NEUROACTION: ALIGNING EEG TO DISCRETE MOTION PRIMITIVES FOR CONTINUOUS MOTION DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-to-motion decoding remains challenging because noisy scalp EEG is difficult to map directly to continuous kinematics. We present NeuroAction, a framework that aligns EEG representations to discrete motion primitives for continuous motion decoding across heterogeneous paradigms, including full-body free movement, lower-limb treadmill locomotion, and upper-limb freewill reaching and grasping. A VQ-based motion prior discretizes continuous motion into reusable primitives that serve as both alignment prototypes and intermediate prediction targets. Prototypical and supervised contrastive objectives align EEG embeddings to motion-derived prototypes and organize them according to primitive assignments. An E2M Transformer predicts motion-code sequences from the learned EEG representations, while a final geometric refinement stage improves continuous trajectory reconstruction through motion-level training. Under a last-session test protocol, NeuroAction achieves the best mean reconstruction performance among the compared EEG decoding baselines across all three datasets, with the clearest gains on full-body motion in both reconstruction accuracy and distribution-level quality. Multi-subject ablations further support the contributions of EEG-to-primitive alignment, token prediction, and refinement. Code will be released publicly upon acceptance.

1 INTRODUCTION

Recent years have seen significant advances in brain-computer interfaces (BCIs), enabling applications ranging from communication restoration and neurorehabilitation to closed-loop neuromodulation and human-computer interaction. Among these applications, a central goal is to translate neural activity into actionable motor commands for external effectors, such as robotic arms, prosthetic limbs, exoskeletons, and humanoid agents. Achieving this goal requires robust EEG-to-motion decoding: inferring continuous human movement from non-invasive brain signals.

Despite steady progress, EEG-to-motion decoding remains highly challenging. Most existing methods are developed for specific motor paradigms, such as hand movement, reaching, or simplified gait cycles, and typically regress low-level kinematic variables directly from EEG Bradberry et al. (2010). While effective in constrained laboratory settings, such direct regression becomes difficult to scale across heterogeneous motor behaviors, long temporal horizons, and cross-session variations. A key reason is that EEG-to-motion decoding is not a strict one-to-one mapping. Scalp EEG provides noisy and spatially coarse measurements of population-level neural dynamics, rather than precise muscle-level commands. Neurophysiological studies further suggest that motor control is organized hierarchically, relying on internal forward models and modular motor primitives instead of explicitly specifying all joint trajectories Wolpert et al. (1998); Mussa-Ivaldi (1999). Optimal feedback control theory also suggests that neural control prioritizes task-relevant variables rather than exact reproduction of detailed kinematics Todorov & Jordan (2002). These observations motivate using motion-derived primitives as structured targets for EEG representation learning and as an intermediate representation for continuous motion reconstruction.

Meanwhile, EEG-motion study is moving from controlled single-task settings toward naturalistic and heterogeneous behaviors, including treadmill walking with closed-loop BCI control He et al. (2018), freewill reaching and grasping Thapa et al. (2025), and full-body locomotion across complex

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.

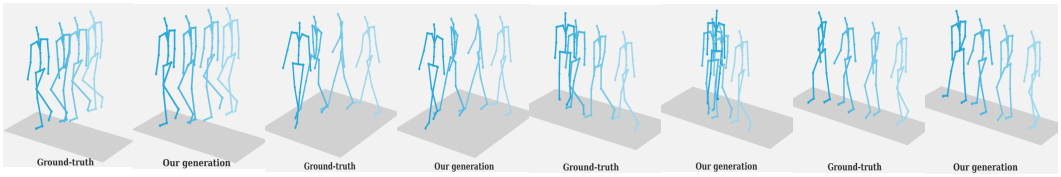

Figure 1: Qualitative EEG-to-motion decoding examples on FBM (subject S10, session 10). We show four representative locomotion patterns: forward walking, downstairs, right turn, and left turn. Each example compares the ground-truth motion with the NeuroAction-generated motion side by side. Each segment contains 40 frames, and we visualize one frame every 10 frames for clarity.

terrains Brantley et al. (2018). The shift requires that EEG-to-motion frameworks go beyond task-specific regressors and can be applied across upper-limb, lower-limb, and full-body motor paradigms.

Recent progress in motion generation and vision-language-action modeling suggests that continuous motion can be discretized into reusable tokens. Vector-quantized autoencoders have been widely used to tokenize high-dimensional human motion for Transformer-based generation van den Oord et al. (2017); Guo et al. (2022b); Zhang et al. (2023), and VQ-VLA further shows that quantization is also useful for low-dimensional action trajectories. However, EEG-to-motion decoding requires a different use of motion quantization: EEG is noisy, subject-dependent, and not an explicit symbolic command, but it is temporally aligned with motion. We therefore use motion codes both as structured motion-primitive prototypes for EEG representation and as discrete intermediate targets for reconstruction.

To this end, we propose NeuroAction, a framework that aligns EEG representations to discrete motion primitives for continuous motion decoding. NeuroAction learns dataset-specific primitives with a VQ tokenizer, aligns EEG embeddings to their codebook representations through prototypical and supervised contrastive learning, predicts motion-code sequences with an E2M Transformer, and refines decoded trajectories through continuous motion-level supervision. We evaluate NeuroAction on three EEG-motion datasets covering full-body movement, treadmill locomotion, and freewill reaching and grasping. Under a strict last-session test protocol, NeuroAction achieves the best mean reconstruction performance among the compared EEG decoding baselines across all three datasets, with the clearest gains on full-body motion. Multi-subject ablations further support the contributions of EEG-to-primitive alignment, token prediction, and geometric refinement.

Our key contributions are summarized as follows:

- We adopt EEG-to-primitive alignment, using motion-derived codebook prototypes and primitive assignments to structure continuous EEG representations via prototypical and contrastive learning.
- We develop NeuroAction, a hierarchical EEG-to-motion framework that integrates motion tokenization, representation alignment, and motion-level refinement for continuous motion reconstruction.
- Across three heterogeneous EEG-motion datasets, NeuroAction achieves the best mean reconstruction performance over baselines under a strict last-session test protocol, with the clearest gains on full-body motion and multi-subject ablations supporting alignment and staged decoding.

2 Method

2.1 Problem Formulation

Let $X \in \mathbb{R}^{C \times T_e}$ be an EEG segment with C electrodes and T_e time samples, and let $Y \in \mathbb{R}^{T_m \times D}$ be its paired motion sequence, where D depends on the target motion representation. NeuroAction employs a discrete latent motion sequence $S = (s_1, \ldots, s_N)$, where each $s_i \in \{1, \ldots, K\}$ indexes one code from a learned motion codebook of size K. Each code corresponds to a short kinematic primitive in the motion manifold. We denote EEG-to-motion decoding as a latent generation process:

$$p(Y \mid X) \approx \sum_{S} p_\theta(Y \mid S)\, p_\phi(S \mid X), \tag{1}$$

where $p_\phi(S \mid X)$ predicts motion primitives from EEG, and $p_\theta(Y \mid S)$ reconstructs continuous motion from the selected primitives.

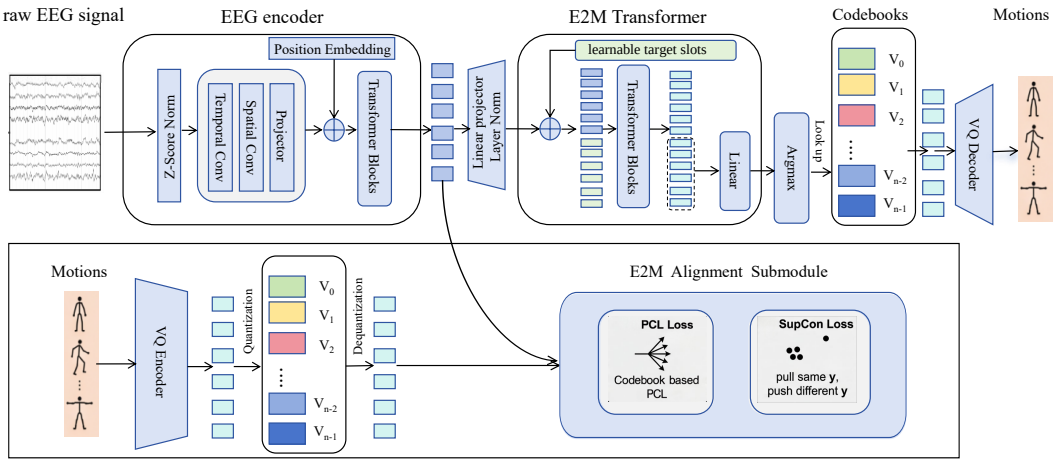

Figure 2: NeuroAction architecture. EEG segments are encoded into continuous representations trained to align with discrete motion primitives learned by a VQ tokenizer. The E2M Transformer predicts motion-code sequences from these representations, and the motion decoder reconstructs continuous trajectories from the selected code vectors.

Under this formulation, EEG provides coarse neural evidence for selecting motion primitives, while the motion decoder synthesizes the final continuous trajectory from the learned motion prior. This yields a unified interface for heterogeneous EEG-motion settings with different output dimensions and temporal resolutions.

2.2 Discrete Motion Prior

We first train a lightweight CNN-based VQ-VAE as the motion tokenizer. Given a motion sequence Y, the encoder maps it to latent vectors $Z = (z_1, \ldots, z_N)$, where $z_i \in \mathbb{R}^{d_c}$. A codebook $\mathcal{C} = \{c_1, \ldots, c_K\}$ defines a discrete set of motion primitives. Each latent vector is assigned to its nearest codebook entry, i.e., $s_i = \arg\min_k \|z_i - c_k\|_2^2$, and the quantized latent is $\hat{z}_i = c_{s_i}$.

The VQ decoder reconstructs the continuous motion sequence from the quantized latent sequence $\hat{Z} = (\hat{z}_1, \ldots, \hat{z}_N)$. The tokenizer is trained with reconstruction and commitment losses:

$$\hat{Y} = D_{\mathrm{VQ}}(\hat{Z}), \qquad \mathcal{L}_{\mathrm{VQ}} = \|Y - \hat{Y}\|_1 + \beta\|Z - \mathrm{sg}[\hat{Z}]\|_2^2, \tag{2}$$

where $\mathrm{sg}[\cdot]$ is stop-gradient. The codebook is updated using EMA statistics of assigned encoder outputs, with dead-code reset to mitigate codebook collapse. After training, the VQ decoder and codebook define a discrete motion-prior space for EEG decoding.

2.3 EEG Encoder

Given a batch of EEG segments $X \in \mathbb{R}^{B\times C\times T}$, the EEG encoder produces token-level neural embeddings $H = f_{\mathrm{EEG}}(X) \in \mathbb{R}^{B\times N_e\times d_e}$. The EEG token count N_e is matched to the motion-token count N_m; we use $N_e = N_m = N$ in the alignment and loss equations. Each EEG token corresponds to a motion-primitive position, and unmasked encoder attention gives it full-window context. The EEG encoder first applies channel-wise z-score normalization, then uses a lightweight convolutional patch embedding to form temporal EEG tokens, followed by a Transformer encoder Vaswani et al. (2017) to model temporal dependencies:

$$\tilde{X} = \mathrm{ZScore}(X), \qquad U = \mathrm{PatchEmbed}(\tilde{X}), \qquad H = \mathrm{TransEnc}(U). \tag{3}$$

2.4 EEG-to-Motion Token Decoder

The EEG-to-motion decoder predicts motion codes from EEG embeddings H. An adapter maps H to $H^{\mathrm{e2m}} \in \mathbb{R}^{B\times N_e\times d_m}$, where d_m is the E2M hidden width. Learnable target slots $Q \in \mathbb{R}^{N_m\times d_m}$

are expanded across the batch to $Q^B \in \mathbb{R}^{B\times N_m\times d_m}$. Concatenation is along the sequence axis:

$$\begin{aligned} U^{\mathrm{e2m}} &= [H^{\mathrm{e2m}}; Q^B], \qquad Z^{\mathrm{e2m}} = \mathcal{T}_{\mathrm{e2m}}(U^{\mathrm{e2m}}), \\ P_{b,i,:} &= \mathrm{Linear}(Z^{\mathrm{e2m}}_{b,N_e+i,:}), \qquad i = 1,\dots,N_m. \end{aligned} \tag{4}$$

Here $Z^{\mathrm{e2m}} \in \mathbb{R}^{B\times(N_e+N_m)\times d_m}$ and $P \in \mathbb{R}^{B\times N_m\times K}$, with the motion-token logits read from the target-slot positions.

