 2025; Yang et al., 2025) that need not reflect neurophysiological structure, and *contrastive* objectives (Kostas et al., 2021) depend on how positives and negatives are constructed, so trials sharing a class but differing in subject or montage are liable to be pushed apart. *Latent-prediction self-distillation*, the family TAHLP belongs to, replaces reconstruction and negatives with a teacher–student objective over continuous features (Baevski et al., 2022; Assran et al., 2023; Grill et al., 2020; Caron et al., 2021); the closest relative inside EEG is EEG-X (Foumani et al., 2025), which also uses a teacher–student formulation but still denoises the waveform, so its target remains signal-level. Hierarchical targets have been explored before, but always over a discrete or causal space—autoregressive prediction of residual vector-quantized codes (Cui et al., 2026) or a topology-induced hierarchy modelled autoregressively (Yang et al., 2025). TAHLP predicts *continuous* residuals with no quantization step, no separately fitted vocabulary and no causal factorization.

Motor-imagery specialization. Generic EEG foundation models fit one encoder to corpora spanning several paradigms (Kostas et al., 2021; Yang et al., 2023; Wang et al., 2024; Jiang et al., 2024b; Wang et al., 2025; Zhou et al., 2025), and nothing in a reconstruction or contrastive objective selects one paradigm over another. This has an empirical counterpart: benchmarks report that specialist decoders trained from scratch remain competitive and that larger models do not necessarily generalize better (Liu et al., 2026a). In the per-subject regime that most MI benchmarks report—where a decoder is fitted on the same subject it is later tested on—compact convolutional (Schirrmeister et al., 2017) and classical decoders such as filter-bank common spatial patterns and Riemannian classifiers (Ang et al., 2008; Yger et al., 2017) remain strong references that foundation models do not uniformly surpass. We invoke this literature as motivation only and deliberately do not compare against those pipelines in Table 1: they are fitted on test-subject data by construction, which our subject-independent protocol excludes (Section 3.2), and no published numbers exist for them under our split and preprocessing. For MI specifically, (Liu et al., 2026b) show that a paradigm-specific model beats generic ones while consuming far less pre-training data, and (Liu et al., 2025) argue that MI corpora need dedicated curation rather than being absorbed into a generic pool. TAHLP follows

Table 3: Design choices of the generalist baselines compared in Table 1. ✓/✗ mark whether a property holds (green: yes; red: no)—**rel.**: the encoding depends on the displacement between two electrodes rather than on absolute position; **CB-free**: no separately fitted codebook; **MI-only**: the pre-training corpus is restricted to motor imagery. Pre-training data size and parameter counts are transcribed from (Liu et al., 2026a) Table 1, except LaBraM (2,534 h, from (Jiang et al., 2024b)) and BrainRVQ (9,200 h and 5.82 M, from (Cui et al., 2026)); where a model ships several sizes we print the range as published, and a dash means the source reports no value.

Model	Venue	Electrode encoding	Pre-training objective	rel.	CB-free	MI-only	Pre-train data	Params (M)
BENDR	Front. HNS 2021	fixed montage, index PE	contrastive + token reconstruction	✗	✓	✗	1.5 TB	4.0
BIOT	NeurIPS 2023	channel-wise tokens	masked token reconstruction	✗	✓	✗	58,021 h	3.2
LaBraM	ICLR 2024	absolute PE, 10–20 numbering	masked codebook prediction	✗	✗	✗	2,534 h	5.8–369
CBraMod	ICLR 2025	ACPE, conv over channel index	masked raw reconstruction	✓	✓	✗	27,062 h	4.0
BrainRVQ	arXiv 2026	channel-identity tokens	hierarchical AR over RVQ codes	✗	✗	✗	9,200 h	5.82
CSBrain	NeurIPS 2025	anatomical region partition	masked reconstruction	✗	✓	✗	9,000 h	4.9
TAHLP (ours)	this work	**name embedding + I3D-ERoPE**	**multi-level latent-prediction self-distillation**	✓	✓	✓	**221 h**	**6.2**

this line: the corpus, the objective and the inductive bias are all chosen for MI, and no label is used at any point.

Table 3 summarizes the design choices of the generalist baselines we compare against.

B Pre-training Monitoring

Self-supervised pre-training has no labelled objective, so the training loss alone cannot tell us whether the encoder is learning useful representations or silently degrading. We therefore monitor two possible failure modes. *Collapse*: if the teacher's target tokens converge onto a single direction, the loss becomes trivially small while carrying no information. *Context copying*: the student may reconstruct a masked token by copying its visible neighbours rather than inferring the missing signal. To catch both, we track four readouts every epoch:

1. **Validation loss** is the masked-prediction objective evaluated on the held-out validation split. *Lower is better*, but it is insufficient on its own. *Healthy behaviour*: it falls steeply in the early epochs and then plateaus at a low value (in our run, from ≈ 0.5 to ≈ 0.32); a rising validation loss would signal divergence.
2. **Alignment: cos and cm** are the mean cosine between the student prediction and the EMA-teacher target, averaged over *all* positions (`cos`) and over *masked* positions only (`cm`). *Higher is better*. `cm` is the harder, more informative readout because it is scored exactly where the student saw nothing, whereas `cos` also benefits from visible positions the student can copy. *Healthy behaviour*: both rise toward ≈ 0.9 and stay high, with `cm` tracking `cos` at a small, near-constant offset (in our run `cm` ≈ 0.89 alongside `cos` ≈ 0.91).
3. **Visibility gap** $\texttt{gap} = 2(\texttt{cos} - \texttt{cm})$ is the advantage of having seen a position. *A small, positive value is desirable*. *Healthy range*: roughly $+0.02$ to $+0.05$, stable or shrinking (in our run it settles near $+0.034$); a large or growing gap signals context copying.
4. **Collapse probe** `pwc` is the mean pairwise cosine among the teacher's own target tokens—a student-independent readout. *Far from* 1 *is desirable*. *Healthy range*: well below 1 and roughly flat, ideally low (near 0 means maximally diverse) and certainly below ≈ 0.5 (in our run it stays in $[0.40, 0.44]$ after epoch 1); approaching 1 means the targets have collapsed onto a single direction.

These readouts are deliberately redundant: because each target token is ℓ_2-normalized along the feature dimension, a collapsed teacher would drive loss, cos and cm to spuriously good values at once, so `pwc` is the independent control that catches this, and `gap` catches the complementary context-copying shortcut—neither failure is visible from the loss curve alone.

