MIND WORLD MODELS: JOINT LEARNING OF MOTOR EEG REPRESENTATIONS AND ACTION-CONDITIONED WORLD DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motor decoding from electroencephalography (EEG) supports non-invasive brain–computer interfaces, yet existing methods either rely on costly cue-based trials and coarse categorical labels or use pretraining objectives that do not directly capture how movement unfolds, limiting their ability to learn transferable motor representations. To this end, we propose the Mind World Model (MWM), a framework that learns motor EEG representations from scalable naturalistic EEG and egocentric video without action annotations. MWM treats EEG observed up to the current frame as the latent action of a world model: a mind model encodes the EEG, and a world model uses it to predict the representation of a future frame, yielding a movement-grounded pretraining objective. We further introduce the Lateralization-Aware Transformer (LAT), which uses frame averaging to separate shared bilateral from side-specific lateral features, encoding hemispheric correspondence to reduce the burden of learning left–right relationships from data. MWM pretraining produces two models: a mind model for downstream motor decoding and a world model that can be adapted to prediction conditioned on real actions, opening a scalable and flexible path to collect egocentric data for embodied AI. Across 9 classification and 7 regression tasks, the mind model improves task-averaged balanced accuracy by 5.9% and Pearson correlation by 16.2% relative to the strongest baseline in each task family. Compared with an inverse-dynamics latent-action baseline, the world model reduces L1 error by 23.2% within recordings and 11.0% on held-out subjects.

1 INTRODUCTION

Motor decoding from electroencephalography (EEG) aims to infer intended or executed movements from non-invasive recordings of brain activity. It encompasses both recognizing movement categories and estimating continuous quantities such as movement trajectories and force. This capability underpins brain–computer interfaces for assistive control and neurorehabilitation, allowing movement-related neural activity to guide external devices or rehabilitation feedback (Tangermann et al., 2012; Ramos-Murguialday et al., 2013). For these applications, useful EEG representations must capture both which movement is intended and how it unfolds, and transfer across motor tasks with limited labeled data. Learning such representations presents three challenges (Figure 1, top).

First, task-specific motor decoding faces limitations in both data scalability and the information conveyed by supervision. Supervised motor-decoding research is largely organized around labeled examples of predefined tasks, commonly collected through repeated, cue-based laboratory trials (Tangermann et al., 2012; Schirrmeister et al., 2017; Lawhern et al., 2018). Such protocols make it costly to expand the volume and diversity of training data. Moreover, categorical annotations such as “left hand” and “right hand” compress motor behavior into a small vocabulary. Movements with different trajectories, speeds, or forces can share the same label, leaving their within-class variation unspecified by the training target. Direct measurements of continuous kinematics or force retain such within-class variation, but collecting them at scale typically requires task-specific instrumentation and careful synchronization with EEG. Thus, scaling motor representation learning requires both more accessible collection paradigm and supervision that retains information beyond discrete action categories.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

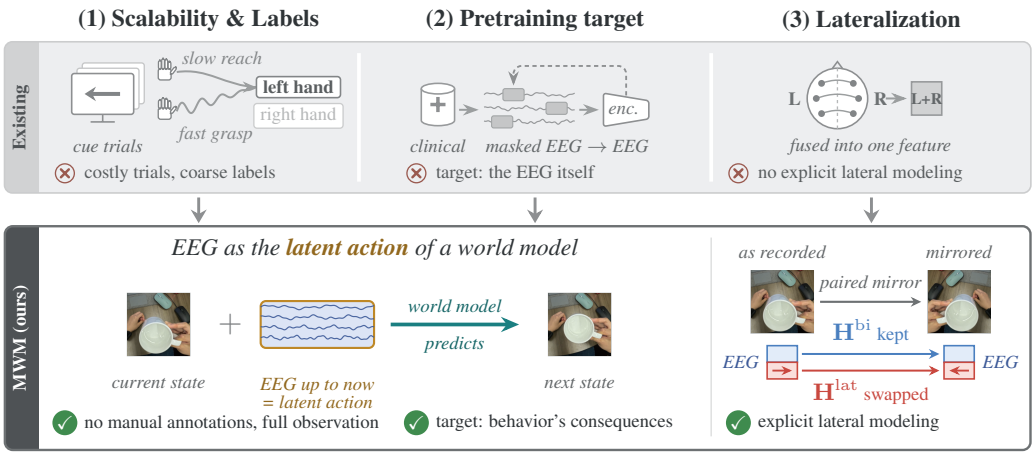


Figure 1: **Existing approaches (top) versus MWM (bottom) on the three challenges of motor EEG representation learning.** **(1)** Cue-based trials are costly to scale, and categorical labels give one class to movements that unfold differently (Tangermann et al., 2012; Schirrmeister et al., 2017; Lawhern et al., 2018). **(2)** General-purpose EEG foundation models (FMs) pretrain with EEG-only objectives (Kostas et al., 2021; Jiang et al., 2024; Wang et al., 2025a). MWM answers both with one mechanism: it treats the EEG recorded up to the current frame as the latent action of a world model that predicts the next visual state. It therefore learns from naturalistic EEG and egocentric video without action annotations, and its target is the visual consequence of behavior rather than the EEG itself. **(3)** Existing methods do not explicitly separate bilateral from lateral activity (Schirrmeister et al., 2017; Lawhern et al., 2018; Wang et al., 2025a); MWM does, and pairs each horizontal video mirror with an exchange of only the lateral component $\mathbf{H}^{\text{lat}}$.

Second, general-purpose EEG foundation models address label scarcity through self-supervised objectives such as contrastive learning, masked signal reconstruction, and neural token prediction (Kostas et al., 2021; Chien et al., 2022; Jiang et al., 2024; Wang et al., 2024b; 2025a). Although these objectives learn useful signal structure, they do not directly supervise the relationship between neural activity and movement. Their pretraining corpora are also assembled primarily for broad or clinical EEG analysis rather than for capturing continuous, naturalistic motor behavior and its consequences (Jiang et al., 2024; Wang et al., 2025a; Obeid & Picone, 2016; El Ouahidi et al., 2025). These differences motivate pretraining that aligns both the learning signal and the recording context with motor decoding.

Third, motor EEG exhibits *lateralization*: similar movement-related patterns can indicate different acting sides depending on which hemisphere they occur in. For example, sensorimotor rhythm suppression is often stronger over the left hemisphere for right-hand movements and vice versa (Pfurtscheller & Neuper, 1997; Nam et al., 2011). Without this structural prior, existing encoders (Schirrmeister et al., 2017; Lawhern et al., 2018; Wang et al., 2025a) must learn the relationship between patterns and hemispheric location from data, potentially requiring larger training datasets and yielding suboptimal decoding performance when data are limited.

World-model training faces a similar bottleneck in scaling data collection. Instrumented interfaces can constrain natural behavior and capture only selected aspects of action. For example, handheld grippers and motion-capture gloves primarily record hand kinematics, without directly measuring head and leg movements, interaction forces, or hand stiffness (Chi et al., 2024; Wang et al., 2024a). Video-based approaches instead infer pseudo-actions from current and future frames (Bruce et al., 2024; Gao et al., 2025). However, access to future frames introduces a risk of target-information leakage: pseudo-actions may encode future visual content, allowing the world model to exploit prediction shortcuts rather than learn transferable action-conditioned dynamics (Garrido et al., 2026).

To address these challenges, we propose the Mind World Model (MWM), which jointly learns a mind model and a world model from synchronized EEG and egocentric video in EgoBrain (Lin et al., 2026). The mind model maps EEG observed up to the current frame to a latent action, and the world model uses it to predict a future visual representation. Using egocentric recordings of natural motor behavior, rather than general-purpose EEG corpora, aligns the training data with motor decoding.

Visual prediction supplies supervision beyond categorical labels without task-specific annotation, enabling scalable pretraining. Encoding actions from EEG avoids explicit motion measurements and future-frame leakage. Joint learning thus produces action-related EEG representations and a world-model initialization for future prediction conditioned on real actions.

To capture motor lateralization, we introduce the Lateralization-Aware Transformer (LAT), an EEG encoder that uses frame averaging (Puny et al., 2022) over original and mirrored representations to separate shared bilateral features from side-specific lateral features. When a video is mirrored horizontally, LAT exchanges lateral EEG features across homologous electrodes while preserving the bilateral component. This prior links corresponding movement patterns across hemispheres, reducing the burden of learning their left–right relationship from limited data.