The E2M Transformer applies a causal attention mask over the full ordered input sequence $[H^{\mathrm{e2m}}_1,\dots,H^{\mathrm{e2m}}_{N_e}, Q_1,\dots,Q_{N_m}]$. Slot Q_j can attend to all EEG tokens and slots $Q_1,\dots,Q_j$, but not future slots. Since the upstream EEG encoder is bidirectional, each conditioning token already contains full-window context. All logits are produced in one forward pass: slots are dependent, but predicted IDs are never fed back as inputs. Appendix E.2 gives the mask and tensor shapes.

The predicted token is selected by maximum logit, and the token prediction loss is the cross-entropy against the ground-truth motion tokens:

$$\hat{s}_{b,i} = \arg\max_k P_{b,i,k}, \qquad \mathcal{L}_{\mathrm{tok}} = -\frac{1}{BN}\sum_{b=1}^{B}\sum_{i=1}^{N}\log\frac{\exp(P_{b,i,s_{b,i}})}{\sum_{k=1}^{K}\exp(P_{b,i,k})}. \tag{5}$$

This loss trains the E2M decoder to generate the discrete motion-token sequence obtained from the paired ground-truth motion.

2.5 EEG-to-Primitive Alignment

In addition to token prediction, we align continuous EEG embeddings to motion-codebook prototypes and organize them according to the primitive assignments of paired motion. Let $h_{b,i}$ denote the EEG embedding at token position i of sample b, and let $s_{b,i}$ be the corresponding ground-truth motion token obtained from the frozen motion tokenizer. The configured alignment width satisfies $d_e = d_c$, so the EEG embeddings and motion-code vectors can be compared directly. We normalize EEG embeddings and codebook vectors, and define the codebook-based prototypical contrastive loss as:

$$\bar{h}_{b,i} = \frac{h_{b,i}}{\|h_{b,i}\|_2}, \qquad \bar{c}_k = \frac{c_k}{\|c_k\|_2}, \qquad \mathcal{L}_{\mathrm{pcl}} = -\frac{1}{BN}\sum_{b=1}^{B}\sum_{i=1}^{N}\log\frac{\exp(\bar{h}_{b,i}^{\top}\bar{c}_{s_{b,i}}/\tau_c)}{\sum_{k=1}^{K}\exp(\bar{h}_{b,i}^{\top}\bar{c}_k/\tau_c)}. \tag{6}$$

We further apply supervised contrastive learning over EEG tokens using the same motion-token labels. Let $m = (b,i)$ index each EEG token instance, with embedding $\bar{h}_m$ and label s_m. For each anchor m, the positive set is $\mathcal{P}(m) = \{n \neq m \mid s_n = s_m\}$, and valid anchors are collected in $\mathcal{V} = \{m : |\mathcal{P}(m)| > 0\}$. The supervised contrastive loss is:

$$\mathcal{L}_{\mathrm{supcon}} = -\frac{1}{|\mathcal{V}|}\sum_{m\in\mathcal{V}}\frac{1}{|\mathcal{P}(m)|}\sum_{p\in\mathcal{P}(m)}\log\frac{\exp(\bar{h}_m^{\top}\bar{h}_p/\tau_s)}{\sum_{n\neq m}\exp(\bar{h}_m^{\top}\bar{h}_n/\tau_s)}. \tag{7}$$

The PCL objective aligns each EEG embedding to the motion-code prototype indexed by its paired motion token. SupCon further encourages EEG embeddings sharing the same primitive label to form compact and separable clusters. Together, these objectives structure the EEG representation space for subsequent motion-token prediction.

2.6 Differentiable Motion-Level Refinement

After EEG-to-motion token generation, NeuroAction performs a final geometric refinement stage to improve continuous trajectory accuracy. Token-level supervision encourages correct primitive prediction, but it does not directly optimize the decoded motion sequence. We therefore introduce a differentiable decoding path from token logits to continuous motion: given token logits $P \in \mathbb{R}^{B\times N\times K}$, we use straight-through Gumbel-Softmax to obtain differentiable hard token selections:

$$G_{b,i} = \mathrm{GumbelSoftmax}(P_{b,i}; \tau, \mathrm{hard} = \mathrm{True}), \qquad G_{b,i} \in \mathbb{R}^K. \tag{8}$$

The forward selections are one-hot, while the backward pass uses relaxed gradients. Thus $G \in \mathbb{R}^{B\times N_m\times K}$ retrieves code embeddings of shape $B \times N_m \times d_c$. The selected motion-code embedding

is computed as $\tilde{z}_{b,i} = G_{b,i}\mathcal{C}$, where $\mathcal{C} \in \mathbb{R}^{K \times d_c}$ is the motion codebook. The VQ decoder then reconstructs the continuous motion $\tilde{Y} = D_{\mathrm{VQ}}(\tilde{Z})$. The refinement stage is optimized only with a frame-level MSE loss:

$$\mathcal{L}_{\mathrm{GR}} = \mathrm{MSE}(\tilde{Y}, Y). \tag{9}$$

During this stage, gradients flow through the straight-through token-selection path and the VQ decoder, allowing the EEG encoder, E2M Transformer, and motion decoder to be refined jointly. This completes the coarse-to-fine hierarchy by converting discrete primitive predictions into more accurate continuous motion trajectories.

2.7 Training Procedure

NeuroAction is trained in three stages. In Stage 1, we train the motion VQ-VAE on motion sequences only, producing a dataset-specific motion codebook and decoder that define the discrete motion-prior space. Each paired motion sequence is then quantized into a token sequence using nearest-neighbor codebook assignment.

In Stage 2, the motion tokenizer is frozen, and the EEG encoder and E2M Transformer are trained to predict motion tokens from EEG. The training objective combines token prediction with codebook-based PCL and supervised contrastive alignment:

$$\mathcal{L}_{\mathrm{stage2}} = \mathcal{L}_{\mathrm{tok}} + \lambda_{\mathrm{pcl}}\mathcal{L}_{\mathrm{pcl}} + \lambda_{\mathrm{supcon}}\mathcal{L}_{\mathrm{supcon}}. \tag{10}$$

In Stage 3, we initialize from the Stage 2 model and apply the geometric refinement objective in Sec. 2.6. The VQ encoder and codebook remain fixed, while the EEG encoder, E2M Transformer, and VQ decoder are updated using continuous motion-level supervision. Overall, training proceeds from learning discrete motion primitives to EEG-to-primitive alignment and token prediction, followed by continuous motion-level refinement.

2.8 Inference

At inference time, only EEG is required. Given an EEG segment, the EEG encoder first extracts neural embeddings, and the E2M Transformer predicts a discrete motion-token sequence. Hard argmax token IDs are mapped to their frozen codebook embeddings and decoded by the VQ decoder into the final continuous motion trajectory. Geometric refinement is a training stage, so inference requires no iterative motion optimization.

3 Experiments

3.1 Experimental Settings

Datasets. We evaluate NeuroAction on three public EEG-motion datasets spanning heterogeneous motor paradigms: full-body multi-terrain locomotion (FBM) Brantley et al. (2018), treadmill lower-limb locomotion (TM) He et al. (2018), and freewill reaching and grasping (FR) Thapa et al. (2025). For FBM, we use subjects `s02`–`s10`, excluding `s01` due to synchronization issues, and 59 EEG channels after removing `FCz`; for TM, we evaluate six subjects using all 60 channels, excluding `S03` and `S04` due to data quality issues; and for FR, we use all 31 channels from `s05`, `s07`, `s13`, `s15`, and `s22`, the five subjects with complete three-session recordings. Dataset acquisition schedules and the distinction between cross-day and session-disjoint evaluation are detailed in Appendix B.1.

Motion representation. We use dataset-specific motion targets according to the available kinematic signals. For FBM, we first select 22 joints from the original 23-joint representation by removing the L3 joint and reorder them to match the SMPL-style HumanML3D joint convention; the resulting motion is then converted to the 263-D HumanML3D-style representation. For TM, the target is a 6-D lower-limb joint-angle sequence measured from bilateral hip, knee, and ankle goniometers. For FR, the target is the synchronized three-axis wrist acceleration sequence.

Data split. We use a strict session-wise split to avoid temporal leakage. For each subject, the last available session is used for testing, while earlier sessions are used for training. The motion VQ-VAE is trained separately for each dataset using only kinematic data from the training sessions; held-out EEG-to-motion test sessions are never used for VQ-VAE training.

Table 1: Comparison with adapted EEG-to-motion baselines under the last-session test protocol. Values are cohort mean ± between-subject standard deviation. FBM PCC/R^2/RMSE evaluate 30-D end-effector positions; FID and Diversity use full 263-D motion. GT denotes real-motion diversity. Subject-level results and paired tests are reported in Appendix D.6.

	FBM ($n = 9$)					**FR** ($n = 5$)			**TM** ($n = 6$)		
Method	PCC↑	R^2 ↑	RMSE↓	FID↓	DIV→	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓
GT	–	–	–	–	$5.86_{\pm 2.93}$	–	–	–	–	–	–
MTRT	$0.033_{\pm 0.012}$	$-0.223_{\pm 0.216}$	$0.898_{\pm 0.251}$	$33.3_{\pm 27.3}$	$3.43_{\pm 1.81}$	$0.505_{\pm 0.242}$	$0.111_{\pm 0.365}$	$0.254_{\pm 0.044}$	$0.071_{\pm 0.154}$	$-0.058_{\pm 0.107}$	$0.468_{\pm 0.025}$
DeepConvNet	$0.032_{\pm 0.041}$	$-0.262_{\pm 0.154}$	$0.951_{\pm 0.231}$	$62.5_{\pm 33.7}$	$2.92_{\pm 1.60}$	$0.506_{\pm 0.232}$	$-0.015_{\pm 0.468}$	$0.264_{\pm 0.040}$	$0.389_{\pm 0.232}$	$0.024_{\pm 0.239}$	$0.443_{\pm 0.055}$
STTransformer	$0.017_{\pm 0.008}$	$-0.365_{\pm 0.244}$	$0.941_{\pm 0.257}$	$38.1_{\pm 33.3}$	$2.52_{\pm 2.53}$	$0.489_{\pm 0.215}$	$-0.066_{\pm 0.533}$	$0.272_{\pm 0.051}$	$0.377_{\pm 0.229}$	$0.063_{\pm 0.123}$	$0.434_{\pm 0.047}$
NeuroDyGait	$0.049_{\pm 0.028}$	$-0.212_{\pm 0.203}$	$0.897_{\pm 0.251}$	$65.2_{\pm 38.2}$	$3.94_{\pm 1.34}$	$0.607_{\pm 0.164}$	$\underline{0.162}_{\pm 0.370}$	$\underline{0.240}_{\pm 0.033}$	$0.450_{\pm 0.165}$	$\underline{0.127}_{\pm 0.134}$	$\underline{0.418}_{\pm 0.039}$
EEGConformer	$0.026_{\pm 0.013}$	$-0.307_{\pm 0.133}$	$0.930_{\pm 0.233}$	$22.2_{\pm 23.4}$	$2.14_{\pm 2.12}$	$0.474_{\pm 0.235}$	$-0.154_{\pm 0.574}$	$0.280_{\pm 0.048}$	$0.333_{\pm 0.241}$	$0.032_{\pm 0.129}$	$0.443_{\pm 0.041}$
ContraWR	$0.026_{\pm 0.031}$	$-0.261_{\pm 0.145}$	$0.904_{\pm 0.229}$	$56.9_{\pm 28.4}$	$2.82_{\pm 1.56}$	$0.513_{\pm 0.306}$	$-0.006_{\pm 0.481}$	$0.266_{\pm 0.050}$	$-0.006_{\pm 0.015}$	$-0.058_{\pm 0.089}$	$0.470_{\pm 0.028}$
Ours-stage2	$\underline{0.412}_{\pm 0.241}$	$\underline{-0.117}_{\pm 0.392}$	$\underline{0.821}_{\pm 0.245}$	$\mathbf{3.17}_{\pm 2.04}$	$\mathbf{4.48}_{\pm 2.90}$	$\underline{0.643}_{\pm 0.146}$	$0.094_{\pm 0.525}$	$0.247_{\pm 0.052}$	$\underline{0.459}_{\pm 0.224}$	$-0.102_{\pm 0.308}$	$0.463_{\pm 0.077}$
Ours-stage3	$\mathbf{0.522}_{\pm 0.238}$	$\mathbf{0.110}_{\pm 0.354}$	$\mathbf{0.718}_{\pm 0.229}$	$\underline{4.14}_{\pm 4.04}$	$\underline{4.36}_{\pm 2.89}$	$\mathbf{0.657}_{\pm 0.152}$	$\mathbf{0.250}_{\pm 0.340}$	$\mathbf{0.230}_{\pm 0.038}$	$\mathbf{0.485}_{\pm 0.211}$	$\mathbf{0.184}_{\pm 0.148}$	$\mathbf{0.404}_{\pm 0.052}$