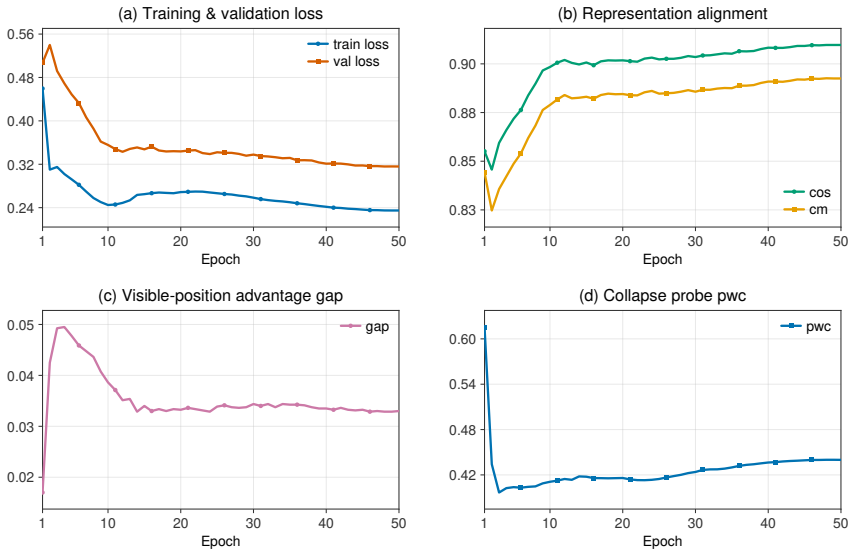

Figure 4: Pre-training monitoring: loss, alignment, and health diagnostics.

We train TAHLP for 50 epochs and, for downstream fine-tuning, select the checkpoint with the minimum validation loss of the pretext objective; this minimum is reached at epoch 48. Across the run, validation loss decreases steadily and cos/cm rise together, while gap stays small and positive and pwc remains well below 1 and essentially flat— evidence that the student is genuinely aligning to the teacher rather than collapsing or copying context. Validation loss plateaus after epoch 44 while training loss is still falling, *i.e.* a saturation rather than overfitting regime, so we release the epoch-48 checkpoint.

C Probing Electrode-Topology Awareness in the Pretrained Encoder

To test what the geometry-native position encoding has taught the model, we probe the self-supervised pretrained encoder for genuine *electrode-topology awareness*: does its internal geometry respect scalp position, or does it merely memorize electrode coordinates without spatial organization? Figure 5 summarizes the evidence.

Protocol. We ask whether the geometry of the pretrained encoder respects scalp position or merely memorizes electrode coordinates, using three readouts over $n = 300$ MI trials (153 left- and 147 right-hand): (i) *coordinate specificity*, the attention–geometry alignment ρ (Spearman) under true versus permuted electrode coordinates, shown as scalp topographies (Figure 5a–c) and tested by a paired-permutation test (Figure 5d); (ii) *CLS sensitivity*, the cosine between the CLS token under true and under permuted coordinates (Figure 5e); and (iii) *laterality*, the laterality index $\mathrm{LI} = (\mathrm{attn}_{\mathrm{right}} - \mathrm{attn}_{\mathrm{left}})/(\mathrm{attn}_{\mathrm{right}} + \mathrm{attn}_{\mathrm{left}})$ over left- vs. right-hand trials, tested against zero by a sign-flip permutation test (Figure 5f).

Alignment is small in absolute terms but reliable under true coordinates ($\rho_{\mathrm{mean}} \approx 0.099$, std 0.028) versus ≈ 0.017 (std 0.037) under permutation—a paired difference of ≈ 0.082 with paired-permutation $p = 0.0005$; the CLS token is sensitive to perturbation ($\bar{\rho}_{\mathrm{CLS}} = 0.936$, $\Delta = 0.064$); and LI is reliably non-zero though small ($+0.052$, 95% CI $\approx [0.036, 0.068]$, for left-hand and ≈ -0.049 for right-hand trials, SE ≈ 0.008, $p = 0.0005$) and symmetric—the left- and right-hand magnitudes do not differ ($\Delta|\mathrm{LI}| = 0.003$, $p = 0.80$).

How much these numbers can bear. We do not present this probe as evidence that TAHLP has discovered sensorimotor lateralization, and we report both effect sizes as small. Three considerations keep the inference within what $n = 300$ trials can support. *First, the design is paired within trial.* The permutation control holds the signal, the weights and the mask fixed and changes only the coordinates fed to I3D-ERoPE, so the quantity under test is the coordinate-dependent component of