Our contributions are as follows:

- We introduce MWM, which learns motor EEG representations from naturalistic EEG–video by predicting the visual consequences of behavior. This objective provides supervision beyond categorical action labels and enables scalable data collection without explicit action annotations.
- We propose LAT to encode hemispheric correspondence by separating bilateral and lateral EEG features, reducing the burden of learning this structure from data.
- We show that the pretrained world model transfers from EEG-derived actions to prediction conditioned on real actions, opening a path to scalable world-model data collection without relying on manipulation interfaces that constrain natural behavior and capture only selected aspects of action.

We evaluate the mind model on 9 classification and 7 regression tasks, and the world model on future prediction conditioned on real actions. Relative to the strongest baseline in each task family, the mind model improves task-averaged balanced accuracy by 5.9% and Pearson correlation by 16.2%. Ablations show that alignment, prediction, and frame averaging each improve average transfer, with effects that vary by task. Relative to the latent-action baseline, the world model reduces L1 error by 23.2% within recordings and 11.0% on held-out subjects.

2 Related Works

EEG Foundation Models. EEG foundation models learn transferable representations through contrastive learning, masked reconstruction, and token prediction (Kostas et al., 2021; Chien et al., 2022; Jiang et al., 2024; Wang et al., 2024b). Such objectives exploit structure in unlabeled recordings, reducing dependence on task-specific annotations and enabling transfer across EEG analysis tasks. Other work improves spatiotemporal modeling and electrode-layout flexibility (Wang et al., 2025a;b; Döner et al., 2025), scales pretraining (El Ouahidi et al., 2025), or incorporates language and complementary neural modalities (Jiang et al., 2025; Xiao et al., 2025). These advances target general EEG representation learning; MWM uses visual consequences of natural behavior to supervise motor representations and explicitly models hemispheric correspondence.

World Models. World models support prediction and control through observation generation (Ha & Schmidhuber, 2018; Micheli et al., 2023; Alonso et al., 2024) or compact latent dynamics (Hafner et al., 2020; Hansen et al., 2024). Recent approaches predict frozen or jointly learned visual features (Zhou et al., 2025a; Sobal et al., 2025; Assran et al., 2025; Maes et al., 2026). Predicting in feature space lets the dynamics model focus on abstract state changes without reconstructing every detail of the observation. Video-based methods infer latent actions from frame transitions (Bruce et al., 2024; Gao et al., 2025; Garrido et al., 2026). However, collecting explicit action labels can be difficult to scale, while inferring actions from future frames risks leaking target information into the action representation (Garrido et al., 2026). MWM pretrains a world model from EEG and video, deriving actions without explicit motion measurements or future frames. The pretrained model transfers to prediction conditioned on real actions.

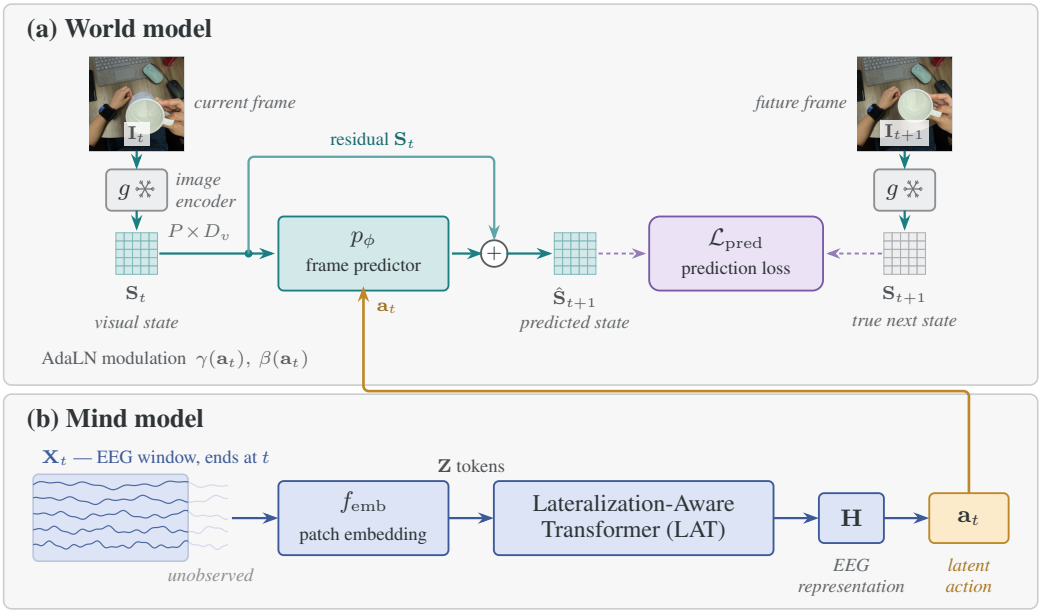


Figure 2: **The Mind World Model. (a)** A frozen encoder g maps the current frame to patch features that a residual predictor advances; the same frozen g builds the target from the next frame. **(b)** The mind model encodes the EEG window, which ends on the current frame; its LAT encoder is detailed in Figure 3. The EEG reaches the predictor only as AdaLN modulation, never as an input token: it occupies the slot a recorded action would.

3 METHOD

Our central design is to use neural activity in place of an externally recorded action within a predictive world model. The mind model observes EEG only up to the current visual state and must encode the information needed to predict what happens next. We first formulate this prediction task and its two model components, then describe the Lateralization-Aware Transformer (LAT) and training objectives.

3.1 WORLD MODEL

As illustrated in Figure 2, the Mind World Model (MWM) jointly learns motor-related EEG representations and visual dynamics from synchronized EEG and egocentric video without action annotations. Given an EEG segment $\mathbf{X}_t \in \mathbb{R}^{C\times T}$ and a current RGB frame $\mathbf{I}_t \in \mathbb{R}^{3\times H_v\times W_v}$, the mind model uses LAT (Section 3.2) to encode the EEG into a latent action $\mathbf{a}_t \in \mathbb{R}^{D_a}$, which conditions the world model's prediction of the future visual state. Here, C is the number of EEG channels, T the number of time samples, H_v and W_v the image height and width, and D_a the action dimension.

Image Encoder. As shown in Figure 2(a), the current frame $\mathbf{I}_t$ and its future frame $\mathbf{I}_{t+1} \in \mathbb{R}^{3\times H_v\times W_v}$ are independently encoded by a frozen DINOv2 encoder g (Oquab et al., 2024):

$$\mathbf{S}_t = g(\mathbf{I}_t) \in \mathbb{R}^{P\times D_v}, \qquad \mathbf{S}_{t+1} = g(\mathbf{I}_{t+1}) \in \mathbb{R}^{P\times D_v}, \tag{1}$$

where P is the number of visual patches and D_v is their feature dimension. We use DINOv2 to balance performance and computational efficiency. Freezing g provides stable targets and reduces training compute and memory by eliminating encoder backpropagation.

Frame Predictor. The frame predictor p_ϕ advances the current visual state using the EEG-derived latent action:

$$\hat{\mathbf{S}}_{t+1} = p_\phi(\mathbf{S}_t, \mathbf{a}_t) \in \mathbb{R}^{P\times D_v}. \tag{2}$$

The mind-model pathway in Figure 2(b) aggregates EEG tokens into $\mathbf{a}_t$, which enters the frame predictor only through modulation. The current-frame tokens are then processed by a lightweight Transformer whose layers are modulated by $\mathbf{a}_t$ through adaptive layer normalization (AdaLN) (Peebles & Xie, 2023). For a hidden token $\mathbf{h} \in \mathbb{R}^{D_p}$, where D_p is the predictor width, the action

produces a scale $\gamma(\mathbf{a}_t) \in \mathbb{R}^{D_p}$ and a shift $\beta(\mathbf{a}_t) \in \mathbb{R}^{D_p}$:

$$\mathrm{AdaLN}(\mathbf{h}, \mathbf{a}_t) = \gamma(\mathbf{a}_t) \odot \mathrm{LN}(\mathbf{h}) + \beta(\mathbf{a}_t). \tag{3}$$

This design treats EEG as a latent action that controls the predicted visual transition. The residual connection in Figure 2(a) adds a predicted change tensor $\Delta_\phi(\mathbf{S}_t, \mathbf{a}_t) \in \mathbb{R}^{P \times D_v}$ to the current state:

$$\hat{\mathbf{S}}_{t+1} = \mathbf{S}_t + \Delta_\phi(\mathbf{S}_t, \mathbf{a}_t), \tag{4}$$

such that the predictor focuses on dynamic changes instead of reconstructing the static scene.

To avoid temporal leakage, the EEG window ends at the current frame, while the prediction target is sampled from the subsequent time interval. Therefore, the target visual change is not already contained in the EEG observation window.