Training details. Across datasets, kinematic signals are low-pass filtered at 5 Hz, downsampled to 20 Hz, and temporally aligned with EEG, while EEG is band-pass filtered at 0.1–48 Hz, re-referenced using common average reference (CAR) Delorme & Makeig (2004), and downsampled to 100 Hz. Paired signals are segmented into 2 s windows with a 0.5 s stride. TM and FR motion targets are min-max normalized to $[0, 1]$, Channel-wise z-score normalization is applied to EEG as the first operation of the EEG encoder. The default VQ codebook sizes are 32/16/12 and the motion-token counts are 10/10/5 for FBM/TM/FR, respectively. Dataset-specific model widths, optimizer settings, learning-rate schedules, iteration budgets, and training costs are provided in Appendix B.2.

3.2 Comparison to State-of-the-Art Approaches

Baselines. Since EEG-to-motion decoding has been studied under different motor paradigms and output formats, we compare NeuroAction with representative baselines including the upper-limb decoder MTRT Wang et al. (2023), the gait-specific NeuroDyGait (DY) Fu et al. (2025), and general-purpose EEG models EEGConformer (EC) Song et al. (2023), STTransformer (ST), DeepConvNet (DCN) Schirrmeister et al. (2017), and ContraWR Yang et al. (2023). All methods use the same dataset splits, preprocessing, and prediction targets; detailed baseline adaptation and training controls are provided in Appendix B.4. Table 1 reports cohort-level mean±std performance under the last-session test protocol.

Overall performance and uncertainty. Figures 4 and 3 complement the aggregate metrics with representative qualitative reconstructions across the three datasets. NeuroAction captures the dominant temporal structure of full-body motion on FBM and follows the major joint-angle and wrist-acceleration trends on TM and FR, while fine-scale deviations remain visible.

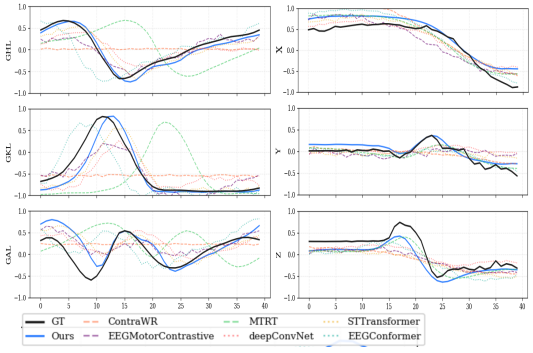

Figure 3: Comparison on TM (GHL/GKL/GAL, left hip/knee/ankle z-axis angles) and FR (right X/Y/Z wrist accelerations). NeuroAction better follows ground-truth trends and amplitudes.

Table 1 shows that NeuroAction Stage 3 attains the highest mean reconstruction performance among the compared methods on all three datasets. On FBM, PCC increases from NeuroDyGait's 0.049 ± 0.028 to 0.522 ± 0.238, R^2 from -0.212 ± 0.203 to 0.110 ± 0.354, and RMSE decreases from 0.897 ± 0.251 to 0.718 ± 0.229. Positive mean R^2 implies non-trivial variance capture, although absolute accuracy and between-subject variability are limited.

The gains on TM and FR are smaller and accompanied by substantial between-subject variability, especially given the limited cohort sizes. Complete subject-level results, improved-subject counts, and paired statistical comparisons are provided in Appendix D.6. Overall, the performance gains are clearest on FBM, while conclusions on TM and FR remain more limited.

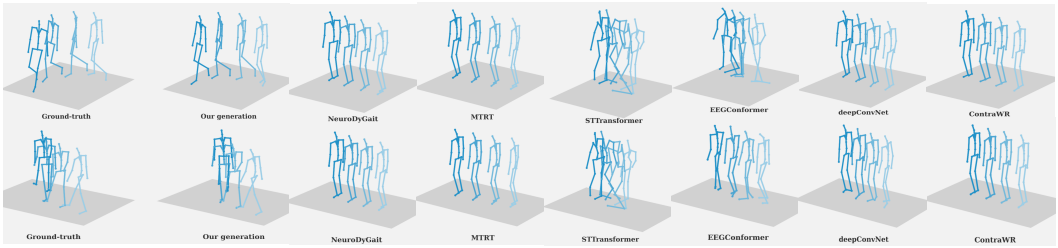

Figure 4: Qualitative comparison on FBM. We visualize decoded full-body motion from NeuroAction and representative baselines against the ground truth. Compared with baselines that often produce over-smoothed or disordered poses, NeuroAction better preserves the locomotion pattern, articulated body structure, and overall motion coherence.

Hard-token decoding and refinement. Stage 2 already provides effective motion reconstruction from predicted hard token IDs alone, achieving PCC/R^2/RMSE of $0.412/-0.117/0.821$ on FBM, $0.459/-0.102/0.463$ on TM, and $0.643/0.094/0.247$ on FR. This shows that the discrete motion-token stage captures useful motion structure before any continuous refinement. Stage 3 then further improves the reconstructed trajectories: all three metrics improve for every FBM and TM subject, while on FR PCC improves for 5/5 subjects and R^2/RMSE for 4/5, with S07 unchanged on the latter two metrics. Together, these results support complementary roles for the two stages, with Stage 2 providing coarse motion structure and Stage 3 refining it in continuous kinematic space.

Distributional quality. On FBM, Stage 2 has the lowest FID (3.169 ± 2.037) and Diversity closest to the real-motion reference. Stage 3 improves paired reconstruction while FID increases to 4.139 ± 4.036. The default Stage-3 objective supervises 30-D end effectors, whereas FID measures full-263-D sequences using the pretrained motion evaluator of Guo et al. Guo et al. (2022a), which is widely used for text-to-motion evaluation. A controlled full-263-D refinement objective does not resolve this trade-off (Appendix D.1.4). Given evaluator-domain mismatch and substantial subject variability, FID is complementary evidence rather than a direct measure of EEG-conditioned trajectory accuracy.

3.3 Ablation Study

We conduct multi-subject component ablations on four subjects per dataset: FBM S02/S07/S09/S10, TM S01/S02/S06/S07, and FR S05/S07/S13/S22. Table 2 reports mean $\pm$ between-subject standard deviation. Removing either alignment objective or both degrades mean performance across all three datasets, supporting the contribution of codebook-guided EEG alignment. Skipping Stage 2 causes the largest overall degradation, supporting the importance of learning the discrete intermediate task before motion-level optimization.

w/o SupCon removes supervised contrastive learning, *w/o PCL* removes prototypical contrastive learning, and *w/o Alignment* removes both. *w/o E2M* removes the token-generation branch and cross-entropy supervision: EEG embeddings are trained using alignment losses alone, tokens are selected by nearest-codebook cosine similarity, and Stage 3 is omitted. Thus, this setting ablates the learned token-generation pathway rather than only the Transformer architecture. *w/o Stage 2* retains the Stage-1 motion prior and E2M architecture, but skips token-prediction and alignment training and directly optimizes the EEG encoder, E2M module, and VQ decoder with motion-space MSE. The effect of Stage 3 is evaluated separately on the complete cohorts in Table 1 and Appendix D.6.

Appendix D.1 evaluates codebook capacity, temporal token granularity, alignment weighting, codebook fine-tuning, and refinement supervision scope. The experiments use explicitly specified subject subsets or the complete cohorts, with comparisons made only within matched experimental settings.

EEG input controls. Table 3 compares standard EEG with three input controls designed to probe different information sources in the decoding pipeline. Zero EEG evaluates how much performance can be attributed to the pretrained VQ-VAE motion prior alone, without informative neural input. Random temporal shuffling removes temporal order while preserving instantaneous multichannel values, testing how much decoding can rely on non-temporal EEG information. Temporal reversal provides an intermediate control that reverses temporal orientation while preserving local adjacency,

Table 2: Component ablations on four subjects per dataset: FBM (S02/S07/S09/S10), TM (S01/S02/S06/S07), and FR (S05/S07/S13/S22). Entries are mean ± between-subject standard deviation. Ablation definitions are given in Sec. 3.3.

	FBM			FR			TM		
Variant	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓
Full model	$\mathbf{0.584}_{\pm 0.231}$	$\mathbf{0.259}_{\pm 0.377}$	$\mathbf{0.746}_{\pm 0.291}$	$\mathbf{0.717}_{\pm 0.087}$	$\mathbf{0.350}_{\pm 0.297}$	$\mathbf{0.216}_{\pm 0.024}$	$\mathbf{0.522}_{\pm 0.189}$	$\mathbf{0.187}_{\pm 0.098}$	$\mathbf{0.404}_{\pm 0.066}$
w/o SupCon	$\underline{0.373}_{\pm 0.349}$	$\underline{0.163}_{\pm 0.289}$	$\underline{0.836}_{\pm 0.296}$	$0.612_{\pm 0.144}$	$0.174_{\pm 0.400}$	$0.242_{\pm 0.028}$	$0.338_{\pm 0.320}$	$\underline{0.062}_{\pm 0.091}$	$\underline{0.434}_{\pm 0.066}$
w/o PCL	$0.186_{\pm 0.329}$	$-0.015_{\pm 0.287}$	$0.940_{\pm 0.297}$	$\underline{0.648}_{\pm 0.165}$	$\underline{0.283}_{\pm 0.286}$	$\underline{0.229}_{\pm 0.023}$	$0.188_{\pm 0.377}$	$0.012_{\pm 0.081}$	$0.446_{\pm 0.065}$
w/o Alignment	$0.259_{\pm 0.301}$	$0.033_{\pm 0.219}$	$0.920_{\pm 0.263}$	$0.596_{\pm 0.152}$	$0.150_{\pm 0.479}$	$0.244_{\pm 0.029}$	$0.164_{\pm 0.328}$	$-0.037_{\pm 0.028}$	$0.458_{\pm 0.041}$
w/o E2M	$0.317_{\pm 0.244}$	$-0.295_{\pm 0.476}$	$1.000_{\pm 0.341}$	$0.598_{\pm 0.151}$	$0.091_{\pm 0.447}$	$0.255_{\pm 0.037}$	$\underline{0.490}_{\pm 0.194}$	$-0.075_{\pm 0.201}$	$0.461_{\pm 0.094}$
w/o Stage 2	$0.018_{\pm 0.022}$	$-0.396_{\pm 0.274}$	$1.078_{\pm 0.236}$	$0.170_{\pm 0.221}$	$-0.196_{\pm 0.155}$	$0.307_{\pm 0.053}$	$0.184_{\pm 0.137}$	$-0.165_{\pm 0.050}$	$0.492_{\pm 0.025}$

Table 3: EEG input controls, averaged over FBM (S03/S10), TM (S07/S08), and FR (S05/S13). Reversal changes temporal orientation; shuffling destroys temporal order within each window. Stage 1 is reused and Stages 2–3 are retrained with the same settings.