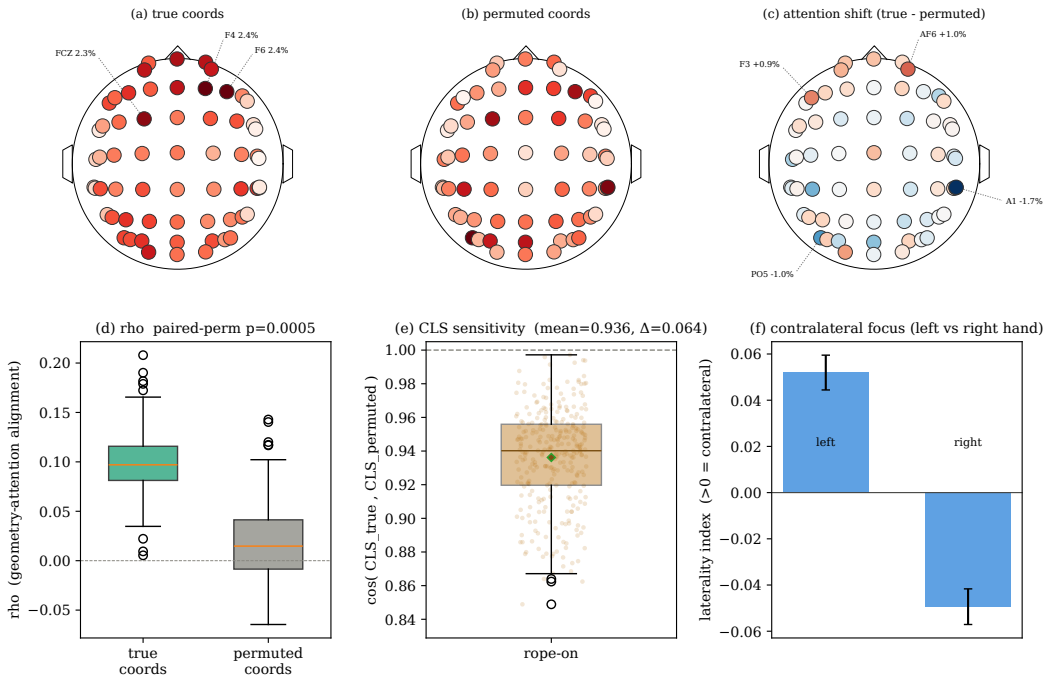

Figure 5: Probing electrode-topology awareness in the TAHLP pretrained encoder. **Top row** (**a**–**c**, scalp topography): (**a**) attention–geometry alignment under *true* electrode coordinates; (**b**) the same under *permuted* coordinates; (**c**) attention shift (true − permuted). **Bottom row** (**d**–**f**): (**d**) coordinate-specificity control—a paired-permutation test of $\cos(\mathrm{CLS}_{\mathrm{true}}, \mathrm{CLS}_{\mathrm{permuted}})$ under the rope-on model, $p = 0.0005$; (**e**) CLS sensitivity: the CLS token's cosine similarity to its unperturbed reference drops to $\bar{\rho}_{\mathrm{CLS}} = 0.936$ ($\Delta = 0.064$); (**f**) contralateral focus: laterality index ($\mathrm{LI} > 0$ = contralateral) for left- vs. right-hand trials, with a sign-flip permutation test and a left-vs-right magnitude contrast. See the protocol below.

attention, not its overall magnitude. Absolute values such as $\rho \approx 0.10$ or $\mathrm{LI} \approx 0.05$ are reported for completeness and are not the basis of the claim. *Second, for LI the informative quantity is the sign, not the size.* $\mathrm{LI} = +0.052$ versus -0.049 is a left-minus-right difference of ≈ 0.10 ($\mathrm{SE} \approx 0.011$ by error propagation), and it is the sign flip together with its mirror symmetry ($\Delta|\mathrm{LI}| = 0.003$, $p = 0.80$) that a coordinate lookup table would not be expected to produce; an artefact driven by, say, an asymmetry in the montage or in the mask would not reverse cleanly between hands. *Third, a small magnitude is the expected outcome here and is therefore a weak falsification target.* The probe reads the attention of a *self-supervised* encoder whose objective is to predict masked tokens from their context, not of a task-trained decoder, so nothing in the pretext task rewards moving attention mass between hemispheres; $\mathrm{LI} \approx 0.05$ means that roughly 5% of the bilateral attention mass shifts contralaterally, which is a plausible order of magnitude for a representation that has never seen a class label. The claim we draw is correspondingly narrow: the encoder's attention is a function of the electrode coordinates it is given, and the ordering it induces agrees in sign with the known contralateral organization of motor imagery. We do not use it to argue that the encoder decodes MI on its own. The evidence that I3D-ERoPE matters for decoding is causal rather than correlational—the component ablation of Table 2 ($+1.20$, $+1.20$ and $+1.55$ Balanced Accuracy on PhysioNet-MI, SHU-MI and BCICIV-2a over a learned coordinate embedding) and the data-efficiency curves of Figure 2—and the probe is offered as a mechanistic gloss on those results, not as an independent argument.

D Heterogeneity of the Pre-training Corpus

The pre-training corpus aggregates 22 publicly released MI-EEG datasets whose *raw* (source) specifications differ along four axes, summarized in Figure 6 and Table 4. Every number in this section

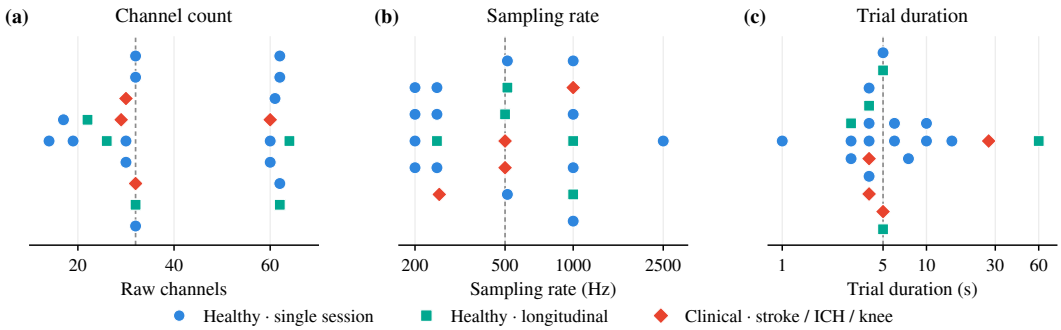

Figure 6: Raw-source heterogeneity of the MI-EEG pre-training corpus. Each point denotes one dataset. Panels show (a) raw channel count, (b) raw sampling rate (log scale), and (c) raw single-trial duration (log scale). Marker color and shape encode cohort type: healthy single-session (blue circles), healthy longitudinal (teal squares), and clinical stroke/ICH/knee (coral diamonds). Dashed vertical lines mark medians. All values are pre-processing (raw) specifications.

describes the *original* recordings, *prior* to resampling to 200 Hz, channel selection and trial windowing; conflating the two is a common misreading we avoid here.

(a) Channel count and montage. Retained channels range from 14 (Zhou2016) to 62 (Forenzo2023, Lee2019_OpenBMI) over nominal 64-channel caps. The layouts themselves are heterogeneous rather than one standard cap: standard 10–20 (most datasets), 10–05 high-density (Forenzo2023, Stieger2021), Biosemi-32 (Ofner2017MI, Rozado2015, Tavakolan2017, Zenodo-5882500) and the g.Nautilus 17-channel cap (TrianaGuzman2024). Because channel *identity*—the electrode name and its 3D position—is preserved, this heterogeneity is exactly what our topology-aware encodings exploit.

(b) Sampling rate. Acquisition hardware spans more than an order of magnitude, 200 Hz to 2 500 Hz, over seven bands: 200 Hz (Kaya2018, Shin2017A/B, Weibo2014); 250 Hz; 256 Hz (HefmiIch2025, a hybrid EEG–fNIRS rig); 500 Hz; 512 Hz; 1 000 Hz (Forenzo2023, GuttmannFlury2025, Lee2019/OpenBMI, Liu2025, Stieger2021, Tavakolan2017, Yi2025); and 2 500 Hz (Jeong2020MI). All recordings are resampled to a common 200 Hz axis before training, so the native rate is a property of the *source*, not of the processed representation—which is why a rate-agnostic front end is part of our design. *SHU-2022* appears only in the raw candidate pool: it is excluded from pre-training (Section 3.1) because the SHU-MI benchmark is drawn from the same cohort.

(c) Trial duration. Single-trial lengths vary by roughly 60×, from 1.0 s (Kaya2018, a tight pre-cue imagery segment) to 60.0 s (Forenzo2023, an entire BCI game round including cue, imagery and feedback); most datasets use 3–5 s imagery windows. These are raw, source-defined extents before any windowing, and they preclude any fixed-duration assumption.