3.2 Mind Model

Transformers. Figure 3(a) shows how each channel of $\mathbf{X}_t$ is divided into non-overlapping temporal patches. A channel-wise CNN embeds each patch by combining its temporal waveform and spectral representation, producing

$$\mathbf{Z}_t = f_{\mathrm{emb}}(\mathbf{X}_t) \in \mathbb{R}^{C \times N \times D_e}, \tag{5}$$

where N is the number of temporal patches per channel and D_e is the EEG token width.

The criss-cross Transformer block in Figure 3(c) alternates temporal attention across patches with spatial attention across EEG channels. This factorization preserves the channel–time structure of EEG while reducing the cost of full attention. The resulting patch tokens are aggregated into the latent action $\mathbf{a}_t$ used by the frame predictor.

LAT and Paired Mirror Augmentation. LAT encodes the lateralization prior described in the introduction: corresponding movement patterns over opposite hemispheres indicate opposite acting sides. Modeling this correspondence explicitly allows LAT to learn shared movement structure across left–right counterparts.

Raw EEG mixes bilateral activity with lateralized motor activity; swapping entire channels would exchange both, although only the latter should change sides under our motor prior. We therefore learn the decomposition in feature space, as shown in Figure 3(b). Let $\Pi \in \{0,1\}^{C \times C}$ exchange homologous electrodes, such as C3 and C4, along the channel axis, and write $\mathbf{Z} = \mathbf{Z}_t \in \mathbb{R}^{C \times N \times D_e}$. An elementwise gate $\mathbf{G} \in [0,1]^{C \times N \times D_e}$ separates bilateral and lateral components $\mathbf{Z}^{\mathrm{bi}}, \mathbf{Z}^{\mathrm{lat}} \in \mathbb{R}^{C \times N \times D_e}$:

$$\mathbf{Z}^{\mathrm{lat}} = \mathbf{G} \odot \mathbf{Z}, \qquad \mathbf{Z}^{\mathrm{bi}} = (1 - \mathbf{G}) \odot \mathbf{Z}. \tag{6}$$

We define the mirror operation as

$$\mathcal{P}(\mathbf{Z}) = \mathbf{Z}^{\mathrm{bi}} + \Pi \mathbf{Z}^{\mathrm{lat}} \in \mathbb{R}^{C \times N \times D_e}. \tag{7}$$

Thus, bilateral activity is preserved, while lateral activity is exchanged between the two hemispheres.

Bilateral features capture movement structure shared by both sides and should remain invariant under mirroring, while lateral features encode the acting side and should transform equivariantly. Inspired by frame averaging (FA) (Puny et al., 2022), we wrap a shared Transformer with invariant and equivariant readouts (Figure 3(d)). The same Transformer $\mathcal{T}$ encodes the original and mirrored inputs into $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{C \times N \times D_e}$:

$$\mathbf{Z}_1 = \mathcal{T}(\mathbf{Z}), \qquad \mathbf{Z}_2 = \mathcal{T}(\mathcal{P}(\mathbf{Z})). \tag{8}$$

The invariant head averages these outputs; the equivariant head takes their signed difference after aligning homologous channels. Each head uses a linear feature readout $r_{\mathrm{inv}}, r_{\mathrm{eq}} : \mathbb{R}^{D_e} \to \mathbb{R}^{D_e/2}$, applied at each channel and time location, to produce $\mathbf{H}^{\mathrm{bi}}, \mathbf{H}^{\mathrm{lat}} \in \mathbb{R}^{C \times N \times D_e/2}$:

$$\mathbf{H}^{\mathrm{bi}} = r_{\mathrm{inv}}\left(\frac{\mathbf{Z}_1 + \mathbf{Z}_2}{2}\right), \qquad \mathbf{H}^{\mathrm{lat}} = r_{\mathrm{eq}}\left(\frac{\mathbf{Z}_1 - \Pi \mathbf{Z}_2}{2}\right). \tag{9}$$

Concatenating them yields the final EEG representation $\mathbf{H} = \left[\mathbf{H}^{\mathrm{bi}}; \mathbf{H}^{\mathrm{lat}}\right] \in \mathbb{R}^{C \times N \times D_e}$.

Egocentric recordings can overrepresent dominant-hand movements because people preferentially use one hand, with right-handedness more common in the population (Papadatou-Pastou et al., 2020). To address this sampling bias, we randomly mirror each video horizontally during training and apply the corresponding operation $\mathcal{P}$ to its EEG representation, as illustrated in Figure 3(e). This produces paired left–right counterparts while preserving their cross-modal correspondence.

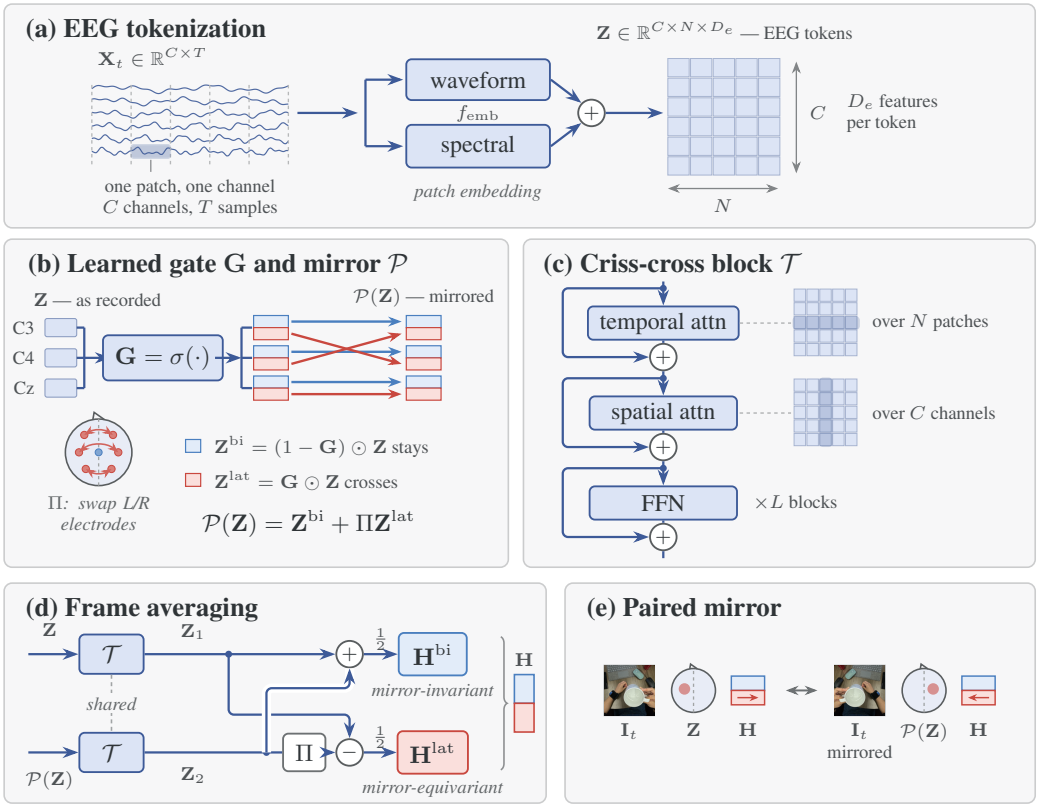


Figure 3: **The Lateralization-Aware Transformer (LAT) for EEG encoding.** **(a)** Patch embedding, applied within each channel, **(b)** the learned gate $\mathbf{G} \in [0,1]^{C\times N\times D_e}$ and the electrode permutation $\Pi \in \{0,1\}^{C\times C}$, which act on the embedded tokens before any Transformer block, **(c)** a criss-cross block, **(d)** a shared Transformer with invariant and equivariant readout heads, and **(e)** the paired EEG–video mirror. Only the learned lateral component is exchanged across hemispheres; the raw signal is never permuted.

3.3 Training Objective

Alignment Loss. We align global EEG and synchronized-frame representations using symmetric InfoNCE (van den Oord et al., 2018), denoted $\mathcal{L}_{\text{align}}$. EEG features are projected into the visual feature space, and both modalities are ℓ_2-normalized before computing similarities.

Prediction Loss. The prediction loss in Figure 2(a) compares the predicted state with the frozen representation of the future frame. For each sample, we average the ℓ_1 distance over the P patches:

$$\mathcal{L}_{\text{pred}} = \frac{1}{P}\sum_{p=1}^{P}\left\|\hat{\mathbf{S}}_{t+1}^{p} - \mathbf{S}_{t+1}^{p}\right\|_1. \tag{10}$$

Because the DINOv2 encoder g is frozen, the visual target remains fixed throughout training. Minimizing this loss therefore requires the mind model to extract EEG information that explains the upcoming visual transition.