	FBM			FR			TM		
Input	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓
Standard EEG	**0.802**	**0.612**	**0.493**	**0.775**	**0.572**	**0.207**	**0.718**	**0.351**	**0.349**
Reversed EEG	0.672	0.402	0.606	0.736	0.446	0.237	0.648	0.102	0.411
Shuffled EEG	0.096	-0.311	0.981	0.534	0.093	0.306	0.0005	-0.702	0.588
Zero EEG	0.013	-0.615	1.044	0.000	-0.247	0.352	0.000	-0.653	0.591

Fourier magnitude, and cross-channel relations. For all controls, we reuse the Stage-1 motion prior and retrain Stages 2–3 under the same protocol.

For the order-destruction control, each 2-s window receives an independent random temporal permutation shared across channels. This preserves instantaneous multichannel vectors and marginal values while destroying temporal order and local continuity, with the EEG–motion window pairing unchanged. Shuffling reduces PCC to 0.096 on FBM and 0.0005 on TM. On FR, PCC remains 0.534, but R^2 drops from 0.572 to 0.093 and RMSE increases from 0.207 to 0.306. Zero EEG performs poorly on all datasets. These controls indicate that ordered temporal structure contributes substantially to decoding on FBM and TM, while FR retains additional information in instantaneous multichannel structure despite a marked loss in reconstruction quality.

3.4 Representation Alignment Analysis

We further analyze whether the learned EEG representations are structurally aligned with motion primitives. We perform token-level RSA Kriegeskorte et al. (2008) between EEG embeddings and motion representations by comparing their representational dissimilarity matrices (RDMs) at each token position. We report representative results on FBM `S10` and FR `S05`, with more details in Appendix D.3.

As shown in Fig. 5, the real EEG–motion RSA curves remain above permutation-based chance baselines , supporting structural correspondence between the learned EEG representations and the motion-primitive space. The latent and quantized RSA curves compare EEG embeddings with pre-quantized motion latents and quantized codebook embeddings, respectively; chance curves are obtained through permutation controls, and the ceiling estimates the attainable similarity given representational variability.

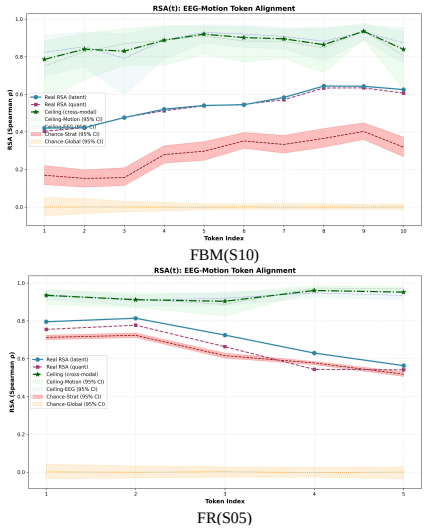

Figure 5: Token-level RSA on FBM `S10` and FR `sub-05`.

We also visualize token-level EEG embeddings with t-SNE on FBM `S10` and FR `S05`. We select embeddings associated with the four most frequently used motion codes and color each point by its paired codebook label. As shown in Fig. 6, the embeddings exhibit code-dependent organization, providing qualitative evidence of correspondence with recurring motion primitives.Because the codebook is learned solely from motion, these results reflect task-specific EEG–motion correspondence rather than a direct physiological categorization of motor states.

4 Discussion

NeuroAction uses discrete motion primitives as an intermediate representation between noisy EEG and continuous kinematics. Its main contribution is aligning continuous EEG embeddings to motion-derived prototypes and integrating this alignment with token prediction and motion-level refinement. A common formulation supports FBM, TM, and FR, using a dataset-specific motion tokenizer and subject-specific EEG-to-motion models.

Multi-subject ablations support the contributions of alignment and staged learning, while EEG input controls demonstrate sensitivity to ordered temporal dynamics. Frequency-band analysis (Appendix D.2) and representation analyses provide complementary diagnostics. Because the motion codebook is learned solely from kinematics, the resulting tokens should be interpreted as reusable motion primitives with task-specific EEG correspondence rather than independently established physiological motor states.

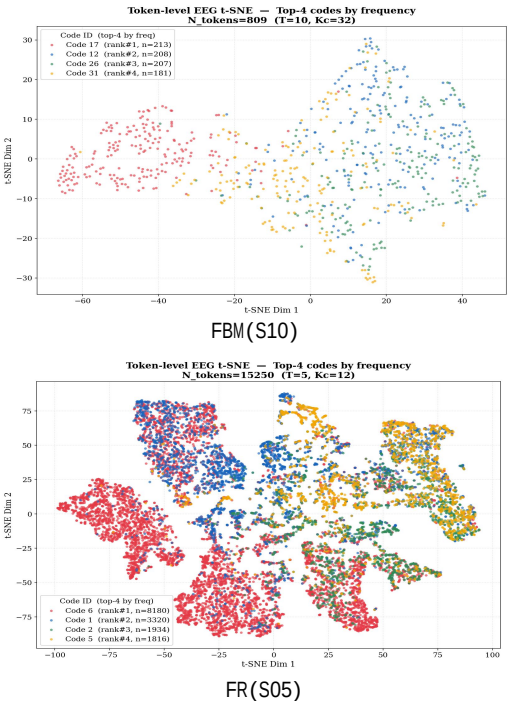

Figure 6: Token-level t-SNE visualization of EEG embeddings on FBM `S10` and FR `S05`.

Limitations. The study is limited by small subject cohorts and offline evaluation. Although TM and FR include cross-day testing, longer-term EEG nonstationarity remains unresolved, while FBM lacks verified day-level metadata. The discrete VQ motion prior may introduce quantization error and suppress fine-grained kinematic variation. Reported inference timings characterize model-side computation only and do not establish end-to-end closed-loop performance. Larger population studies, longer-term evaluation, and online validation remain important future directions.

5 Conclusion

We presented NeuroAction, a framework that aligns EEG representations to discrete motion primitives and integrates motion-token prediction with geometric refinement for continuous motion decoding. Across three dataset-specific instantiations, the method improves mean reconstruction metrics over the compared baselines, with the strongest statistical evidence on FBM end-effector reconstruction. Multi-subject ablations, temporal controls, and representation analyses support structured EEG–motion alignment. Cross-day testing and offline profiling characterize current capabilities, while broader population studies and online closed-loop evaluation remain necessary.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and readability of the manuscript. They were not used to generate scientific ideas, experimental designs, code, data, analyses, results, or references. All technical content and scientific claims were produced and verified by the authors.

ETHICS STATEMENT

This work uses only previously published public EEG–motion datasets and involves no new human-subject data collection. We follow the data-use conditions of the original datasets and use the recordings solely for offline research on EEG-to-motion decoding. Because EEG is a sensitive biosignal, practical deployment of such systems should require informed consent, appropriate privacy and data-security safeguards, and careful safety validation, particularly in closed-loop or assistive-control settings.

REPRODUCIBILITY STATEMENT

We provide detailed experimental information in the appendix, including dataset acquisition and session protocols, preprocessing and normalization, motion representations, session-wise train/test splits, model architectures and tensor shapes, codebook and token configurations, optimization hyperparameters, baseline adaptation procedures, evaluation metrics, subject-level statistics, ablation settings, and inference-cost measurements. All reported experiments use explicitly specified subject subsets or complete cohorts, and held-out test sessions are excluded from motion-prior training. The code will be made publicly available upon acceptance.

REFERENCES

Trent J. Bradberry, Rodolphe J. Gentili, and José L. Contreras-Vidal. Reconstructing three-dimensional hand movements from noninvasive electroencephalographic signals. *The Journal of Neuroscience*, 30(9):3432–3437, 2010. doi: 10.1523/JNEUROSCI.6107-09.2010.

Justin A. Brantley, Trieu Phat Luu, Sho Nakagome, Fangshi Zhu, and Jose L. Contreras-Vidal. Full body mobile brain-body imaging data during unconstrained locomotion on stairs, ramps, and level ground. *Scientific Data*, 5:180133, 2018. doi: 10.1038/sdata.2018.133. URL https://www.nature.com/articles/sdata2018133.pdf.

Arnaud Delorme and Scott Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134 (1):9–21, 2004. doi: 10.1016/j.jneumeth.2003.10.009.

Yiqun Duan, Qiang Zhang, Jinzhao Zhou, Jingkai Sun, Xiaowei Jiang, Jiahang Cao, Jiaxu Wang, Yiqian Yang, Wen Zhao, Gang Han, Yijie Guo, and Chin-Teng Lin. E2H: A two-stage non-invasive neural signal driven humanoid robotic whole-body control framework. *arXiv preprint arXiv:2410.02141*, 2024. URL https://arxiv.org/abs/2410.02141.

Xi Fu, Weibang Jiang, Rui Liu, Gernot R. Müller-Putz, and Cuntai Guan. Zero-shot EEG-to-gait decoding via phase-aware representation learning. *arXiv preprint arXiv:2506.22488v1*, 2025. URL https://arxiv.org/abs/2506.22488v1.

Xi Fu, Rui Liu, Aung Aung Phyo Wai, Hannah Pulferer, Neethu Robinson, Gernot R. Müller-Putz, and Cuntai Guan. EEG2GAIT: A hierarchical graph convolutional network for EEG-based gait decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 34:441–454, 2026. doi: 10.1109/TNSRE.2025.3647101. URL https://arxiv.org/abs/2504.03757.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, 2022a. doi: 10.1109/CVPR52688.2022.00509. URL https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In *Computer Vision – ECCV 2022*, volume 13695 of *Lecture Notes in Computer Science*, pp. 580–597. Springer, 2022b. doi: 10.1007/978-3-031-19833-5_34. URL https://arxiv.org/abs/2207.01696.

Joseph T. Gwin, Klaus Gramann, Scott Makeig, and Daniel P. Ferris. Electrocortical activity is coupled to gait cycle phase during treadmill walking. *NeuroImage*, 54(2):1289–1296, 2011. doi: 10.1016/j.neuroimage.2010.08.066.

Yongtian He, Trieu Phat Luu, Kevin Nathan, Sho Nakagome, and Jose L. Contreras-Vidal. A mobile brain-body imaging dataset recorded during treadmill walking with a brain-computer interface. *Scientific Data*, 5:180074, 2018. doi: 10.1038/sdata.2018.74.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. NeuroLM: A universal multi-task foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.00101.

Brianna M. Karpowicz, Yahia H. Ali, Lahiru N. Wimalasena, Andrew R. Sedler, Mohammad Reza Keshtkaran, Kevin Bodkin, Xuan Ma, Daniel B. Rubin, Ziv M. Williams, Sydney S. Cash, Leigh R. Hochberg, Lee E. Miller, and Chethan Pandarinath. Stabilizing brain-computer interfaces through alignment of latent dynamics. *Nature Communications*, 16(1):4662, 2025. doi: 10.1038/s41467-025-59652-y.

Kristel Knaepen, Andreas Mierau, Helio Fernandez Tellez, Dirk Lefeber, and Romain Meeusen. Temporal and spatial organization of gait-related electrocortical potentials. *Neuroscience Letters*, 599:75–80, 2015. doi: 10.1016/j.neulet.2015.05.036. URL https://doi.org/10.1016/j.neulet.2015.05.036.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

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.

Trieu Phat Luu, Yongtian He, Sho Nakagame, Jeffrey Gorges, Kevin Nathan, and Jose L. Contreras-Vidal. Unscented kalman filter for neural decoding of human treadmill walking from non-invasive electroencephalography. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1548–1551. IEEE, 2016. doi: 10.1109/EMBC.2016.7591006.

Ferdinando A. Mussa-Ivaldi. Modular features of motor control and learning. *Current Opinion in Neurobiology*, 9(6):713–717, 1999. doi: 10.1016/S0959-4388(99)00029-X.

João Ruivo Paulo, Teresa Sousa, João Perdiz, Lara Pereira, Mariette Vasen, Susana Mouga, Gabriel Pires, and Miguel Castelo-Branco. A multimodal dataset addressing motor function in autism. *Scientific Data*, 12:959, 2025. doi: 10.1038/s41597-025-05313-0. URL https://www.nature.com/articles/s41597-025-05313-0.pdf.