(d) Session structure. Most datasets are single-session, but six are longitudinal: Zhou2020 (7 days), Kumar2024 (6 sessions over months), Zuo2025 and Liu2025 (5 sessions each), Stieger2021 (7–11 sessions) and Forenzo2023. Stieger2021 and Forenzo2023 are the only large-scale sources recorded with real-time neurofeedback, making them uniquely suited to studying user learning during BCI control.

E Preprocessing Details

We standardize every dataset independently, so that no statistic is shared across corpora, and we keep each recording in its native channel space. For each dataset the following steps are applied to the raw per-subject, per-session recordings:

1. **Channel selection.** Non-EEG channels are removed: EOG, ECG, EMG, stimulus, and miscellaneous auxiliary channels are dropped, retaining only EEG channels.
2. **Re-referencing.** A common average reference (CAR) is applied across the retained EEG channels.

3. **Band-pass and notch filtering.** A zero-phase FIR filter realizes a 0.5–$40\,\mathrm{Hz}$ band-pass, together with $50\,\mathrm{Hz}$ and $60\,\mathrm{Hz}$ notch filters to suppress line noise from both European/Chinese and Korean/American acquisition sites. Filtering is performed at the native sampling rate, prior to any resampling, so that the pass-band is interpreted on the original time base.
4. **Unit normalization.** All signals are scaled to microvolts (μV) by applying the appropriate per-dataset gain correction (*e.g.* V$\rightarrow$ μV, mV-as-V$\rightarrow$ μV, or nV$\rightarrow$ μV), so that every dataset shares a common amplitude unit.
5. **Resampling.** Every recording is resampled to a unified $200\,\mathrm{Hz}$ time axis using polyphase resampling.
6. **Trial segmentation.** Around each motor-imagery cue, a fixed-length window is cut from the continuous recording, using a length equal to that dataset's native trial duration (1–$10\,\mathrm{s}$ across datasets), preserving the original per-trial semantic label. Trial length is thus retained at the dataset level and is not forced to a single global value, so the corpus keeps its original temporal heterogeneity.

Channel-space representation. The standardized signals are kept in their native-channel form with no interpolation applied (HDF5 attribute `interpolation="none"`): each dataset retains its native electrode set and physical channel identity, rather than being resampled onto a fixed standard montage. This preserves the electrode geometry that our electrode-aware encoders exploit.

Cross-subject normalization and geometric encoding. The per-dataset HDF5 files are merged into a unified corpus. Each subject is augmented with a per-subject z-score normalization, in which per-channel mean and standard deviation are computed within the subject and stored alongside the data, enabling cross-subject amplitude alignment without a shared global scale; Euclidean alignment (He & Wu, 2020; Wu, 2025) is additionally applied to whiten each subject's covariance toward a shared reference. In addition, the 3D positions of all electrodes are normalized onto a common unit sphere via a head-sphere fit to the standard_1005 template, $(p-c)/r$, yielding scale- and translation-invariant coordinates for I3D-ERoPE. The resulting unified HDF5 corpus stores, per subject, the $200\,\mathrm{Hz}$ μV EEG tensor together with its per-subject z-score statistics, unit-sphere electrode coordinates, and semantic labels, consumable directly by the pre-training encoder.

Batching. Each trial is represented as $X \in \mathbb{R}^{C\times T}$ with its ordered channel-name list. Following Section 2.1, the trial is cut into non-overlapping patches of $t = 200$ samples and $A = \lfloor T/t \rfloor$ patches are retained per channel, so any trailing remainder shorter than one full patch is discarded. A batch is dynamically padded to the maximum channel count and maximum number of patches within the batch. Binary masks $\mathbf{b}^{\mathrm{ch}}$ and $\mathbf{b}^{\mathrm{time}}$ mark valid channels and valid time patches; they are passed to the attention layer so that padded positions do not affect attention, and they also gate the application of I3D-ERoPE.

Reading the per-dataset inventory. Table 4 lists all 22 datasets with the numbers that matter for reproducing the corpus. **Trial** is the raw source-defined trial extent, before windowing, and f_s is the native sampling rate; both describe the *original* recordings, and they are the quantities discussed in the four axes above. Every other column is read back from the unified corpus and therefore reports what is actually fed to the model: **Ch** is the number of EEG electrodes retained after non-EEG channels (EOG, ECG, EMG, stimulus) are removed—hence one or two below the nominal montage size for several datasets—and **Subjects** is the number of participants contributing at least one retained trial. **Win** is the analysis window on the common $200\,\mathrm{Hz}$ time axis; it differs from **Trial** only for the five datasets whose source trials are longer than the 4–$10\,\mathrm{s}$ analysis window (Forenzo2023, GuttmannFlury2025, HefmiIch2025, TrianaGuzman2024, Zenodo-5882500), and the durations are *not* full session times—they exclude inter-trial rest and calibration. **Trials** is the number of motor-imagery trials retained and **Dur (h)** converts it to hours through the analysis window, $\mathrm{Dur} = N_{\mathrm{trials}} \times \mathrm{Win}/3600$; the two columns sum to the 207 887 trials and $\approx 221\,\mathrm{h}$ quoted in the main text. **Split** gives the resulting *subject-level* train/validation/test trial counts (448/57/54 subjects) and sums to 169 430 / 18 450 / 20 007 (178 h/20 h/22 h), so **Dur (h)** covers all three partitions; the training subset alone holds $\approx 178\,\mathrm{h}$ and pre-training uses $\approx 199\,\mathrm{h}$ of training plus validation data. Kaya2018 and Zhou2016 have no test partition because their cohorts are too small to split three ways. **Cohort** is H for healthy controls and C for a clinical cohort (HefmiIch2025, 20 intracerebral-hemorrhage patients plus 17 healthy controls; Liu2024, 50 stroke patients; Liu2025, 27 stroke patients; Zuo2025, 30 knee-pain patients), and **Long.** marks a longitudinal design with ≥ 5

Table 4: Per-dataset inventory of the pre-training corpus: 22 publicly released MI-EEG datasets. **Trial** and f_s describe the original recordings; every other column is read back from the unified corpus. See Appendix D for column definitions, cohort codes and provenance.