The overall training objective combines the standard masked EEG patch reconstruction loss $\mathcal{L}_{\text{mask}}$ with visual alignment and prediction, weighted by nonnegative coefficients λ_{align} and λ_{pred}:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}. \tag{11}$$

4 Experiments

We first evaluate the pretrained mind model on motor regression and classification benchmarks. We then analyze lateralization in its representations through hand-motion readouts and motor-imagery

Table 1: Pearson correlation on regression tasks. Mean ± sample s.d. over three seeds.

Model	ForceGameReg	SrisReachME	OfnerApertureME	OfnerKinME	WAYGALReg	BiomechMag	ForenzoFutVec
BIOT	$0.044_{\pm 0.025}$	$0.102_{\pm 0.014}$	$0.032_{\pm 0.008}$	$-0.021_{\pm 0.003}$	$0.217_{\pm 0.001}$	$0.039_{\pm 0.001}$	$0.041_{\pm 0.003}$
LaBraM	$0.112_{\pm 0.003}$	$0.173_{\pm 0.002}$	$0.059_{\pm 0.007}$	$0.001_{\pm 0.002}$	$0.226_{\pm 0.001}$	$0.119_{\pm 0.001}$	$0.094_{\pm 0.002}$
CBraMod	$0.083_{\pm 0.027}$	$\underline{0.217}_{\pm 0.001}$	$0.075_{\pm 0.007}$	$0.078_{\pm 0.011}$	$0.324_{\pm 0.003}$	$0.138_{\pm 0.001}$	$0.096_{\pm 0.006}$
CSBrain	$\underline{0.113}_{\pm 0.003}$	$0.185_{\pm 0.001}$	$0.053_{\pm 0.006}$	$0.062_{\pm 0.011}$	$0.295_{\pm 0.001}$	$0.136_{\pm 0.002}$	$\underline{0.114}_{\pm 0.002}$
REVE	$0.089_{\pm 0.021}$	$\underline{0.217}_{\pm 0.001}$	$\underline{0.153}_{\pm 0.015}$	$\underline{0.175}_{\pm 0.010}$	$\mathbf{0.460}_{\pm \mathbf{0.003}}$	$\underline{0.153}_{\pm 0.005}$	$0.097_{\pm 0.002}$
MWM (ours)	$\mathbf{0.165}_{\pm \mathbf{0.010}}$	$\mathbf{0.241}_{\pm \mathbf{0.002}}$	$\mathbf{0.233}_{\pm \mathbf{0.018}}$	$\mathbf{0.206}_{\pm \mathbf{0.006}}$	$\underline{0.433}_{\pm 0.002}$	$\mathbf{0.157}_{\pm \mathbf{0.002}}$	$\mathbf{0.127}_{\pm \mathbf{0.001}}$

topographies, and use ablations to assess the contributions of the lateralization-aware architecture, prediction, and alignment. Finally, we evaluate the pretrained world model's transfer from EEG-derived actions to future prediction conditioned on real actions.

4.1 Experimental Setup

Pretraining. We pretrain MWM on 63.28 hours of synchronized EEG and egocentric video from all 40 EgoBrain participants (Lin et al., 2026). All regression and classification benchmarks use the same fixed pretrained checkpoint. Dataset preprocessing details are provided in Appendix A.

Linear-probe evaluation. We evaluate every backbone on the same task-specific one-second EEG input sampled at 200 Hz, freezing the backbone and training only a linear prediction head. We report test balanced accuracy for classification and test Pearson correlation for regression, averaging correlations across output dimensions for vector targets. Each table entry reports the mean and standard deviation over three evaluation seeds shared across models for that task. The participant-disjoint splits and evaluation settings are specified in Appendices A and B.

Baselines. We compare with recent EEG foundation models: BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025a), CSBrain (Zhou et al., 2025b), and REVE (El Ouahidi et al., 2025).

4.2 Regression Tasks

Tasks. We evaluate seven continuous decoding tasks in Table 1. ForceGameReg (Li et al., 2026) predicts plateau force; SrisReachME (Srisrisawang & Müller-Putz, 2024), OfnerApertureME, OfnerKinME (Ofner et al., 2017), and BiomechMag (Perez-Blanco et al., 2026) predict future reaching speed, hand-aperture change, hand speed, and wrist-speed magnitude, respectively. WAYGALReg (Luciw et al., 2014) predicts grip and load forces, while ForenzoFutVec (Forenzo et al., 2024) predicts two-dimensional cursor velocity in a closed-loop motor-imagery interface. Dataset statistics, target definitions, preprocessing, and participant-disjoint splits are detailed in Appendix A.

Results. In Table 1, MWM improves task-averaged Pearson correlation by 16.2% relative to REVE, the strongest baseline by this measure. Gains span force, kinematic, and cursor-velocity targets. WAYGALReg is the exception, with REVE performing best. This task involves grip and load forces under varying object weight and surface friction (Luciw et al., 2014). We hypothesize that visual prediction provides a less direct supervisory signal for these force variations, which are not uniquely determined by visible movement.

4.3 Classification Tasks

Tasks. We evaluate nine classification tasks in Table 2: PhysioNet-MI (Schalk, 2009), Schwarz (Schwarz et al., 2020), Kaya5F (Kaya et al., 2018), Forenzo2024 (Forenzo et al., 2024), FineMI (Yi et al., 2025), Ofner2017ME and Ofner2017MI (Ofner et al., 2017), Weibo2014 (Yi et al., 2014), and Jeong2020 (Jeong et al., 2020). They cover motor execution, motor imagery, pre-movement grasp decoding, and cursor-control direction. Schwarz evaluates grasp type from EEG recorded 1–2 s before movement onset. Dataset statistics, target definitions, preprocessing, and participant-disjoint splits are detailed in Appendix A.

Results. In Table 2, MWM improves task-averaged balanced accuracy by 5.9% relative to CBraMod, the strongest baseline by this measure.

Table 2: Balanced accuracy on classification tasks. Mean ± sample s.d. over three seeds.

Model	PhysioNet-MI	Schwarz	Kaya5F	Forenzo2024	FineMI	Ofner2017ME	Weibo2014	Ofner2017MI	Jeong2020
BIOT	0.262±0.008	0.507±0.007	0.207±0.008	0.272±0.004	0.172±0.009	0.149±0.020	0.161±0.003	0.147±0.005	0.097±0.003
LaBraM	0.470±0.006	0.522±0.018	0.231±0.007	0.297±0.002	0.179±0.014	0.160±0.008	0.185±0.020	0.158±0.001	0.118±0.008
CBraMod	0.581±0.007	0.519±0.003	0.266±0.008	0.313±0.003	0.199±0.002	0.177±0.026	0.183±0.002	0.147±0.003	0.146±0.000
CSBrain	0.549±0.003	0.551±0.025	0.244±0.005	0.303±0.001	0.183±0.002	0.174±0.008	0.162±0.001	0.148±0.005	0.149±0.006
REVE	0.560±0.006	0.529±0.012	0.266±0.005	0.294±0.001	0.190±0.004	0.187±0.009	0.186±0.008	0.154±0.001	0.138±0.004
MWM (ours)	**0.606**±0.007	**0.575**±0.035	**0.269**±0.006	**0.318**±0.004	**0.200**±0.005	**0.202**±0.007	**0.187**±0.005	**0.165**±0.002	**0.158**±0.004

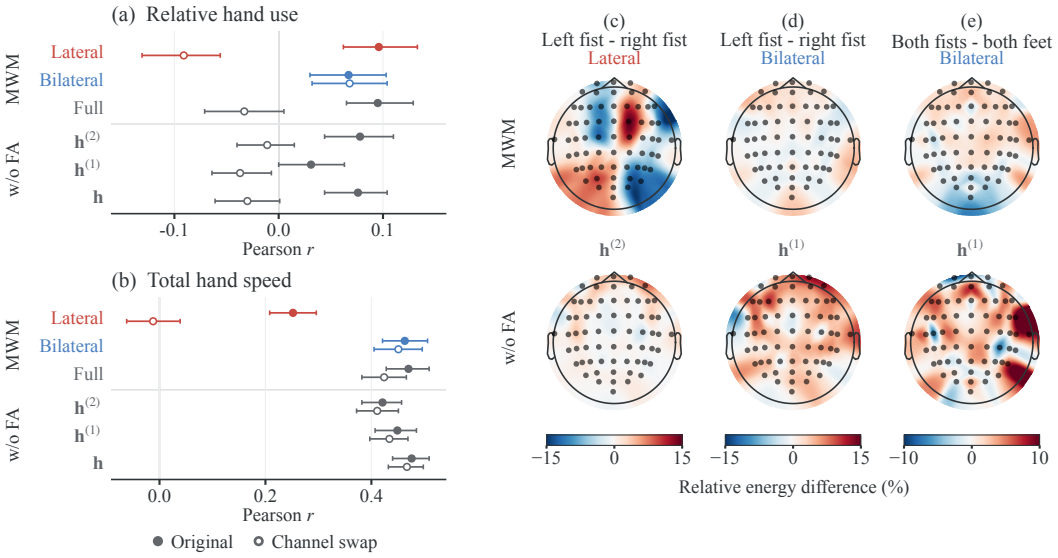


Figure 4: **Lateralization analysis. (a,b)** Hand-motion correlations with 95% confidence intervals estimated by bootstrap resampling, using participants as the resampling unit. **(c–e)** Participant-averaged motor-imagery feature-energy contrasts, normalized by channel-wise mean energy; left–right contrasts share a color scale. Control features $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$ are equal-sized halves of $\mathbf{h}$.