Neethu Robinson and A. P. Vinod. Noninvasive brain–computer interface: Decoding arm movement kinematics and motor control. *IEEE Systems, Man, and Cybernetics Magazine*, 2(4):4–16, October 2016. doi: 10.1109/MSMC.2016.2576638. URL https://www.ieeesmc.org/wp-content/uploads/2014/03/SMC-Magazine-BCI-Article-October-2016.pdf.

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

Paul Steven Scotti, Mihir Tripathy, Cesar Torrico, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq Mathew Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44038–44059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/scotti24a.html.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Bhoj Raj Thapa, John Boggess, and Jihye Bae. A large electroencephalogram database of freewill reaching and grasping tasks for brain machine interfaces. *Scientific Data*, 12:1760, 2025. doi: 10.1038/s41597-025-06039-9.

Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002. doi: 10.1038/nn963.

Stefano Tortora, Stefano Ghidoni, Carmelo Chisari, Silvestro Micera, and Fiorenzo Artoni. Deep learning-based BCI for gait decoding from EEG with LSTM recurrent neural network. *Journal of Neural Engineering*, 17(4):046011, July 2020. doi: 10.1088/1741-2552/ab9842. URL https://doi.org/10.1088/1741-2552/ab9842.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00937.

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)*, volume 30, pp. 5998–6008, 2017. URL https://arxiv.org/abs/1706.03762.

Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.393. URL https://aclanthology.org/2024.acl-long.393/.

Pengpai Wang, Zhongnian Li, Peiliang Gong, Yueying Zhou, Fang Chen, and Daoqiang Zhang. MTRT: Motion trajectory reconstruction transformer for EEG-based BCI decoding. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2349–2358, 2023. doi: 10.1109/TNSRE.2023.3275172.

Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. VQ-VLA: Improving vision-language-action models via scaling vector-quantized action tokenizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11089–11099, 2025. doi: 10.1109/ICCV51701.2025.01032. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_VQ-VLA_Improving_Vision-Language-Action_Models_via_Scaling_Vector-Quantized_Action_Tokenizers_ICCV_2025_paper.html.

Daniel M. Wolpert, R. Chris Miall, and Mitsuo Kawato. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9):338–347, 1998. doi: 10.1016/S1364-6613(98)01221-2.

Chaoqi Yang, Cao Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised electroencephalogram representation learning for automatic sleep staging: model development and evaluation study. *JMIR AI*, 2(1):e46769, 2023. doi: 10.2196/46769. URL https://ai.jmir.org/2023/1/e46769.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14730–14740, 2023. doi: 10.1109/CVPR52729.2023.01415. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_Generating_Human_Motion_From_Textual_Descriptions_With_Discrete_Representations_CVPR_2023_paper.html.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1551c01d7a3d0bf21e2518331e9f7074-Abstract-Conference.html.

APPENDIX

A RELATED WORK

A.1 EEG-BASED MOTION DECODING

Decoding movement-related information from non-invasive EEG has long been studied in BCI research. Early studies showed that low-dimensional kinematic variables, such as movement direction, hand velocity, or joint trajectories, can be decoded from EEG in controlled paradigms Bradberry et al. (2010); Robinson & Vinod (2016). Deep learning further enabled end-to-end spatiotemporal modeling of minimally processed EEG using CNN, RNN, and compact convolutional architectures Schirrmeister et al. (2017); Lawhern et al. (2018), while recent Transformer-based models improve long-range temporal modeling for EEG decoding Song et al. (2023).

EEG-based motor decoding has been studied across a range of movement paradigms. Upper-limb decoding commonly focuses on hand movement, reaching, or low-dimensional arm trajectories Bradberry et al. (2010); Robinson & Vinod (2016), with sequence models such as MTRT reconstructing continuous arm motion using Transformer-based temporal modeling Wang et al. (2023). Lower-limb and gait decoding methods study cyclic locomotion dynamics, motion artifacts, and nonstationarity, using CNN/LSTM architectures Tortora et al. (2020), hierarchical or graph-based designs Fu et al. (2026), and phase-aware representation learning for cross-session or cross-subject robustness Fu et al. (2025). Recent work has also explored structured interfaces for whole-body or humanoid-oriented control, such as E2H, which uses an EEG-to-keyword and keyword-to-motion pipeline Duan et al. (2024). Together, these studies provide important foundations for modeling neural activity associated with upper-limb, lower-limb, gait, and whole-body movement.

For the gait comparison, NeuroDyGait serves as the later task-specific reference. Our comparison covers direct regression and learned continuous intermediate representations; it does not establish superiority over every possible adapted generative model.

A.2 MOTION AND ACTION TOKENIZATION

Discrete representations have become an effective tool for modeling continuous motion and action sequences. In text-to-motion generation, vector-quantized autoencoders tokenize high-dimensional human motion into compact codebooks van den Oord et al. (2017), enabling Transformer-based autoregressive generation conditioned on textual or semantic inputs Guo et al. (2022b); Zhang et al. (2023). These methods show that learned motion codebooks can absorb low-level kinematic details and provide a structured latent space for long-horizon motion synthesis.

Motion and action tokenization has also been explored in vision-language-action and robotic control models. VQ-style action quantization, as in VQ-VLA Wang et al. (2025), discretizes low-dimensional control trajectories into reusable action tokens. This suggests that quantization can serve as a general mechanism for representing both high-dimensional human motion and lower-dimensional robotic action sequences.

A.3 REPRESENTATION ALIGNMENT FOR NEURAL DECODING

Robust neural decoding often benefits from learning structured latent representations rather than relying solely on direct output supervision. Recent EEG foundation models discretize EEG into neural tokens and train sequence models with language-model-like objectives, enabling unified inference across multiple EEG tasks Jiang et al. (2025). Related studies show that aligning latent neural dynamics across sessions can improve decoding stability and robustness Karpowicz et al. (2025). In brain-to-image reconstruction, approaches such as MindEye demonstrate that contrastive alignment between brain activity and semantic embedding spaces can learn informative neural representations for retrieval and reconstruction Scotti et al. (2024). Similar alignment strategies have also been explored in EEG-to-text decoding to bridge neural activity and language representations Wang et al. (2024).

NeuroAction learns discrete primitives from motion and aligns continuous EEG representations to their codebook vectors. These primitives provide both representation-level supervision and

intermediate prediction targets for continuous motion reconstruction. EEG tokenizers used for pretraining or classification therefore address a different intermediate representation.Likewise, generic motion generation and EEG foundation encoders require an EEG-conditioning or synchronized kinematic decoding interface before they can be evaluated on this task. Testing such adaptations is useful future work; their absence does not imply that they are ineffective.

B Experimental Settings

B.1 Dataset Acquisition and Session Protocols

FBM records EEG and full-body kinematics during continuous overground multi-terrain locomotion, including level walking, ramps, and stairs Brantley et al. (2018). We use subjects `s02`–`s10`, excluding `s01` due to synchronization issues. For each subject, the last available session is held out for testing. The dataset does not provide explicit visit/day metadata sufficient to establish a cross-day split, so we interpret this protocol as session-disjoint cross-trial evaluation rather than verified cross-day generalization.

TM records 60-channel EEG and bilateral lower-limb joint angles during treadmill walking and closed-loop BCI control He et al. (2018). We evaluate six subjects, excluding `S03` and `S04` due to data quality issues. Sessions 01 and 02 were recorded during the first visit, while Session 03 was recorded during a second visit within seven days. Training on Session 01/02 and testing on Session 03 therefore provides a cross-day evaluation.

FR records 31-channel EEG and three-axis wrist acceleration during voluntary reaching-and-grasping Thapa et al. (2025). We evaluate `s05`, `s07`, `s13`, `s15`, and `s22`, the five subjects with complete three-session recordings. Sessions 1–3 were recorded on different days; training on the earlier sessions and testing on Session 3 therefore constitutes cross-day evaluation.

Accordingly, the main TM and FR protocols evaluate session-disjoint cross-day generalization, whereas FBM provides session-disjoint evaluation without verified day separation. None of these protocols establishes stability over multi-week or multi-month intervals.

B.2 Model and Training Details

Signal preprocessing. For all datasets, kinematic signals are low-pass filtered at 5 Hz, downsampled to 20 Hz, and temporally aligned with EEG. FBM motion is converted into the 263-D HumanML3D-style representation described in Sec. B.3, while TM and FR motion targets are min-max normalized to $[0, 1]$. EEG signals are band-pass filtered at 0.1–48 Hz, re-referenced using common average reference (CAR) Delorme & Makeig (2004), and downsampled to 100 Hz.Channel-wise z-score normalization is then applied as the first operation of the EEG encoder rather than as an offline preprocessing step.

Windowing and token length. All datasets are processed into temporally aligned EEG-motion windows of length 2 s with a stride of 0.5 s. Kinematic signals are downsampled to 20 Hz, so each motion window contains 40 frames. The motion tokenizer temporally downsamples each window into a sequence of motion tokens. The default token counts are $N_e = N_m = 10$ for FBM/TM and $N_e = N_m = 5$ for FR, corresponding to nominal durations of 0.2 s and 0.4 s per token. The model predicts one motion-token sequence per EEG window.

Architecture hyperparameters. For FBM, we use a VQ codebook size of 32. The EEG encoder hidden dimension and E2M hidden dimension are both set to 512. The EEG encoder contains 6 Transformer layers, and the E2M Transformer contains 16 layers with 8 attention heads.

For TM and FR, which contain lower-dimensional kinematic targets, we use smaller codebooks and model widths. Specifically, TM uses a VQ codebook size of 16, while FR uses a VQ codebook size of 12. For both datasets, the EEG encoder and E2M hidden dimensions are set to 256. The EEG encoder contains 6 Transformer layers, and the E2M Transformer contains 16 layers with 8 attention heads.

Stage 1: VQ motion prior training. We train a separate motion tokenizer for each dataset using only the kinematic training data. Test sessions used in EEG-to-motion evaluation are excluded from VQ training. The VQ tokenizer is trained with AdamW using learning rate 2×10^{-4}, batch size 1024, weight decay 1×10^{-3}, 100k iterations, 1k warmup iterations, and a cosine learning-rate schedule. The reconstruction loss weight is set to 1.0, and the commitment loss weight is set to 0.02. After training, the VQ encoder, codebook, and decoder define the dataset-specific discrete motion prior.

Stage 2: EEG-to-motion token learning. With the motion tokenizer frozen, we train the EEG encoder, neuro-motion alignment module, and E2M Transformer to predict motion-token sequences from EEG. The training objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{tok}} + \lambda_{\text{pcl}}\mathcal{L}_{\text{pcl}} + \lambda_{\text{supcon}}\mathcal{L}_{\text{supcon}}. \tag{11}$$

We set $\lambda_{\text{pcl}} = 1.0$ and $\lambda_{\text{supcon}} = 1.0$, and use temperature $\tau = 0.07$ for both contrastive objectives. Stage 2 is trained with AdamW using learning rate 1×10^{-3}, batch size 512, weight decay 1×10^{-3}, 12k iterations, 1k warmup iterations, and a cosine learning-rate schedule.

Stage 3: geometric refinement. Starting from the Stage 2 checkpoint, we perform geometric refinement by optimizing the decoded continuous motion with an MSE loss. During this stage, the EEG encoder, E2M Transformer, and VQ decoder are updated, while the VQ encoder and codebook remain fixed. We use AdamW with learning rate 1×10^{-5}, batch size 64, weight decay 1×10^{-3}, 2k iterations, 1k warmup iterations, and a cosine learning-rate schedule.

For FBM, Stage-3 MSE supervises only the 30-D end-effector subset in Appendix C; TM/FR supervise all 6/3 target dimensions. The FBM 263-D supervision variant in Appendix D.1.4 changes only this supervision scope.

Training cost. All experiments are trained on a single NVIDIA L4 GPU. Stage 1 is run once per dataset and takes approximately 6 h for FBM and 2 h for TM/FR. Stages 2–3 are subject-specific. With subject-dependent convergence and early stopping, Stage 2 runs for 2,000–12,000 iterations (10–60 min), and Stage 3 for 500–2,000 iterations (3–12 min), within the nominal budgets above. Subject-specific training therefore takes 13–72 min. Original and additional experiments share the same preprocessing, split, normalization, and training settings except for explicitly controlled factors.