Dataset	Ch	f_s (Hz)	Trial (s)	Win (s)	Classes	Subjects	Cohort	Long.	Trials	Split (tr/va/te)	Dur (h)
Forenzo2023 ‡	62	1000	60.0	4.0	2	25	H	✓	3 661	2931/292/438	4.1
GuttmannFlury2025 ‡	60	1000	7.5	4.0	2	31	H		2 520	2040/200/280	2.8
HefmiIch2025 ‡	32	256	27.0	10.0	2	37	C		2 940	2550/150/240	8.2
Jeong2020MI ‡	60	2500	4.0	4.0	11	25	H		39 902	32251/3301/4350	44.3
Kaya2018	19	200	1.0	1.0	3	7	H		16 126	15168/958/0	4.5
Kumar2024 ‡	22	512	5.0	5.0	2	18	H	✓	7 156	5567/792/797	9.9
Lee2019_OpenBMI	62	1000	4.0	4.0	2	54	H		21 600	17200/2000/2400	24.0
Liu2024 ‡	29	500	4.0	4.0	2	50	C		2 000	1600/200/200	2.2
Liu2025 ‡	59	1000	5.0	5.0	2	27	C	✓	8 350	7090/840/420	11.6
Ofner2017MI	61	512	3.0	3.0	7	15	H		900	720/120/60	0.8
Rozado2015 ‡	32	512	6.0	6.0	2	30	H		1 550	1225/175/150	2.6
Shin2017A	30	200	10.0	10.0	2	29	H		1 720	1360/180/180	4.8
Shin2017B	30	200	10.0	10.0	2	29	H		1 740	1380/180/180	4.8
Stieger2021	60	1000	3.0	3.0	4	56	H	✓	49 942	40164/4047/5731	41.6
Tavakolan2017 ‡	30	1000	3.0	3.0	3	12	H		2 760	2280/240/240	2.3
TrianaGuzman2024 ‡	17	250	15.0	4.0	4	32	H		7 679	6239/720/720	8.5
Weibo2014	60	200	4.0	4.0	7	10	H		5 540	4480/500/560	6.2
Yi2025 ‡	60	1000	4.0	4.0	8	18	H		5 801	4520/640/641	6.4
Zenodo-5882500 ‡	32	250	6.0	4.0	2	12	H		5 760	4800/480/480	6.4
Zhou2016	14	250	5.0	5.0	3	4	H		1 800	1365/435/0	2.5
Zhou2020 ‡	26	500	5.0	5.0	4	8	H	✓	6 640	5000/800/840	9.2
Zuo2025 ‡	30	500	4.0	4.0	2	30	C	✓	11 800	9500/1200/1100	13.1
Total (22)	14–62	200–2500	1.0–60.0	1.0–10.0	2–11	**559**	4 C / 18 H	6	**207 887**	**169 430/18 450/20 007**	≈**221**

sessions, matching the black bar outlines in Figure 6. Entries marked ‡ have no dedicated publication and are cited from their release record; several datasets are accessed through MOABB (Jayaram & Barachant, 2018). For space the table itself carries no reference markers—the sources, in table order, are (Forenzo et al., 2023; Guttmann-Flury et al., 2025; Hefmi Ich study group, 2025; Jeong et al., 2020; Kaya et al., 2018; Kumar et al., 2024; Lee et al., 2019; Liu et al., 2024; 2025; Ofner et al., 2017; Rozado et al., 2015; Shin et al., 2017; Stieger et al., 2021; Tavakolan et al., 2017; Triana-Guzman et al., 2024; Yi et al., 2014; 2025; unknown, 2022; Zhou et al., 2016; 2020; Zuo et al., 2025).

F Implementation Details

F.1 Spatiotemporal Backbone

A TemporalConv stem followed by a 12-layer Transformer. The stem is a 2-D convolution stack `Conv2d`$(1 \to 8,$ `kernel` $(1, 15),$ `stride` $(1, 8),$ `padding` $(0, 7))$ plus two $(1, 3)$ refinements. A 200-sample patch ($\approx$ 1 s at 200 Hz) is compressed to 8 feature maps of 25 time positions, which are flattened and linearly projected to the model dimension 200 (the projection enforces embed_dim $=$ out_chans $\times$ 25 $=$ 200), yielding exactly one token per patch, consistent with Equation 2 and with $N = CA$ in Equation 8. The Transformer (Vaswani et al., 2017) uses hidden dimension 200, 10 attention heads, MLP ratio 4, stochastic depth (drop-path) 0.1, and LayerScale initial value 0.1.

F.2 Heterogeneous Channel Encoding

Rather than interpolating every recording onto a fixed montage, we keep each dataset's native electrode layout and identify channels by name. A learnable channel-identity embedding table assigns a per-channel vector that is added to all tokens of that channel; the vocabulary is built as described in Appendix F.3 and is extended at run time with any electrode name discovered in a new dataset. Spatial geometry is injected through *I3D-ERoPE*, described in Appendix F.4. A dynamic-padding scheme with an XNOR-equivalent attention mask lets the same model process variable channel counts and variable-length trials without layout interpolation.

F.3 Channel Vocabulary

Channels are identified by name through one global lookup table, built once and extended at run time: a curated 10–05 list contributes its hand-written part, a standard 3D electrode position table

contributes every named electrode that is not already in the core list, names occurring in both are merged by name deduplication (so no name occupies two slots), and any name first seen in a new dataset is appended dynamically. Table 5 summarizes the construction.

Table 5: Construction of the channel-identity vocabulary. No electrode name appears twice: the deduplication step merges the names shared by the curated list and the 3D position table (FCC3, POO1, POO2, POO7, POO8). The third row is obtained by difference, $475 - 87 - 257$, because the construction script does not log it as a separate component; it is the standard 10–05 core list.

Component	Size
Curated `_EXT_105` hand-written part	87
Merged from `ELECTRODE_3D` (not in core)	257
Remaining `standard_105` core names (by difference)	131
`standard_105` subtotal	475
Dynamically discovered channels (`discovered_channels.json`)	10
`CHANNEL_VOCAB` total	485

F.4 Electrode Coordinates and I3D-ERoPE

Electrode information enters the model in two disjoint roles. *Absolute identity* comes from the learned name-embedding table of Appendix F.3, indexed by channel name and independent of geometry. *Relative geometry* comes only from I3D-ERoPE: the 3D scalp coordinate $\mathbf{u}_i$ rotates queries and keys and is never added to the token embedding.

Coordinate preparation. The collated batch carries $\mathbf{U}^{\mathrm{ch}} \in \mathbb{R}^{B\times C\times 3}$, broadcast along the time-patch axis so that every patch of a channel shares one coordinate; the `[CLS]` token has $\mathbf{u} = \mathbf{0}$ and is not rotated. Coordinates are mapped onto the unit head sphere by a head-sphere fit to the `standard_1005` template, $\mathbf{u}_i = (\mathbf{p}_i - c)/r$, with center c and radius r estimated from the template (Appendix E), and then scaled by the coordinate scale $\gamma = 10$. The fit removes the translation and scale difference between pre-training coordinates given in metric units and downstream coordinates taken from a unit-sphere name table—without it, rotation angles would differ by an order of magnitude between the two stages—and γ places several periods of the highest-frequency plane inside the unit sphere, which avoids angle aliasing between distant electrodes.