4.4 Lateralization Analysis

We examine hand-motion readouts and motor-imagery topographies of the bilateral and lateral representations. The control, *w/o FA*, removes the frame-averaging (FA) wrapper, leaving a plain Transformer. Protocols and additional analyses are provided in Appendix D.

Hand-motion readouts. We fit ridge readouts to frozen EgoBrain representations for relative hand use and total hand speed, using targets extracted from WiLoR (Potamias et al., 2025) hand tracks in synchronized video. In Figure 4(a,b), exchanging homologous raw EEG channels under the same readout reverses the lateral correlation for relative hand use from 0.096 to -0.091, versus 0.078 to -0.011 for the corresponding control block. The bilateral total-speed correlation changes little (0.463 to 0.451). Both blocks encode both targets, but their contrasting responses support LAT's design of invariant bilateral and equivariant lateral readouts.

Motor-imagery topographies. Figure 4(c–e) visualizes channel-wise feature-energy contrasts on held-out PhysioNet-MI data (Schalk, 2009). The left–right fist contrast is predominantly mirror-antisymmetric in the lateral block (fraction 0.811), compared with 0.152 for the bilateral block and 0.329 for the corresponding control features. The bilateral both-fists versus both-feet contrast is predominantly symmetric. These patterns are consistent with LAT's different treatment of shared and side-specific activity.

4.5 Ablation Studies

Table 3 compares the full model with variants without frame averaging (w/o FA), prediction, or alignment. Removing any of these components reduces average transfer, supporting the combined use of hemispheric structure, predictive supervision, and cross-modal alignment.

Table 3: Ablations of frame averaging (FA), prediction, and alignment. Balanced accuracy for classification and Pearson correlation for regression; mean ± sample s.d. over three seeds. The last row averages per-task relative changes with respect to MWM over the listed tasks.

Task	**MWM (ours)**	w/o FA	w/o prediction	w/o alignment
Classification				
PhysioNet-MI	**0.606±0.007**	**0.606±0.007**	0.600±0.009	0.596±0.004
Schwarz	**0.575±0.035**	0.570±0.015	0.545±0.027	0.549±0.009
Kaya5F	0.269±0.006	0.261±0.004	0.266±0.004	**0.278±0.004**
FineMI	0.200±0.005	**0.201±0.003**	0.198±0.008	0.174±0.004
Ofner2017MI	0.165±0.002	0.161±0.009	0.160±0.007	**0.174±0.003**
Jeong2020	**0.158±0.004**	**0.158±0.004**	0.149±0.006	0.148±0.001
Regression				
ForceGameReg	0.165±0.010	0.164±0.007	**0.171±0.014**	0.107±0.010
SrisReachME	**0.241±0.002**	0.235±0.007	0.240±0.003	0.166±0.000
OfnerApertureME	0.233±0.018	**0.236±0.021**	0.228±0.012	0.056±0.012
WAYGALReg	**0.433±0.002**	0.391±0.003	0.387±0.004	0.311±0.002
BiomechMag	0.157±0.002	0.152±0.002	**0.158±0.000**	0.144±0.001
ForenzoFutVec	**0.127±0.001**	0.118±0.004	0.121±0.003	0.108±0.002
Avg. rel. change	–	−2.3%	−2.6%	−17.6%

4.6 Transfer World Models for Real-Action Conditioning

We evaluate whether the pretrained world models transfer to future prediction conditioned on real actions. On EgoBrain (Lin et al., 2026), synchronized IMU measurements serve as a proxy for real actions in this evaluation. We train the same lightweight IMU-adapter architecture to map these measurements to each model's frozen action interface and predict frozen DINOv2 (Oquab et al., 2024) features 0.2 s ahead under within-recording and held-out-subject splits.

Baselines and Evaluation. We compare MWM with a predictor pretrained using inverse-dynamics latent actions (Garrido et al., 2026) and a randomly initialized predictor. All models use the same inputs, targets, predictor architecture, IMU-adapter architecture, and action interface. Unlike the checkpoint used in the previous sections, MWM and the latent-action model are pretrained separately for each split using only its training portion. We report L1 error, cosine similarity, and the increase in error after shuffling the IMU inputs. Further details are provided in Appendix C.

Table 4: Evaluation of real-action conditioning in held-out next-step prediction, with IMU measurements as the action proxy. All pretrained models use the same predictor and action interface.

	Within recording			Held-out subjects		
Model	L1↓	Cos.↑	Δ_{act} ↑	L1↓	Cos.↑	Δ_{act} ↑
Random init.	0.7334	0.8334	+0.0000	0.7224	0.8407	+0.0000
Latent action	0.6520	0.8463	+0.0020	0.5543	0.8737	+0.0029
MWM (ours)	**0.5009**	**0.8773**	**+0.0079**	**0.4933**	**0.8850**	**+0.0057**

Results. Table 4 shows that MWM achieves lower L1 error and a larger action gap than the latent-action model on both splits. Relative to the latent-action model, these L1 errors represent reductions of 23.2% within recordings and 11.0% on held-out subjects. The randomly initialized predictor has a zero action gap, whereas MWM remains sensitive to matched IMU inputs on both splits.

5 Conclusion

We introduced MWM, which uses EEG as the latent action of a world model to learn motor representations without explicit action annotations. The Lateralization-Aware Transformer incorporates hemispheric correspondence through frame averaging. Relative to the strongest baseline in each task family, the mind model improves task-averaged balanced accuracy by 5.9% and Pearson correlation by 16.2%, while the world model supports future prediction conditioned on real actions. These results support synchronized neural and visual experience as supervision for motor decoding and world modeling.

Limitations. Our world-model evaluation tests action-conditioned prediction error; establishing end-to-end evaluation, such as robotic planning, remains future work.

AI USE STATEMENT

We used generative AI tools solely to edit and polish the manuscript text, edit figures, and adjust the layout. All AI-assisted modifications were reviewed by the authors to ensure that the revised text preserved the intended technical meaning and that figure revisions preserved the underlying data and scientific content. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies motor EEG representation learning and action-conditioned visual prediction using the existing datasets described in Appendix A. We do not anticipate additional ethical concerns specific to the methods and experiments presented here.

REPRODUCIBILITY STATEMENT

The Method section describes the model architecture and training objectives. Appendix A documents the pretraining corpus and all 16 downstream tasks, including preprocessing and data splits. Appendix B provides model configurations, optimization settings, and linear-probe protocols. Additional protocols and metrics for world-model evaluation and lateralization analysis are provided in Appendices C and D, respectively. To facilitate reproduction, anonymized code is provided in the supplementary material.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems*, volume 37, pp. 58757–58791, 2024. doi: 10.52202/079017-1873. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6bdde0373d53d4a501249547084bed43-Abstract-Conference.html.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Robotics: Science and Systems*, 2024. URL https://roboticsproceedings.org/rss20/p045.pdf.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked auto-encoder for EEG representation learning. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL https://openreview.net/forum?id=kttuLV59ZuJ.

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*, volume 38, pp. 78772–78798, 2025. doi: 10.52202/085713-2377. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/66969a9e6bd7a26dfeccea7227178ca7-Abstract-Conference.html.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, pp. 25583–25619, 2025. doi: 10.52202/085713-0760. URL https://papers.neurips.cc/paper_files/paper/2025/hash/20a917f77773ac0fa8bea2bdd6606b66-Abstract-Conference.html.

Dylan Forenzo, Hao Zhu, and Bin He. A continuous pursuit dataset for online deep learning-based EEG brain-computer interface. *Scientific Data*, 11:1256, 2024. doi: 10.1038/s41597-024-04090-6. URL https://www.nature.com/articles/s41597-024-04090-6.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18744–18771. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25u.html.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild, 2026. URL https://arxiv.org/abs/2601.05230.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2450–2462, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Ji-Hoon Jeong, Jeong-Hyun Cho, Kyung-Hwan Shim, Byoung-Hee Kwon, Byeong-Hoo Lee, Do-Yeun Lee, Dae-Hyeok Lee, and Seong-Whan Lee. Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions. *GigaScience*, 9 (10):giaa098, 2020. doi: 10.1093/gigascience/giaa098. URL https://doi.org/10.1093/gigascience/giaa098.