B.3 Motion representation (263-D)

Following prior tokenized motion generation pipelines Zhang et al. (2023), each FBM motion frame uses the standard 263-D HumanML3D-style representation:

$$\mathbf{x} = \left(\dot{r}_a, \dot{r}_x, \dot{r}_z, r_y, \mathbf{p}_{\neg r}, \mathbf{R}_{\neg r}, \mathbf{v}, \mathbf{c}_f\right) \in \mathbb{R}^{263}. \tag{12}$$

For $j = 22$ joints, the root block contains yaw angular velocity (1), planar root velocity expressed in the root frame (2), and root height (1). Root-relative positions $\mathbf{p}_{\neg r} \in \mathbb{R}^{3(j-1)}$ and continuous 6D rotations $\mathbf{R}_{\neg r} \in \mathbb{R}^{6(j-1)}$ exclude the root joint. Local joint velocities $\mathbf{v} \in \mathbb{R}^{3j}$ include all joints, and $\mathbf{c}_f \in \mathbb{R}^4$ contains foot-contact indicators. The concatenation order is root features, non-root positions, non-root rotations, all-joint velocities, and contacts. Consequently,

$$1 + 2 + 1 + 3(j-1) + 6(j-1) + 3j + 4 = 4 + 63 + 126 + 66 + 4 = 263.$$

B.4 Baseline Adaptation and Training

For a controlled comparison, all baselines use the same EEG preprocessing, session-wise train/test splits, dataset-specific target representations, normalization, and evaluation metrics as NeuroAction. Each baseline retains its original backbone, while only the dataset-specific input interface and regression output head are adapted where necessary.

On FBM, all methods predict the same 263-D HumanML3D-style representation over 40 motion frames. Every regression baseline uses a shared MLP readout with widths feature $\rightarrow$ 512 $\rightarrow$ 2048 $\rightarrow$ 10,520, followed by reshaping to 40×263. We use this readout for all FBM regression baselines because a single linear projection was unstable for the high-dimensional output. On TM and FR, the corresponding outputs are the normalized 6-D lower-limb joint-angle sequence and 3-D wrist-acceleration sequence, respectively.

For each baseline, dropout is selected from $\{0.0, 0.1, \ldots, 0.9\}$. Training uses learning rate 10^{-4}, cosine scheduling, a 200-epoch budget, and early-stopping patience 10. These controls standardize the prediction task and baseline optimization protocol; backbone capacities and the staged NeuroAction training procedure are not forced to be identical across methods.

Regarding EEG2GAIT, NeuroDyGait directly compares against it, reports stronger gait-decoding performance, and shares the same first author; we therefore use NeuroDyGait as the later and stronger gait-specific reference. Moreover, the EEG2GAIT release lacks components required for full reproduction without undocumented assumptions.

C Evaluation Metrics

We evaluate EEG-to-motion decoding with RMSE, PCC, and R^2 on all three datasets. For FBM, these reconstruction metrics are computed on the 30-D end-effector position subset of the 263-D HumanML3D-style motion vector, corresponding to the XYZ coordinates of 10 limb joints: LKnee, RKnee, LHeel, RHeel, LFoot, RFoot, LElbow, RElbow, LWrist, and RWrist. The 0-based indices are

$$\mathcal{D}_{\mathrm{FBM}} = [13{:}19,\ 22{:}28,\ 31{:}37,\ 55{:}67],$$

where each range is a zero-based, half-open slice, giving 30 dimensions in total. For TM, the metrics are computed on the 6-D bilateral lower-limb joint-angle sequence from hip, knee, and ankle goniometers. For FR, the metrics are computed on the 3-D wrist-acceleration sequence. For FBM, we additionally report FID and Diversity in a pretrained motion-embedding space. We compute each metric per subject and then average across subjects; reported standard deviations describe between-subject variability. Paired comparisons use two-sided paired t-tests across subjects, with uncorrected p-values reported in Appendix D.6.

C.1 Root Mean Squared Error (RMSE)

Given predicted motion $\hat{\mathbf{Y}} \in \mathbb{R}^{B\times T\times D}$ and ground-truth motion $\mathbf{Y} \in \mathbb{R}^{B\times T\times D}$, where $D = 30$ for FBM reconstruction metrics, $D = 6$ for TM, and $D = 3$ for FR, we first flatten the batch and temporal axes into $n = 1, \ldots, BT$. RMSE is computed independently for each evaluated dimension and then averaged:

$$\mathrm{RMSE}^{(d)} = \sqrt{\frac{1}{BT}\sum_{n=1}^{BT}\left(\hat{y}_n^{(d)} - y_n^{(d)}\right)^2}, \qquad \mathrm{RMSE} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{RMSE}^{(d)}. \tag{13}$$

Lower RMSE indicates more accurate continuous motion reconstruction.

C.2 Pearson Correlation Coefficient (PCC)

PCC measures temporal correlation between predicted and ground-truth motion signals. We compute PCC separately for each evaluated motion dimension and report the average across dimensions:

$$\mathrm{PCC}^{(d)} = \frac{\sum_n (\hat{y}_n^{(d)} - \bar{\hat{y}}^{(d)})(y_n^{(d)} - \bar{y}^{(d)})}{\sqrt{\sum_n (\hat{y}_n^{(d)} - \bar{\hat{y}}^{(d)})^2}\sqrt{\sum_n (y_n^{(d)} - \bar{y}^{(d)})^2}}, \qquad \mathrm{PCC} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{PCC}^{(d)}. \tag{14}$$

Here, n indexes all sample-frame pairs. Higher PCC impies better temporal agreement with ground truth.

C.3 Coefficient of Determination (R^2)

R^2 is also computed independently for each evaluated motion dimension and then averaged:

$$R^{2,(d)} = 1 - \frac{\sum_n \left(y_n^{(d)} - \hat{y}_n^{(d)}\right)^2}{\sum_n \left(y_n^{(d)} - \bar{y}^{(d)}\right)^2}, \qquad R^2 = \frac{1}{D}\sum_{d=1}^{D} R^{2,(d)}. \tag{15}$$

Higher R^2 indicates better explained variance, and negative values indicate performance worse than predicting the mean for that dimension.

Table 4: Codebook capacity on three subjects per dataset: FBM (S02/S07/S10), TM (S02/S06/S07), and FR (S05/S07/S22). Entries are mean ± between-subject standard deviation. The default sizes are $K = 32/16/12$ for FBM/TM/FR.

Dataset	K	PCC↑	R^2 ↑	RMSE↓
FBM	12	$0.449_{\pm 0.382}$	$0.120_{\pm 0.445}$	$0.785_{\pm 0.354}$
FBM	16	$0.454_{\pm 0.403}$	$0.135_{\pm 0.470}$	$0.771_{\pm 0.354}$
FBM	32	$\mathbf{0.568}_{\pm 0.280}$	$\mathbf{0.232}_{\pm 0.457}$	$\mathbf{0.734}_{\pm 0.355}$
FBM	64	$0.488_{\pm 0.351}$	$0.168_{\pm 0.491}$	$0.755_{\pm 0.361}$
FBM	128	$0.476_{\pm 0.321}$	$0.123_{\pm 0.437}$	$0.790_{\pm 0.339}$
TM	12	$0.474_{\pm 0.276}$	$0.118_{\pm 0.116}$	$0.411_{\pm 0.083}$
TM	16	$\mathbf{0.545}_{\pm 0.225}$	$\mathbf{0.194}_{\pm 0.119}$	$\mathbf{0.395}_{\pm 0.078}$
TM	32	$0.470_{\pm 0.306}$	$0.164_{\pm 0.090}$	$0.405_{\pm 0.080}$
TM	64	$0.414_{\pm 0.372}$	$0.115_{\pm 0.146}$	$0.411_{\pm 0.085}$
TM	128	$0.407_{\pm 0.361}$	$0.118_{\pm 0.091}$	$0.421_{\pm 0.084}$
FR	12	$\mathbf{0.698}_{\pm 0.096}$	$\mathbf{0.272}_{\pm 0.311}$	$\mathbf{0.211}_{\pm 0.027}$
FR	16	$0.605_{\pm 0.173}$	$0.060_{\pm 0.545}$	$0.232_{\pm 0.044}$
FR	32	$0.601_{\pm 0.151}$	$-0.009_{\pm 0.685}$	$0.239_{\pm 0.050}$
FR	64	$0.576_{\pm 0.173}$	$-0.016_{\pm 0.703}$	$0.240_{\pm 0.055}$
FR	128	$0.561_{\pm 0.170}$	$-0.047_{\pm 0.717}$	$0.246_{\pm 0.054}$

C.4 Fréchet Inception Distance (FID) for FBM

For FBM, we evaluate distribution-level full-body motion quality using FID in a fixed pretrained motion-embedding space. Specifically, we use the motion encoder from the evaluator released by Guo et al. Guo et al. (2022a), which is commonly used in text-to-motion evaluation. Unlike RMSE, PCC, and R^2, FID is computed on the full 263-D HumanML3D-style motion sequence after evaluator preprocessing. Given generated motion embeddings $\{\hat{\mathbf{z}}_n\}_{n=1}^N$ and real motion embeddings $\{\mathbf{z}_n\}_{n=1}^N$, we compute empirical means and covariances $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$. FID is defined as:

$$\mathrm{FID} = \left\|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_r - 2(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_r)^{1/2}\right). \tag{16}$$

Lower FID indicates closer distributions in this evaluator space. Because the evaluator was trained on a different skeleton/data domain from FBM, it is a complementary measure and should not be interpreted as direct evidence of paired trajectory accuracy; small differences with large subject variability warrant caution.

C.5 Diversity for FBM

For FBM, we report Diversity to measure the variability of generated full-body motions in the same Guo et al. motion-encoder embedding space Guo et al. (2022a). Diversity is computed on full 263-D HumanML3D-style motion sequences after evaluator preprocessing. Given generated embeddings $\{\hat{\mathbf{z}}_n\}_{n=1}^N$, Diversity is computed as:

$$\mathrm{Diversity} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \left\|\hat{\mathbf{z}}_i - \hat{\mathbf{z}}_j\right\|_2. \tag{17}$$

We also report the Diversity of ground-truth FBM motions as a reference. A generated set with Diversity closer to the real-motion reference indicates better coverage of the natural motion distribution.

D Additional Experimental Results

D.1 Extended Ablation Studies

Effect of VQ codebook size. Table 4 expands capacity analysis to three subjects per dataset: FBM S02/S07/S10, TM S02/S06/S07, and FR S05/S07/S22. Under the same end-to-end evaluation, the default $K = 32/16/12$ gives the best mean PCC, R^2, and RMSE for FBM/TM/FR. Increasing K to

Table 5: Model-capacity configurations used in the size ablation. L_{EEG} and L_{E2M} denote the numbers of EEG-encoder and total E2M Transformer layers, respectively.

Dataset	Size	d_{EEG}	d_{E2M}	L_{EEG}	L_{E2M}	H_{EEG}	H_{E2M}
FBM	Large	512	512	6	16	8	8
	Middle	512	256	5	12	8	8
	Small	512	256	4	8	8	4
	Extra Small	512	256	3	6	8	4
TM	Large	256	256	6	16	8	8
	Middle	256	256	5	12	8	8
	Small	256	128	4	8	8	4
	Extra Small	256	128	3	6	8	4
FR	Large	256	256	6	16	8	8
	Middle	256	256	5	12	8	8
	Small	256	128	4	8	8	4
	Extra Small	256	128	3	6	8	4

Table 6: Illustrative single-subject model-size ablation across three datasets. Results are reported on FBM (`S10`), TM (`S07`), and FR (`S05`).