Rotation. For head dimension d_h, plane $c = 0, \ldots, d_h/2 - 1$ is assigned the axis $a = c \bmod d_u$ with $d_u = 3$ (the three spatial axes; no time axis is used) and rotates the query/key pair by the angle $\gamma\, \mathbf{u}_i[a]\, \omega_c$, where $\omega_c = \mathrm{base}^{-2c/d_h}$ and $\mathrm{base} = 10^4$, exactly as in Equation 5; values are left unrotated. The $d_h/2$ planes are distributed round-robin over the three axes, so each axis spans the full frequency range. Because RoPE acts multiplicatively, the attention score depends only on the relative displacement $\mathbf{u}_{i'} - \mathbf{u}_i$ (Equation 6), so spatially nearby electrodes receive higher compatibility.

Masking and augmentation policy. Consistently with Section 2.1, masking is token-wise—each *valid* token is independently drawn for masking, with no channel-level or block-level structure—with probability $r = 0.7$, giving Ω; the symmetric branch then masks the complementary set $I \setminus \Omega$, so that every token is predicted in exactly one pass. Masked positions in the student are replaced by a learnable `mask_token`, so roughly 30% of the valid tokens remain visible in each pass. When data augmentation is used, it is applied once to the shared input before masking, so that the teacher and student remain positionally aligned. The level weights in the hierarchical prediction loss are $w = [1, 0.5, 0.25]$.

Pre-training hyper-parameters. We pre-train for 50 epochs at batch size 64 with AdamW (peak learning rate $5{\times}10^{-4}$, weight decay 0.05, cosine decay after a 2% warm-up), in bf16 with gradient checkpointing; the run takes about 19 h on $8{\times}$A800. The EMA momentum follows a cosine schedule $0.996 \to 1.0$, the teacher target averages the top $K = 6$ blocks after instance normalization, the hierarchical loss uses three prediction levels with weights $(1.0, 0.5, 0.25)$, and the token mask ratio is $r = 0.70$, leaving about 30% of valid tokens visible per pass.

F.5 Downstream Datasets

All four benchmarks are held out from pre-training and evaluated under a strictly subject-independent protocol: no subject contributes to both pre-training and fine-tuning, and the subject-level partition is fixed once and shared by every method. **PhysioNet-MI** (Schalk et al., 2004) is the largest pool (109 subjects, 64 channels, 160 Hz) and the only one with four effector classes that include both fists; **SHU-MI** (Ma et al., 2022) (25 subjects, 32 channels, 250 Hz) and **BNCI2015001** (Faller et al., 2012) (12 subjects, 13 channels, 512 Hz) are binary, the first contrasting left versus right hand and the second right hand versus both feet from the sparsest montage in the group; **BCICIV-2a** (Tangermann et al., 2012) is the smallest pool (9 subjects, 22 channels, 250 Hz) and the most fine-grained, with four classes that include tongue imagery. Together the four span a $12\times$ range in subject pool, a $5\times$ range in channel count and two different binary contrasts, which is what makes the subject-pool and montage effects in Section 3.3 readable.

Preprocessing and splits follow (Cui et al., 2026) and (Zhou et al., 2025) so that our numbers can be compared with theirs directly: every dataset is resampled to 200 Hz and cut into the analysis window listed in Table 6, and BCICIV-2a additionally keeps the 4–40 Hz band-pass those papers use. The last subjects of each dataset are reserved for test, the preceding ones for validation, and model selection uses κ on the 4-class tasks and Balanced Accuracy on the binary ones (Section 3.2).

Table 6: Summary of the downstream MI-EEG datasets.

Property	**PhysioNet-MI**	**SHU-MI**	**BCICIV-2a**	**BNCI2015001**
Task type	4-class MI	2-class MI	4-class MI	2-class MI
Classes	L/R fist, both fists, both feet	L/R hand	L/R hand, feet, tongue	R hand, both feet
Subjects	109	25	9	12
Channels	64	32	22	13
Original rate	160 Hz	250 Hz	250 Hz	512 Hz
Resampled rate	200 Hz	200 Hz	200 Hz	200 Hz
Trial window	4 s	4 s	$[2, 6]$ s (4 s)	$[3, 8]$ s (5 s)
Total samples	9,837	11,988	5,088	2,400
Train subjects	1–70	1–15	1–5	1–7
Val subjects	71–89	16–20	6–7	8–9
Test subjects	90–109	21–25	8–9	10–12

G Same-Corpus Control: Re-pre-training the Generalist Baselines

Table 1 compares TAHLP against the *released* checkpoints of the generalist baselines. Those checkpoints were pre-trained on corpora that differ from ours by more than an order of magnitude (Table 3), so a raw comparison conflates the pre-training objective and the electrode encoding with the data budget. To separate them, we re-pre-train every generalist baseline *from its own published pre-training recipe, using its own released pre-training code, on our 221 h motor-imagery corpus*, then fine-tune under exactly the same downstream protocol as in Section 3.2. The hyperparameters of each baseline are kept at their published values; only the corpus changes.

Tables 7 and 8 report the paired comparison over all four benchmarks (SHU-MI and BNCI2015001 for the binary task; BCICIV-2a and PhysioNet-MI for the 4-class tasks). **Scope of the control.** Re-pre-training a released baseline means reproducing its full published recipe on our data—for CSBrain, whose published recipe alone costs about 9,000 h of TUEG and 101 GPU-hours on four A100s, the most expensive of the six—and we re-pre-train all six generalists on the same 221 h corpus. Retraining on the smaller corpus does not cost *every* baseline accuracy: all six lose ground on the two binary benchmarks, but five of six *improve* on BCICIV-2a and only CSBrain declines (-2.67 Balanced Accuracy points), and the ordering among the generalists is broadly preserved. TAHLP leads on all four benchmarks by a wider margin than it does against the released checkpoints ($+1.42$ versus $+0.86$ Balanced Accuracy points on SHU-MI, $+2.56$ versus $+0.58$ on BCICIV-2a, $+2.75$ versus $+2.10$ on PhysioNet-MI and $+1.52$ versus $+1.06$ on BNCI2015001). The gain therefore comes from the pre-training objective and the electrode encoding rather than from the data budget.

Table 7: **Same-corpus control, binary benchmarks.** For each generalist baseline we report its released checkpoint and, below it (†), the same architecture re-pre-trained by us on our 221 h MI corpus and fine-tuned under the protocol of Section 3.2; percent, mean over five seeds, standard deviation in subscript, best per column in **bold** and second best underlined.