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

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. URL https://openreview.net/forum?id=Io9yFt7XH7.

Murat Kaya, Mustafa Kemal Binli, Erkan Ozbay, Hilmi Yanar, and Yuriy Mishchenko. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific Data*, 5:180211, 2018. doi: 10.1038/sdata.2018.211. URL https://www.nature.com/articles/sdata2018211.

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

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

Qianliang Li, Aliaksandr Dabranau, Kyveli Kompatsiari, and Ivana Konvalinka. Dataset for article: Friendship modulates physical effort and neural dynamics during group coordination in shared social networks, 2026. URL https://doi.org/10.11583/DTU.31369621. Dataset.

Nie Lin, Yansen Wang, Dongqi Han, Weibang Jiang, Jingyuan Li, Ryosuke Furuta, Yoichi Sato, and Dongsheng Li. EgoBrain: Synergizing minds and eyes for human action understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DGcoJINQ7P.

Matthew D. Luciw, Ewa Jarocka, and Benoni B. Edin. Multi-channel EEG recordings during 3,936 grasp and lift trials with varying weight and friction. *Scientific Data*, 1:140047, 2014. doi: 10.1038/sdata.2014.47. URL https://www.nature.com/articles/sdata201447.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorld-Model: Stable end-to-end joint-embedding predictive architecture from pixels, 2026. URL https://arxiv.org/abs/2603.19312.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Chang S. Nam, Yongwoong Jeon, Young-Joo Kim, Insuk Lee, and Kyungkyu Park. Movement imagery-related lateralization of event-related (de)synchronization (ERD/ERS): Motor-imagery duration effects. *Clinical Neurophysiology*, 122(3):567–577, 2011. doi: 10.1016/j.clinph.2010.08.002. URL https://doi.org/10.1016/j.clinph.2010.08.002.

Iyad Obeid and Joseph Picone. The temple university hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016. doi: 10.3389/fnins.2016.00196. URL https://doi.org/10.3389/fnins.2016.00196.

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. doi: 10.1371/journal.pone.0182578. URL https://doi.org/10.1371/journal.pone.0182578.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Marietta Papadatou-Pastou, Eleni Ntolka, Judith Schmitz, Maryanne Martin, Marcus R. Munafò, Sebastian Ocklenburg, and Silvia Paracchini. Human handedness: A meta-analysis. *Psychological Bulletin*, 146(6):481–524, 2020. doi: 10.1037/bul0000229.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Jorge G. Perez-Blanco, Joel C. Huegel, Luis G. Hernández-Rojas, Amanda Valdez-Calderón, Héctor Lizárraga-Torreblanca, David Cruz-Ortiz, Mariana Ballesteros, Manuela Gomez-Correa, and Javier M. Antelis. An EEG-EMG-kinematics dataset from wrist pointing tasks for biomarker research in neurorehabilitation. *Scientific Data*, 13(1):931, 2026. doi: 10.1038/s41597-026-07287-z. URL https://doi.org/10.1038/s41597-026-07287-z.

Gert Pfurtscheller and Christa Neuper. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2–3):65–68, 1997. doi: 10.1016/S0304-3940(97)00889-6. URL https://doi.org/10.1016/S0304-3940(97)00889-6.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12242–12254, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Potamias_WiLoR_End-to-end_3D_Hand_Localization_and_Reconstruction_in-the-wild_CVPR_2025_paper.html.

Omri Puny, Matan Atzmon, Edward J. Smith, Ishan Misra, Aditya Grover, Heli Ben-Hamu, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zIUyj55nXR.

Ander Ramos-Murguialday, Doris Broetz, Massimiliano Rea, Leonhard Läer, Özge Yilmaz, Fabricio L. Brasil, Giulia Liberati, Marco R. Curado, Eliana Garcia-Cossio, Alexandros Vyziotis, Woosang Cho, Manuel Agostini, Ernesto Soares, Surjo Soekadar, Andrea Caria, Leonardo G. Cohen, and Niels Birbaumer. Brain–machine interface in chronic stroke rehabilitation: A controlled study. *Annals of Neurology*, 74(1):100–108, 2013. doi: 10.1002/ana.23879. URL https://doi.org/10.1002/ana.23879.

Gerwin Schalk. EEG Motor Movement/Imagery Dataset. *PhysioNet*, September 2009. doi: 10.13026/C28G6P. URL https://doi.org/10.13026/C28G6P. Version 1.0.0.

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

Andreas Schwarz, Carlos Escolano, Luis Montesano, and Gernot R. Müller-Putz. Analyzing and decoding natural reach-and-grasp actions using gel, water and dry EEG systems. *Frontiers in Neuroscience*, 14:849, 2020. doi: 10.3389/fnins.2020.00849. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.00849/full.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 49126–49162, 2025. doi: 10.52202/085713-1465. URL https://doi.org/10.52202/085713-1465.

Nitikorn Srisrisawang and Gernot R. Müller-Putz. Simultaneous encoding of speed, distance, and direction in discrete reaching: An EEG study. *Journal of Neural Engineering*, 21(6):066042, 2024. doi: 10.1088/1741-2552/ada0ea. URL https://doi.org/10.1088/1741-2552/ada0ea.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Fei-Fei Li, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems*, 2024a. URL https://roboticsproceedings.org/rss20/p043.pdf.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39249–39280, 2024b. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

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*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

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

Qinfan Xiao, Ziyun Cui, Chi Zhang, SiQi Chen, Wen Wu, Andrew Thwaites, Alexandra Woolgar, Bowen Zhou, and Chao Zhang. BrainOmni: A brain foundation model for unified EEG and MEG signals. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46081–46114, 2025. doi: 10.52202/085713-1375. URL https://papers.neurips.cc/paper_files/paper/2025/hash/3aef4a18c2646b9d11d9ab4d9bec72c8-Abstract-Conference.html.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260, 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract.html.

Weibo Yi, Shuang Qiu, Kun Wang, Hongzhi Qi, Lixin Zhang, Peng Zhou, Feng He, and Dong Ming. Evaluation of EEG oscillatory patterns and cognitive process during simple and compound limb motor imagery. *PLOS ONE*, 9(12):e114853, 2014. doi: 10.1371/journal.pone.0114853. URL https://doi.org/10.1371/journal.pone.0114853.

Weibo Yi, Jiaming Chen, Dan Wang, Xinkang Hu, Meng Xu, Fangda Li, Shuhan Wu, and Jin Qian. A multi-modal dataset of electroencephalography and functional near-infrared spectroscopy recordings for motor imagery of multi-types of joints from unilateral upper limb. *Scientific Data*, 12(1):953, 2025. doi: 10.1038/s41597-025-05286-0. URL https://doi.org/10.1038/s41597-025-05286-0.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025a. URL https://proceedings.mlr.press/v267/zhou25t.html.

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*, volume 38, pp. 96859–96904, 2025b. doi: 10.52202/085713-2915. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7e199ad8ae40eb19b2980f61f659cb07-Abstract-Conference.html.

A Dataset Details

We describe the pretraining corpus and the sixteen downstream tasks reported in Tables 1 and 2: seven regression tasks and nine classification tasks. To evaluate transfer to unseen participants, we use participant-disjoint train, validation, and test sets, shared by all models within each task. Participant assignments are specified below. All models receive the same task-specific one-second EEG interval sampled at 200 Hz, with model-specific channel interfaces. EEG values in μV are divided by 100 before model input.

A.1 Pretraining Corpus

EgoBrain EgoBrain (Lin et al., 2026) records EEG, egocentric video, and wearable IMU from 40 participants performing everyday activities. We use the 32 EEG channels and egocentric video from all participants for pretraining. From each 114-channel recording, we retain the fixed set of 32 EEG channels. We then apply a 0.3–75 Hz band-pass filter and a 50 Hz notch filter, resample the signal from 256 to 200 Hz, convert it to μV, and clip it to $\pm 300\,\mu$V. This preprocessing yields 56,950 non-overlapping four-second clips, totaling 63.28 hours. Video frames are resized to a short side of 256 pixels and center-cropped to 224×224. Each pretraining example contains a one-second EEG interval ending at an anchor frame, the anchor frame itself, and the target frame 0.2 s later. All participants are used for pretraining without a validation split, except in the world-model evaluation, where pretraining uses only the training portion of each split. The 4 arm- and wrist-mounted IMU units are not used for pretraining; they are used only for the world-model evaluation in Appendix C.