Model Size	**FBM** (`S10`)			**TM** (`S07`)			**FR** (`S05`)		
	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓	PCC↑	R^2 ↑	RMSE↓
Large	**0.878**	**0.749**	**0.451**	**0.803**	**0.323**	**0.305**	**0.778**	**0.562**	**0.182**
Middle	<u>0.864</u>	0.722	<u>0.467</u>	0.787	<u>0.288</u>	0.312	0.769	<u>0.554</u>	<u>0.184</u>
Small	0.861	<u>0.726</u>	0.469	<u>0.797</u>	0.282	<u>0.309</u>	<u>0.770</u>	0.550	0.186
Extra Small	0.856	0.714	0.469	0.791	0.283	0.311	0.766	0.545	0.187

128 lowers PCC to 0.476/0.407/0.561. These results are consistent with a trade-off between motion expressiveness and EEG decodability. Smaller vocabularies may merge distinct motions, while finer vocabularies may create harder EEG discrimination tasks.

Effect of model size. Tables 5 and 6 compare four model-capacity settings. Starting from the default Large model, we progressively reduce the EEG-encoder depth and E2M Transformer depth, and for the smaller variants also reduce the E2M hidden width and number of attention heads where applicable. The Large model achieves the best overall performance across FBM, TM, and FR. Smaller models generally remain competitive, although the ordering is not monotonic across all intermediate sizes. The performance reduction is most visible on FBM, while TM and FR exhibit smaller changes. Because these experiments use one representative subject per dataset, they characterize sensitivity to model capacity rather than a population-level scaling law.

D.1.1 Temporal token granularity

On the same three-subject subsets as the codebook study, we vary $N_m \in \{20, 10, 5\}$ and retrain all stages (Table 7). For 40-frame, 2-s windows, these settings yield nominal durations of 0.1/0.2/0.4 s per token. The best mean results occur at 10 tokens for FBM/TM and 5 for FR, matching the defaults. Coarser tokenization can merge temporal phases; finer tokenization creates longer sequences and subtler adjacent distinctions. This sensitivity concerns the joint EEG-to-motion system.

D.1.2 Alignment-weight sensitivity

We jointly vary $\lambda_{\mathrm{pcl}} = \lambda_{\mathrm{supcon}}$ over $\{0.2, 1, 5\}$, with all remaining settings fixed and the same three-subject subsets. Table 8 favors weight 1 on all dataset means. Together with the individual loss removals in Table 2, this tests both component necessity and sensitivity to a common loss scale.

Table 7: Temporal token granularity on the same three-subject subsets as Table 4. All stages are retrained. Each 2-s window contains 40 motion frames; $N_m = 20/10/5$ corresponds to $2\times/4\times/8\times$ compression. Entries are mean $\pm$ between-subject standard deviation.

Dataset	N_m	PCC↑	R^2 ↑	RMSE↓
FBM	20	$0.512_{\pm 0.310}$	$0.195_{\pm 0.461}$	$0.747_{\pm 0.339}$
FBM	10	$\mathbf{0.568}_{\pm 0.280}$	$\mathbf{0.232}_{\pm 0.457}$	$\mathbf{0.734}_{\pm 0.355}$
FBM	5	$0.482_{\pm 0.373}$	$0.205_{\pm 0.463}$	$0.756_{\pm 0.371}$
TM	20	$0.503_{\pm 0.225}$	$0.155_{\pm 0.089}$	$0.405_{\pm 0.069}$
TM	10	$\mathbf{0.545}_{\pm 0.225}$	$\mathbf{0.194}_{\pm 0.119}$	$\mathbf{0.395}_{\pm 0.078}$
TM	5	$0.470_{\pm 0.259}$	$0.128_{\pm 0.106}$	$0.411_{\pm 0.077}$
FR	20	$0.489_{\pm 0.217}$	$0.005_{\pm 0.600}$	$0.249_{\pm 0.044}$
FR	10	$0.511_{\pm 0.208}$	$0.040_{\pm 0.561}$	$0.240_{\pm 0.044}$
FR	5	$\mathbf{0.698}_{\pm 0.096}$	$\mathbf{0.272}_{\pm 0.311}$	$\mathbf{0.211}_{\pm 0.027}$

Table 8: Joint alignment-weight sensitivity on the same three-subject subsets as Table 4, with $\lambda_{\mathrm{pcl}} = \lambda_{\mathrm{supcon}} = \lambda$ and all other settings fixed. Entries are mean $\pm$ between-subject standard deviation.

Dataset	λ	PCC↑	R^2 ↑	RMSE↓
FBM	0.2	$0.459_{\pm 0.407}$	$0.201_{\pm 0.472}$	$0.754_{\pm 0.363}$
FBM	1	$\mathbf{0.568}_{\pm 0.280}$	$\mathbf{0.232}_{\pm 0.457}$	$\mathbf{0.734}_{\pm 0.355}$
FBM	5	$0.443_{\pm 0.417}$	$0.201_{\pm 0.489}$	$0.762_{\pm 0.374}$
TM	0.2	$0.528_{\pm 0.232}$	$0.149_{\pm 0.117}$	$0.403_{\pm 0.077}$
TM	1	$\mathbf{0.545}_{\pm 0.225}$	$\mathbf{0.194}_{\pm 0.119}$	$\mathbf{0.395}_{\pm 0.078}$
TM	5	$0.472_{\pm 0.281}$	$0.126_{\pm 0.131}$	$0.410_{\pm 0.080}$
FR	0.2	$0.598_{\pm 0.159}$	$0.095_{\pm 0.507}$	$0.229_{\pm 0.045}$
FR	1	$\mathbf{0.698}_{\pm 0.096}$	$\mathbf{0.272}_{\pm 0.311}$	$\mathbf{0.211}_{\pm 0.027}$
FR	5	$0.618_{\pm 0.162}$	$0.154_{\pm 0.416}$	$0.225_{\pm 0.035}$

D.1.3 Frozen versus trainable Stage-3 codebook

We compare frozen and trainable Stage-3 codebooks on all 9/6/5 FBM/TM/FR subjects. The two variants share the same initialization, data, seed, motion-space MSE objective, optimizer, scheduler, iteration budget, and checkpoint criterion. Both update the EEG encoder, E2M Transformer, and VQ decoder while keeping the VQ encoder fixed; the trainable variant additionally updates the motion codebook. The frozen-codebook setting achieves higher mean PCC and R^2 and lower mean RMSE on all three datasets (Table 9), supporting the default choice of keeping the motion-prior interface fixed during Stage 3.

Table 9: Stage-3 codebook fine-tuning on the complete cohorts (FBM: 9; TM: 6; FR: 5 subjects). Entries are mean $\pm$ between-subject standard deviation. Both variants update the EEG encoder, E2M Transformer, and VQ decoder; only codebook trainability differs.

Dataset	Codebook	PCC↑	R^2 ↑	RMSE↓
FBM	Frozen	$\mathbf{0.522}_{\pm 0.238}$	$\mathbf{0.110}_{\pm 0.354}$	$\mathbf{0.718}_{\pm 0.229}$
FBM	Trainable	$0.496_{\pm 0.270}$	$0.106_{\pm 0.362}$	$0.720_{\pm 0.230}$
TM	Frozen	$\mathbf{0.485}_{\pm 0.211}$	$\mathbf{0.184}_{\pm 0.148}$	$\mathbf{0.404}_{\pm 0.052}$
TM	Trainable	$0.463_{\pm 0.222}$	$0.180_{\pm 0.147}$	$0.405_{\pm 0.051}$
FR	Frozen	$\mathbf{0.657}_{\pm 0.152}$	$\mathbf{0.250}_{\pm 0.340}$	$\mathbf{0.230}_{\pm 0.038}$
FR	Trainable	$0.638_{\pm 0.153}$	$0.214_{\pm 0.341}$	$0.237_{\pm 0.039}$

Table 10: FBM Stage-3 supervision scope on all nine subjects. Both settings are evaluated with paired 30-D end-effector metrics and full-263-D FID. Entries are mean ± between-subject standard deviation; the remaining settings are identical.

Supervised dimensions	PCC↑	R^2 ↑	RMSE↓	FID↓
30-D (default)	$0.522_{\pm 0.238}$	$0.110_{\pm 0.354}$	$0.718_{\pm 0.229}$	$4.14_{\pm 4.04}$
263-D	$0.427_{\pm 0.268}$	$0.025_{\pm 0.335}$	$0.773_{\pm 0.231}$	$4.84_{\pm 5.49}$

D.1.4 FBM REFINEMENT SUPERVISION SCOPE

We compare default 30-D end-effector supervision with full-263-D supervision on all nine FBM subjects, retaining the same remaining settings. Table 10 shows that full supervision yields PCC/R^2/RMSE of $0.427 \pm 0.268/0.025 \pm 0.335/0.773 \pm 0.231$ and FID 4.844 ± 5.494, versus $0.522 \pm 0.238/0.110 \pm 0.354/0.718 \pm 0.229$ and FID 4.139 ± 4.036 for the default. Thus, full supervision recovers neither Stage-2 FID (3.169 ± 2.037) nor default end-effector accuracy. Paired MSE and evaluator-space FID measure different properties, and the evaluator's domain mismatch further limits interpretation of small FID differences.

D.2 FREQUENCY-BAND ABLATION STUDY

We conduct a frequency-band ablation to examine which EEG spectral components are most informative for motion decoding under our experimental setting. For each subject and session, we apply band-pass filtering to the EEG signals and train/evaluate NeuroAction using the same preprocessing, architecture, data split, and optimization protocol. We compare individual frequency bands, including *delta*, *theta*, *alpha*, *beta*, and *gamma*, together with an *all-band* setting using the full preprocessed EEG spectrum. Reconstruction performance is reported using PCC and RMSE, where higher PCC and lower RMSE indicate better decoding.

Figure 7 summarizes the FBM results. Low-frequency components provide the strongest predictive information under this setting. Delta-band and all-band EEG show higher mean PCC and lower mean RMSE than the higher-frequency bands, whereas alpha, beta, and gamma filtering generally degrades reconstruction performance. These results suggest that NeuroAction relies more strongly on low-frequency EEG structure for naturalistic locomotion decoding, although the present analysis does not isolate neural activity from other low-frequency signal components.

This trend is consistent with prior EEG-based gait decoding studies. For example, Luu et al. Luu et al. (2016) showed that lower-limb joint angles during treadmill walking can be decoded from delta-band / slow-cortical-potential EEG using an Unscented Kalman filter. More broadly, low-frequency EEG has been associated with gait-cycle-related cortical dynamics and long-timescale sensorimotor integration during rhythmic movement Gwin et al. (2011); Knaepen et al. (2015). Higher-frequency components in mobile EEG can also be more affected by muscle activity and motion-related artifacts, which may reduce their reliability when used in isolation Delorme & Makeig (2004).

D.3 RSA ALIGNMENT ANALYSIS

We further analyze neuro-motion alignment using representational similarity analysis (RSA) Kriegeskorte et al. (2008). We compare the representational structure of EEG embeddings and motion representations by computing representational dissimilarity matrices (RDMs) in the two spaces and reporting their Spearman correlation at each token position. Figure 8 reports token-level RSA curves on FBM subjects `S03`, `S06`, `S09`, and `S10`, and FR subjects `S05`, `S07`, `S13`, and `S22`.

In the RSA plots, *Real RSA (latent)* measures the correspondence between EEG embeddings and continuous motion latents before quantization, while *Real RSA (quant)* measures the correspondence with quantized codebook embeddings. *Chance-Global* globally permutes the correspondence between EEG and motion representations, whereas *Chance-Strat* permutes EEG–motion correspondences within motion-code clusters, providing a structure-preserving null baseline. The *Ceiling* curve provides an empirical reliability ceiling estimated from split-half consistency. Across subjects, the real RSA curves are generally above the chance baselines, supporting structural correspondence between the learned EEG representations and the motion-primitive space.