	SHU-MI (2-class)			**BNCI2015001** (2-class)		
Method	**Balanced Acc.**	**AUC-PR**	**AUROC**	**Balanced Acc.**	**AUC-PR**	**AUROC**
Released checkpoints (their own corpora)						
BENDR	$55.70_{\pm 2.30}$	$58.50_{\pm 2.70}$	$58.60_{\pm 2.80}$	$63.37_{\pm 1.67}$	$66.47_{\pm 1.48}$	$65.67_{\pm 1.93}$
BIOT	$61.79_{\pm 1.83}$	$67.70_{\pm 1.19}$	$66.09_{\pm 1.27}$	$73.60_{\pm 1.08}$	$79.61_{\pm 1.61}$	$77.00_{\pm 1.27}$
LaBraM	$61.66_{\pm 1.92}$	$67.61_{\pm 0.83}$	$66.04_{\pm 0.91}$	$69.47_{\pm 1.13}$	$75.12_{\pm 0.95}$	$73.75_{\pm 1.29}$
CBraMod	$63.70_{\pm 1.51}$	$71.39_{\pm 0.88}$	$69.88_{\pm 0.68}$	$73.44_{\pm 2.71}$	$81.33_{\pm 1.86}$	$79.31_{\pm 2.49}$
BrainRVQ	$63.70_{\pm 1.80}$	$72.30_{\pm 1.10}$	$71.70_{\pm 0.70}$	$73.01_{\pm 2.14}$	$82.01_{\pm 1.42}$	$81.21_{\pm 1.95}$
CSBrain	$64.17_{\pm 0.37}$	$72.30_{\pm 0.79}$	$72.00_{\pm 0.58}$	$73.88_{\pm 1.52}$	$82.41_{\pm 1.36}$	$81.91_{\pm 1.04}$
Re-pre-trained by us on our 221 h MI corpus						
BENDR†	$55.65_{\pm 0.93}$	$61.27_{\pm 1.18}$	$61.78_{\pm 0.86}$	$63.19_{\pm 1.58}$	$68.58_{\pm 1.76}$	$69.52_{\pm 1.49}$
BIOT†	$58.18_{\pm 0.91}$	$64.83_{\pm 1.38}$	$63.53_{\pm 1.20}$	$69.80_{\pm 1.40}$	$76.36_{\pm 1.12}$	$74.46_{\pm 1.55}$
LaBraM†	$58.59_{\pm 0.89}$	$69.02_{\pm 1.48}$	$63.59_{\pm 1.08}$	$66.04_{\pm 1.16}$	$77.32_{\pm 1.28}$	$71.23_{\pm 0.87}$
CBraMod†	$60.82_{\pm 0.70}$	$69.45_{\pm 0.98}$	$69.05_{\pm 0.90}$	$70.91_{\pm 2.58}$	$79.46_{\pm 1.49}$	$78.66_{\pm 2.40}$
BrainRVQ†	$61.19_{\pm 0.84}$	$69.69_{\pm 0.89}$	$69.47_{\pm 1.25}$	$70.55_{\pm 2.08}$	$79.55_{\pm 1.35}$	$78.62_{\pm 1.88}$
CSBrain†	$63.61_{\pm 0.43}$	$72.12_{\pm 0.76}$	$71.74_{\pm 0.64}$	$73.42_{\pm 1.24}$	$81.98_{\pm 1.29}$	$81.49_{\pm 1.05}$
TAHLP (ours)	$\mathbf{65.03_{\pm 1.92}}$	$\mathbf{73.60_{\pm 1.28}}$	$\mathbf{73.24_{\pm 0.92}}$	$\mathbf{74.94_{\pm 1.18}}$	$\mathbf{83.81_{\pm 1.45}}$	$\mathbf{82.95_{\pm 0.94}}$

Table 8: **Same-corpus control, 4-class benchmarks.** Same layout as Table 7: for each generalist baseline, the released checkpoint followed by the same architecture re-pre-trained by us on our **221 h** motor-imagery corpus (†). All values are percent, mean over five seeds with the standard deviation in subscript; best per column in **bold**, second best underlined.

	BCICIV-2a (4-class)			**PhysioNet-MI** (4-class)		
Method	**Balanced Acc.**	**Cohen'sκ**	**Weighted F1**	**Balanced Acc.**	**Cohen'sκ**	**Weighted F1**
Released checkpoints (their own corpora)						
BENDR	$49.00_{\pm 0.70}$	$32.00_{\pm 0.90}$	$48.40_{\pm 0.70}$	$42.20_{\pm 2.20}$	$40.10_{\pm 2.50}$	$40.80_{\pm 2.40}$
BIOT	$47.48_{\pm 0.93}$	$29.97_{\pm 1.39}$	$46.07_{\pm 1.25}$	$61.53_{\pm 1.54}$	$48.75_{\pm 2.72}$	$61.58_{\pm 1.97}$
LaBraM	$48.69_{\pm 0.85}$	$31.59_{\pm 1.54}$	$47.58_{\pm 1.03}$	$61.73_{\pm 1.22}$	$49.12_{\pm 1.92}$	$61.77_{\pm 1.41}$
CBraMod	$51.38_{\pm 0.66}$	$35.18_{\pm 0.94}$	$49.84_{\pm 0.85}$	$64.17_{\pm 0.91}$	$52.22_{\pm 1.69}$	$64.27_{\pm 1.00}$
BrainRVQ	$54.10_{\pm 0.80}$	$38.80_{\pm 0.80}$	$53.30_{\pm 1.20}$	$64.20_{\pm 0.80}$	$52.30_{\pm 1.50}$	$64.40_{\pm 0.90}$
CSBrain	$56.57_{\pm 0.71}$	$42.09_{\pm 0.93}$	$56.37_{\pm 0.87}$	$63.04_{\pm 0.90}$	$50.71_{\pm 1.20}$	$63.08_{\pm 0.95}$
Re-pre-trained by us on our 221 h MI corpus						
BENDR†	$50.78_{\pm 0.85}$	$34.37_{\pm 1.15}$	$50.55_{\pm 0.78}$	$59.82_{\pm 2.08}$	$46.43_{\pm 2.40}$	$60.01_{\pm 2.30}$
BIOT†	$51.67_{\pm 0.93}$	$35.56_{\pm 1.35}$	$51.32_{\pm 1.18}$	$60.85_{\pm 1.55}$	$47.80_{\pm 2.62}$	$60.87_{\pm 1.88}$
LaBraM†	$52.74_{\pm 0.88}$	$36.99_{\pm 1.45}$	$52.29_{\pm 1.00}$	$61.59_{\pm 1.30}$	$48.79_{\pm 1.83}$	$61.88_{\pm 1.33}$
CBraMod†	$53.51_{\pm 0.73}$	$38.01_{\pm 0.90}$	$53.06_{\pm 0.85}$	$62.56_{\pm 0.90}$	$50.08_{\pm 1.60}$	$62.60_{\pm 0.98}$
BrainRVQ†	$54.59_{\pm 0.79}$	$39.45_{\pm 0.86}$	$54.13_{\pm 1.22}$	$63.34_{\pm 0.87}$	$51.12_{\pm 1.42}$	$63.53_{\pm 0.88}$
CSBrain†	$53.90_{\pm 0.72}$	$38.53_{\pm 0.90}$	$53.72_{\pm 0.84}$	$63.55_{\pm 0.86}$	$51.40_{\pm 1.18}$	$63.60_{\pm 0.92}$
TAHLP (ours)	$\mathbf{57.15_{\pm 0.71}}$	$\mathbf{42.87_{\pm 0.76}}$	$\mathbf{56.76_{\pm 1.05}}$	$\mathbf{66.30_{\pm 0.86}}$	$\mathbf{55.07_{\pm 1.42}}$	$\mathbf{66.50_{\pm 0.94}}$