A.2 Continuous Motor Regression

Except for ForceGameReg, the regression tasks use one-second EEG windows with a 0.2 s stride and predict task-specific force or movement quantities during the interval from 0.2 to 0.4 s after each window. We standardize every target dimension using statistics estimated from up to 8,000 evenly spaced examples from the training split only.

ForceGameReg ForceGameReg uses the Force Game dataset (Li et al., 2026). The study recruited 31 three-person groups; one non-compliant participant was excluded, leaving 64-channel EEG from 92 participants. The released 500 Hz epochs were filtered at 1–40 Hz by the dataset authors. We use EEG from 0.8 to 1.8 s of each trial and predict the mean plateau force from 2 to 4 s. We convert the EEG to μV, resample it to 200 Hz, apply common-average referencing, and reject examples with $\max|\mathrm{EEG}| > 800\,\mu$V. Groups, rather than individual participants, are assigned to splits so that participants sharing a trial remain together. The train, validation, and test splits contain 21,040, 2,681, and 3,556 examples, respectively.

SrisReachME SrisReachME is derived from a discrete center-out reaching study (Srisrisawang & Müller-Putz, 2024). We use 60 EEG channels from 19 participants, apply common-average referencing and a zero-phase 0.5 Hz high-pass filter, and resample the recordings from 500 to 200 Hz. We reject windows with $\max|\mathrm{EEG}| > 800\,\mu$V. The target is the mean of the instantaneous planar reaching-speed magnitudes over the prediction interval. We use 13 participants for training, participants $\{2, 8, 14\}$ for validation, and participants $\{3, 9, 15\}$ for testing, resulting in 511,120, 118,314, and 115,879 examples.

OfnerApertureME and OfnerKinME Both tasks use the same 61-channel motor-execution recordings from 15 participants (Ofner et al., 2017). We interpolate missing EEG values, apply common-average referencing and a zero-phase 0.5 Hz high-pass filter, resample from 512 to 200 Hz, and construct continuous windows that include inter-trial rest. We remove windows with $\max|\mathrm{EEG}| > 800\,\mu$V. For OfnerApertureME, we standardize the 11–13 valid CyberGlove sensors within each participant, low-pass filter them at 1 Hz, and fit the first principal component. The target is the mean absolute rate of change of this component. For OfnerKinME, we derive velocity from three-dimensional hand position sampled at 64 Hz and predict the mean of the instantaneous three-dimensional speed magnitudes. Participants $\{1, 4$–$13\}$, $\{2, 14\}$, and $\{3, 15\}$ form the train, validation, and test splits for both tasks, containing 175,950, 32,327, and 31,577 examples.

WAYGALReg WAYGALReg uses the WAY-EEG-GAL grasp-and-lift recordings from 12 participants with 32 EEG channels at 500 Hz (Luciw et al., 2014). We use the released per-lift EEG

segments, apply a fourth-order zero-phase Butterworth high-pass filter at 0.5 Hz within each lift, and form sliding windows within lifts. Each window is resampled to 200 Hz and common-average referenced; windows with $\max|\mathrm{EEG}| > 800\,\mu\mathrm{V}$ are removed. The two targets are mean absolute grip force and mean signed load force over the future interval. The train, validation, and test splits contain eight, two, and two participants: {2–6, 10–12}, {7, 8}, and {1, 9}, respectively.

BiomechMag BiomechMag records wrist movements from 45 participants (Perez-Blanco et al., 2026) using eight central electrodes: C3, C1, Cz, C2, C4, CP3, CPz, and CP4. We use runs 3–9, apply common-average referencing, and resample the continuous recordings from 512 to 200 Hz. We remove windows with $\max|\mathrm{EEG}| > 800\,\mu\mathrm{V}$. The target is the magnitude of the mean two-dimensional wrist-velocity vector over the prediction interval. A participant-level split based on participant index gives 27, 9, and 9 participants, corresponding to 369,642, 125,663, and 125,654 train, validation, and test examples.

ForenzoFutVec ForenzoFutVec uses the same 28-participant continuous-pursuit motor-imagery BCI recordings as Forenzo2024 (Forenzo et al., 2024), with 62 EEG channels sampled at 1,000 Hz. From each trial start, we construct one-second windows with a 0.2 s stride, resample each window to 200 Hz, apply common-average referencing, and reject windows with $\max|\mathrm{EEG}| > 800\,\mu\mathrm{V}$. We limit each participant to approximately 20,000 windows, retaining the complete run in which this threshold is reached. The two outputs are the mean horizontal and vertical cursor velocities over the future interval, in workspace widths per second. These are closed-loop outputs of the online BCI decoder. The split uses S01–S12 and S21–S28 for training, S13–S16 for validation, and S17–S20 for testing, giving 20, 4, and 4 participants. This task and Forenzo2024 use distinct held-out participant sets.

A.3 Motor Classification

Each classification task uses a one-second EEG window at the task-specific time relative to the cue or movement onset specified below. All splits are participant-disjoint.

PhysioNet-MI PhysioNet-MI (Schalk, 2009) contains 64-channel motor-imagery EEG from 109 participants. We use the runs for left fist, right fist, both fists, and both feet, giving four classes. The EEG is common-average referenced, high-pass filtered at 0.3 Hz, notch filtered at 60 Hz, and resampled to 200 Hz. We extract the four seconds following each cue and provide the first second to the model. The train, validation, and test splits contain 6,300, 1,734, and 1,803 examples from 70, 19, and 20 participants, respectively.

Schwarz Schwarz uses the gel-electrode cohort of the BNCI 2020-001 reach-and-grasp dataset (Schwarz et al., 2020): 15 participants and 58 EEG channels at 256 Hz. We evaluate pre-movement decoding of palmar versus lateral grasp from the $[-2, -1)$ s interval relative to movement onset, during instructed target fixation; rest trials are excluded. We apply common-average referencing and resample to 200 Hz. The train, validation, and test splits contain eleven, two, and two participants: {1, 6–15}, {2, 3}, and {4, 5}, respectively.

Kaya5F Kaya5F uses the five-finger motor-imagery paradigm of Kaya et al. (2018), comprising 19 sessions from eight participants. We retain 19 EEG channels after dropping A1, A2, and X5, map legacy temporal-electrode names to their modern equivalents, and apply common-average referencing. We take the first second from each marker-defined segment. Recordings at 1,000 Hz are resampled to 200 Hz; the other recordings are already at 200 Hz. The five classes correspond to the thumb, index, middle, ring, and little fingers. The train, validation, and test splits contain four, two, and two participants: {A, B, H, I}, {C, E}, and {F, G}, respectively.

Forenzo2024 Forenzo2024 uses continuous-pursuit motor-imagery BCI recordings from 28 participants with 62 EEG channels sampled at 1,000 Hz (Forenzo et al., 2024). Acquisition included a 0.1–200 Hz band-pass filter and a 60 Hz notch filter. We extract non-overlapping one-second windows from each trial start, resample each window to 200 Hz, apply common-average referencing, and reject windows with $\max|\mathrm{EEG}| > 800\,\mu\mathrm{V}$. Labels encode the dominant axis and sign of the mean target-to-cursor displacement within the window, giving right, up, left, and down classes. The split uses S01–S08 and S17–S28 for training, S09–S12 for validation, and S13–S16 for testing, giving 20, four, and four participants. The recordings combine the Main and Transfer-Learning study cohorts.

FineMI FineMI (Yi et al., 2025) contains eight classes of unilateral upper-limb motor imagery from 18 participants and 62 EEG channels. The released four-second epochs were common-average referenced and filtered at 4–40 Hz by the dataset authors. Following the benchmark preprocessing, we convert the signals to μV, apply common-average referencing, resample from 250 to 200 Hz, and use the first second of each epoch. The train, validation, and test splits contain 3,840, 960, and 960 examples from 12, 3, and 3 participants, respectively.

Ofner2017ME and Ofner2017MI Ofner2017 (Ofner et al., 2017) provides matched motor-execution and motor-imagery sessions from 15 participants with 61 EEG channels. Both tasks classify elbow flexion, elbow extension, forearm supination, forearm pronation, hand closing, hand opening, and rest. We apply common-average referencing, high-pass filter at 0.5 Hz, resample from 512 to 200 Hz, and reject trials with $\max|\mathrm{EEG}| > 800\,\mu$V. The model receives the first second of the three-second interval following the cue. Participants $\{1, 4\text{–}13\}$, $\{2, 14\}$, and $\{3, 15\}$ form the train, validation, and test splits. These splits contain 4,452, 839, and 773 examples for motor execution and 4,601, 839, and 838 examples for motor imagery.

Weibo2014 Weibo2014 (Yi et al., 2014) contains six classes of simple or compound limb motor imagery and one rest class, recorded from 10 participants with 60 EEG channels. We evaluate the last second of motor imagery, $[6, 7)$ s relative to trial onset. We apply common-average referencing. The train, validation, and test splits contain 3,294, 1,118, and 1,118 examples from 6, 2, and 2 participants, respectively.

Jeong2020 Jeong2020 uses the motor-imagery recordings of participants 1–15 from Jeong et al. (2020). After removing EOG and EMG channels, we retain 60 EEG channels, apply common-average referencing, and resample from 1,000 to 200 Hz. The eleven classes comprise six reaching directions, three grasp types, and two wrist-twisting movements. The model receives the first second after the cue. Participants $\{1, 4\text{–}13\}$, $\{2, 14\}$, and $\{3, 15\}$ form the train, validation, and test splits, containing 16,850, 3,350, and 3,300 examples.

B Implementation Details

Model configuration. The EEG encoder uses 0.2 s temporal patches, a token width of 40, and 12 Transformer blocks with eight attention heads and feed-forward width 800. The first ten blocks produce the representation used for downstream evaluation, while the final two serve as the reconstruction decoder during pretraining. A global channel and a global temporal token augment the token grid. For C EEG channels, the one-second downstream representation has shape $(C + 1) \times 6 \times 40$. The visual encoder is DINOv2-Base with 224×224 inputs and a 16×16 grid of 768-dimensional patch features. The frame predictor has two Transformer layers, width 512, eight attention heads, feed-forward width 1,024, and dropout 0.1, with a 128-dimensional action interface.

Pretraining. We train MWM in FP32 using AdamW with a batch size of 128, a learning rate of 5×10^{-4}, weight decay of 0.05, and gradient clipping at 0.8. The learning rate follows an iteration-wise cosine schedule to 10^{-5} without warmup. All reported downstream MWM evaluations use the same fixed pretrained checkpoint. For full-model pretraining, we set $\lambda_{\mathrm{align}} = \lambda_{\mathrm{pred}} = 1$.

Linear-probe evaluation. We freeze each backbone in evaluation mode and train only a linear prediction head. For MWM, we discard the visual encoder and frame predictor, disable stochastic mirrors, and flatten the frozen EEG token grid before applying the head. For regression, we additionally apply a non-affine LayerNorm and standardize each target dimension using training-split statistics.

We train all probes with AdamW using a learning rate of 5×10^{-5}, weight decay of 0.01, a batch size of 64, an iteration-wise cosine schedule to 10^{-6}, and gradient clipping at 1. Classification probes are trained for 80 epochs with cross-entropy and 0.1 label smoothing, while regression probes are trained for 10 epochs with MSE. We select checkpoints by validation balanced accuracy for classification and validation Pearson correlation for regression.

C World-Model Evaluation Details

Protocol IMU measurements serve as a proxy for real actions in this evaluation. Each example contains frozen DINOv2 patch features for the current frame and the frame $0.2\,\mathrm{s}$ later, together with three consecutive $0.2\,\mathrm{s}$ IMU summaries from four arm- and wrist-mounted sensors. We freeze each frame predictor and train only an IMU adapter, implemented as a three-layer MLP with 512-dimensional hidden features and LayerNorm. Inputs are standardized on the training set, and the adapter is trained for 4,000 steps with AdamW, a batch size of 64, and a learning rate of 10^{-3}. All models use the same inputs, targets, adapter architecture, and training anchors.

Splits The within-recording split reserves the final eleventh of each recording for evaluation and separates it from training by a $300\,\mathrm{s}$ guard interval. The held-out-subject split evaluates on P0006, P0016, P0026, and P0036. For each split, MWM and the latent-action model are pretrained only on the training portion.

Comparisons All compared predictors share the same frame-predictor architecture and an identical 128-dimensional action interface, controlling for architecture and action dimensionality. For pretraining, MWM uses one second of EEG ending at the current frame. The latent-action model (Garrido et al., 2026) instead derives its action from the current and target frames with an inverse-dynamics encoder. We also evaluate a randomly initialized predictor.

Metrics We report the mean L1 error and patch-wise cosine similarity between predicted and target features. To measure use of the action inputs, we fix the current frame, shuffle IMU inputs among held-out examples from the same subject, and compute

$$\Delta_{\mathrm{act}} = E_{\mathrm{shuf}} - E_{\mathrm{true}}, \tag{12}$$

where E_{true} and E_{shuf} are the L1 errors under matched and shuffled action inputs. A positive action gap indicates that matched inputs improve prediction.

D Lateralization Analysis Details

We compare one fixed pretrained MWM model with the *w/o FA* variant, which removes the frame-averaging (FA) wrapper, leaving a plain Transformer. Both encoders are frozen, and preprocessed EEG inputs are multiplied by $1/100$ before encoding.

For a token $\mathbf{h} \in \mathbb{R}^{D_e}$, we write $\mathbf{h} = [\mathbf{h}^{(1)}; \mathbf{h}^{(2)}]$, where each block lies in $\mathbb{R}^{D_e/2}$. In MWM, these blocks correspond to bilateral and lateral features, respectively; in the control model, they are arbitrary partitions of the feature dimension.

D.1 Hand-Motion Readouts

Data and targets. We uniformly sample 300 one-second windows from each of the 40 EgoBrain participants (Lin et al., 2026), retaining the 32-channel montage at 200 Hz. We obtain 3-D hand tracks from synchronized video using WiLoR (Potamias et al., 2025) and compute ego-motion-compensated left- and right-hand speeds. Each window contains five $0.2\,\mathrm{s}$ frame slots. We retain windows with at least three valid slots for each hand, yielding 9,319 windows, and average the valid speeds within each window to obtain L and R. The two targets are relative hand use $y_{\mathrm{side}} = \log(R+\epsilon) - \log(L+\epsilon)$ and total hand speed $y_{\mathrm{speed}} = \log(R+L+\epsilon)$, with $\epsilon = 10^{-3}$. Both targets are centered within each participant. All participants' EEG and video were available during pretraining; the hand labels were used only for this readout analysis.

Features and evaluation. We fit separate ridge regressions to the global token $\mathbf{h}$ and its two blocks $\mathbf{h}^{(1)}$ and $\mathbf{h}^{(2)}$. Features are centered within each participant and standardized using training-fold statistics. We use five participant-disjoint folds, selecting the ridge penalty from $\{0.1, 1, 10, 100\}$ by nested participant-grouped cross-validation. We report Pearson correlation after pooling the held-out predictions from all folds. Confidence intervals use 1,000 participant-bootstrap samples, shared across model and readout comparisons.

Channel-exchange diagnostic. For each fold, we apply the fitted readout to features obtained after exchanging all 14 homologous electrode pairs in the raw EEG, leaving four midline channels and the

Table 5: Mirror-antisymmetric fraction A of the left-fist minus right-fist energy contrast.

Model	Feature block	A
MWM	Lateral	0.811
MWM	Bilateral	0.152
w/o FA	$\mathbf{h}^{(2)}$	0.329
w/o FA	$\mathbf{h}^{(1)}$	0.250

targets unchanged. This diagnostic acts on raw channels; the training operation $\mathcal{P}$ instead exchanges only the learned lateral feature component.

D.2 Motor-Imagery Topographies

Feature-energy contrasts. We use the first second after the cue from the 1,803 PhysioNet-MI test trials of 20 participants (Schalk, 2009): 451 left-fist, 449 right-fist, 450 both-fists, and 453 both-feet trials. Inputs retain 64 channels at 200 Hz with a 0.3 Hz high-pass filter. No prediction head is fitted. For each trial and channel, we sum squared patch-token values over the N temporal patches and $D_e/2$ features of the selected block. After centering within participant, we compute each participant's mean left-fist minus right-fist energy contrast and then average across participants with equal weight. We similarly contrast both-fists with both-feet imagery using $\mathbf{h}^{(1)}$.

Spatial symmetry. Let $\mathbf{m} \in \mathbb{R}^{64}$ be the group-mean energy contrast before channel-wise normalization, and let Π exchange the 27 homologous channel pairs while fixing the ten midline channels. We measure the fraction of mirror-antisymmetric energy over all channels as

$$A(\mathbf{m}) = \frac{\|\mathbf{m} - \Pi\mathbf{m}\|_2^2}{\|\mathbf{m} - \Pi\mathbf{m}\|_2^2 + \|\mathbf{m} + \Pi\mathbf{m}\|_2^2}. \tag{13}$$

Figure 4(c–e) displays channel-normalized maps, while Table 5 reports A computed from the unnormalized contrasts.