Why the re-pre-trained baselines move, and by how much. Shrinking each baseline's corpus to ours is not one manipulation but several that happen to coincide, and separating them matters for how the † rows should be read.

(a) A common component: the budget itself. With architecture, optimizer and schedule held at their published values, every encoder simply sees far fewer distinct samples at 221 h than at 2,534–58,021 h, so the reduction in data is common to all six by construction—but the six do not respond alike. All six lose accuracy on the two binary benchmarks (-0.05 to -3.61 Balanced Accuracy

points on SHU-MI, -0.18 to -3.80 on BNCI2015001); on PhysioNet-MI every change except BENDR's is within 1.61 points; and on BCICIV-2a five of six *improve* by 0.49 to 4.19 points while CSBrain alone loses 2.67. A response that depends this strongly on the benchmark cannot be a pure budget effect, so we attribute the common component to the budget and attribute the spread to the objective, as (b) and (c) do model by model.

(b) Discretized targets are fitted before the encoder is trained (LaBraM, BrainRVQ). LaBraM's target space is a neural codebook of $8{,}192$ codes of dimension 64, fitted by vector-quantized neural spectrum prediction in a separate first stage and then frozen, after which the encoder is trained to predict those codes (Jiang et al., 2024b); the paper reports that accuracy keeps improving with both model size and the volume of pre-training data. A vocabulary of that cardinality is only meaningful if each code receives enough assignments to be estimated reliably, and the data available per code falls by the same factor as the corpus—about $11\times$ for LaBraM ($2{,}534\,\text{h}\rightarrow 221\,\text{h}$). When a codebook is under-supported the degradation is not gradual: the target keeps its full cardinality while losing its semantics, so the encoder is trained against a vocabulary that is increasingly arbitrary, and the quantization error of the tokenizer is inherited by the prediction target. This is the mechanism behind the general point made in the main text—that a useful codebook has to be bought with data. BrainRVQ organizes its codes as a residual hierarchy, which can shed resolution gracefully under the same shrink, and it does correspondingly lose less than LaBraM on the two binary benchmarks (-2.51 and -2.46 against -3.07 and -3.43).

(c) Learned electrode representations have to be re-acquired from data (BIOT, BENDR). Neither model is given electrode geometry: BIOT learns one embedding per channel slot over 18 predefined channels and BENDR uses a fixed montage with an index positional encoding, so the fact that two electrodes are neighbours is something the model must recover from corpus statistics rather than read off its coordinates. BIOT's self-supervised objective—contrastive alignment between a perturbed and an unperturbed view—can only extract as much spatial structure as the data volume supports; at $221\,\text{h}$, i.e. 0.4% of its $58{,}021\,\text{h}$, most of that evidence is gone. BIOT accordingly pays the largest penalty of any baseline on the two binary benchmarks (-3.61 on SHU-MI and -3.80 on BNCI2015001), which are the two benchmarks whose decision rests on a single localized sensorimotor contrast—left versus right hand on SHU-MI, and the hand versus foot areas of the sensorimotor strip on BNCI2015001—rather than on the four-way effector discrimination of the 4-class tasks. BENDR is the mirror image and should not be counted as a degradation at all: its released checkpoint is contrastive-trained on about $1.5\,\text{TB}$ of largely clinical, non-motor-imagery recordings and is the weakest model here on MI to begin with (42.20 Balanced Accuracy on PhysioNet-MI, below every specialist decoder; its published $\kappa = 40.10$ is also internally inconsistent, since $\kappa = (\text{BAcc}-25)/75$ for a balanced 4-class task would imply $\text{BAcc} \approx 55.1$). Re-pre-training on MI-only data therefore repairs a domain mismatch rather than costing accuracy: BENDR *gains* 17.62 points on PhysioNet-MI and 1.78 on BCICIV-2a, and is essentially unchanged on the two binary benchmarks (-0.05 and -0.18)—the one baseline for which our $221\,\text{h}$ corpus is an improvement on its own pre-training data.

Two biases that work against the baselines. We flag two asymmetries that make the † rows a pessimistic estimate of what these methods can do at $221\,\text{h}$, because both favour our own numbers. *(i) Update budget.* Every recipe is re-run at its published epoch count. Because the corpus shrinks by $11\times$–$260\times$ while the epoch count is unchanged, the number of optimizer steps a baseline receives before fine-tuning falls by the same factor, into a regime that its learning rate, warm-up length and cosine horizon were never chosen for. Scaling the epoch count to hold the step budget fixed would remove the confound, but on a corpus this small it trades under-training for over-fitting, so the honest remedy is a per-baseline re-tuning of the schedule—which we cannot afford for six models, and which would in any case raise the question of how much search budget each baseline is entitled to. *(ii) Hyper-parameters and target cardinality.* Choices that are well matched to a $2{,}534\,\text{h}$ or $58{,}021\,\text{h}$ corpus—an $8{,}192$-entry codebook, a fixed 18-channel slot table, a mask ratio, weight decay and capacity tuned at that scale—are not automatically well matched to $221\,\text{h}$. Our own hyper-parameters, by contrast, were selected on this corpus. This asymmetry favours TAHLP, so we do not rest the headline claim of the paper on this control: Table 1 compares against the released checkpoints, and the same-corpus control is offered as evidence about *where* the gain comes from, not as the margin we claim. A budget-matched re-tuning of the two strongest baselines, which would bound the size of this bias, is left to future work.