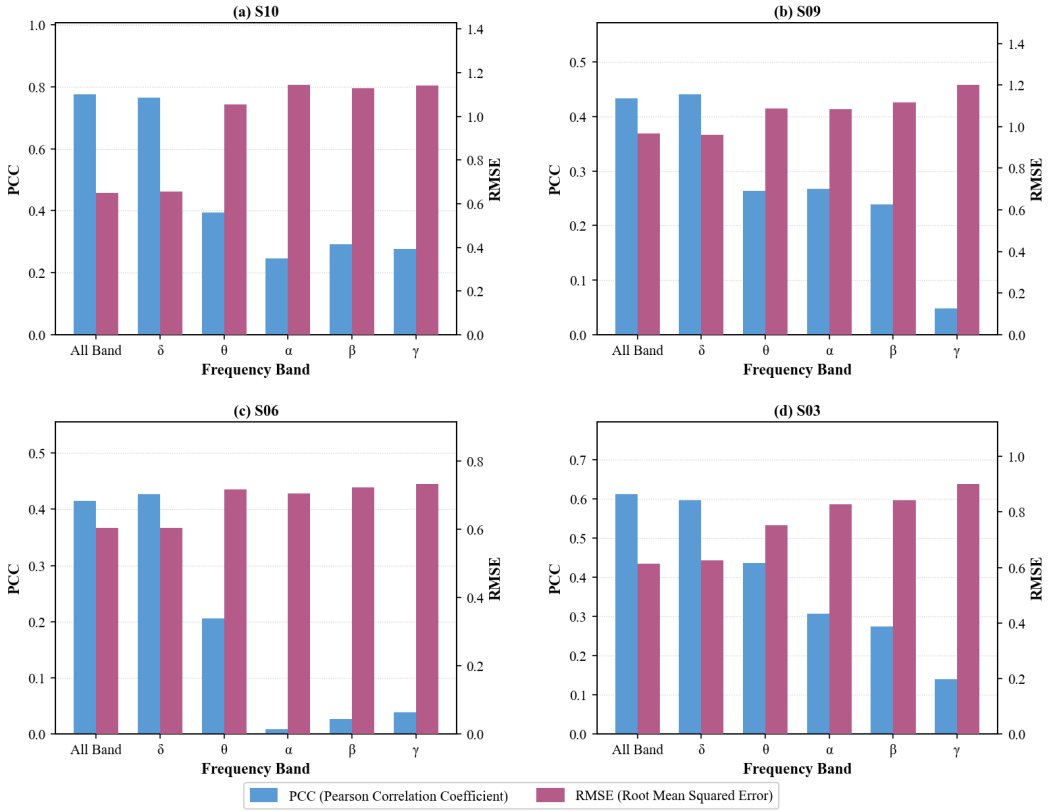

Figure 7: **Frequency-band analysis on FBM.** We report decoding performance under different EEG frequency-band settings, using PCC (higher is better) and RMSE (lower is better). Results for four FBM subjects (S10/S09/S06/S03) are summarized in a single panel, showing that low-frequency bands generally provide stronger decoding performance than higher-frequency bands.

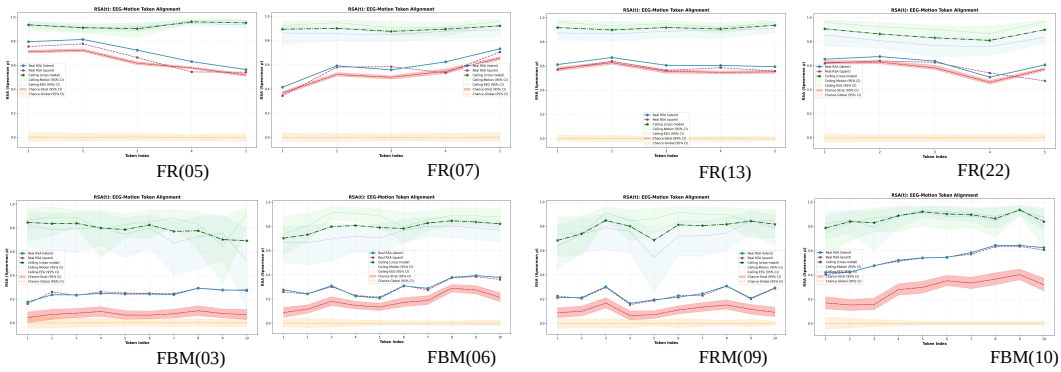

Figure 8: Token-level RSA alignment analysis on FBM and FR. Each curve shows the RSA score over the motion-token timeline, computed between EEG representations and motion-token representations. We visualize results for four FBM subjects (`S03`, `S06`, `S09`, `S10`) and four FR subjects (`sub-05`, `sub-07`, `sub-13`, `sub-22`).

D.4 Trajectory Distribution Visualization

We further visualize the distribution of generated and ground-truth trajectories using t-SNE. For each dataset, predicted and real trajectory windows from the held-out test session are flattened into vectors

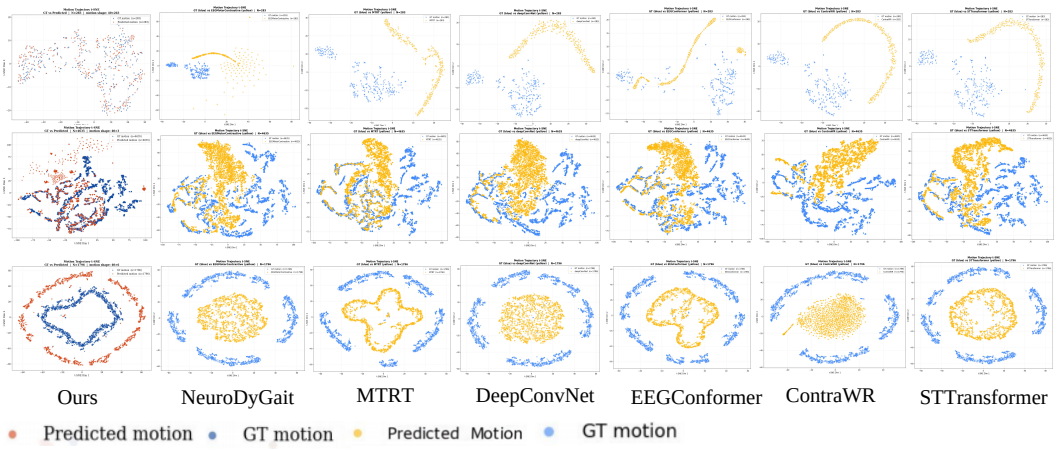

Figure 9: Trajectory distribution visualization using t-SNE on three representative subjects: FBM `S10`, FR `S05`, and TM `S07`. For each dataset, we compare the distributions of ground-truth trajectories, NeuroAction predictions, and baseline predictions in a shared two-dimensional embedding space. NeuroAction generally shows stronger overlap with the ground-truth distribution than the baselines.

and jointly embedded into a two-dimensional space. This visualization is intended as a qualitative diagnostic of distributional overlap rather than a quantitative metric.

Figure 9 compares NeuroAction with the baselines on three representative test subjects: FBM `S10`, FR `S05`, and TM `S07`. Across datasets, baseline predictions often form compact clusters or shifted regions relative to the ground-truth trajectories, indicating limited coverage of the real motion distribution. In contrast, NeuroAction shows stronger overlap with the ground-truth distribution, suggesting that it better preserves the distributional structure of the target motion space.

D.5 Full Subject-wise Baseline Comparisons

To complement the cohort-level results, Tables 11, 12, 13, and 14 report complete subject-wise comparisons. Reconstruction entries are reported as PCC/R^2/RMSE. Bold and underlined values denote the best and second-best results, respectively, determined using the original unrounded values. For FBM distributional metrics, FID is ranked in the same manner, whereas Diversity is reported without ranking because it should be interpreted relative to the real-motion reference.

D.6 Cohort Statistics and Paired Comparisons

Tables 11, 13, and 14 provide the complete subject-level results for all compared methods. Subjects, rather than windows or output dimensions, are treated as the statistical unit. For each dataset and reconstruction metric, we identify the strongest competing baseline according to its cohort-level mean performance and conduct a two-sided paired t-test between NeuroAction Stage 3 and that baseline across subjects.

On FBM, NeuroAction outperforms the strongest baseline, NeuroDyGait, on 9/9 subjects for PCC and 7/9 subjects for both R^2 and RMSE. The corresponding p-values are 0.00027, 0.0229, and 0.0246, respectively. On TM, NeuroAction outperforms NeuroDyGait on 4/6 subjects for each metric, with $p = 0.291$, 0.202, and 0.222 for PCC, R^2, and RMSE, respectively. On FR, NeuroAction outperforms NeuroDyGait on 5/5 subjects for PCC and R^2 and 4/5 subjects for RMSE, with corresponding p-values of 0.0407, 0.0519, and 0.125. The reported p-values are uncorrected for multiple comparisons, and the small TM and FR cohorts limit statistical power.

Relative to Stage 2, Stage 3 improves all three reconstruction metrics for 9/9 FBM and 6/6 TM subjects. On FR, PCC improves for 5/5 subjects and R^2/RMSE for 4/5, with `S07` unchanged on the latter two metrics.

Table 11: Full subject-level reconstruction results on FBM. PCC, R^2, and RMSE are reported in three vertically stacked panels. Bold and underlined values denote the best and second-best results, respectively, determined using the original unrounded values.

(a) PCC ↑

Method	s02	s03	s04	s05	s06	s07	s08	s09	s10
DCN	−0.002	0.010	0.022	0.014	0.006	0.032	0.135	0.037	0.037
EC	0.035	0.023	0.027	0.017	0.012	0.009	0.039	0.024	0.050
ContraWR	0.018	0.002	0.006	−0.012	0.026	0.066	0.074	0.004	0.054
DY	0.091	0.068	0.027	0.029	0.046	0.017	0.054	0.022	0.088
ST	0.029	0.023	0.022	0.009	0.006	0.020	0.020	0.009	0.010
MTRT	0.039	0.055	0.023	0.032	0.014	0.035	0.022	0.041	0.034
Ours	**0.333**	**0.725**	**0.345**	**0.093**	**0.674**	**0.494**	**0.526**	**0.632**	**0.878**

(b) R^2 ↑

Method	s02	s03	s04	s05	s06	s07	s08	s09	s10
DCN	−0.203	−0.114	−0.147	−0.091	−0.381	−0.292	−0.444	−0.515	−0.166
EC	−0.156	−0.220	−0.447	−0.252	−0.493	−0.384	−0.422	−0.123	−0.263
ContraWR	−0.153	−0.092	−0.323	−0.165	−0.372	−0.282	−0.395	−0.489	−0.083
DY	−0.043	−0.039	**−0.107**	−0.093	−0.251	−0.263	−0.684	−0.313	−0.115
ST	−0.138	−0.091	−0.338	−0.205	−0.493	−0.369	−0.498	−0.893	−0.259
MTRT	−0.016	−0.015	−0.148	**−0.088**	−0.311	−0.263	−0.569	−0.546	−0.048
Ours	**0.068**	**0.474**	−0.330	−0.104	**0.143**	**−0.120**	**−0.228**	**0.337**	**0.749**

(c) RMSE ↓

Method	s02	s03	s04	s05	s06	s07	s08	s09	s10
DCN	1.063	0.793	0.762	0.697	0.646	1.313	1.124	1.109	1.051
EC	0.682	0.821	0.837	0.741	0.673	1.340	1.137	1.050	1.092
ContraWR	0.683	0.785	0.816	0.719	0.637	1.300	1.134	1.045	1.017
DY	0.653	0.770	**0.750**	0.698	0.618	1.301	1.171	1.078	1.030
ST	0.678	0.785	0.812	0.729	0.668	1.337	1.155	1.217	1.090
MTRT	0.647	0.763	0.764	**0.697**	0.629	1.310	1.155	1.111	1.005
Ours	**0.619**	**0.536**	0.829	0.702	**0.465**	**1.133**	**0.949**	**0.783**	**0.451**

Table 12: Full subject-level distributional results on FBM. Each entry reports FID/Diversity. Lower FID is better; Diversity is reported without ranking because it is interpreted relative to the real-motion reference.

Method	s02	s03	s04	s05	s06	s07	s08	s09	s10
DCN	73.8/0.339	45.2/2.79	37.8/3.61	48.9/3.03	91.3/3.77	57.0/5.30	10.6/1.07	70.7/1.92	128/4.46
EC	2.66/0.647	5.84/0.429	22.4/2.84	1.91/0.665	23.8/2.51	30.9/1.84	9.60/0.389	25.1/2.84	77.8/7.06
ContraWR	44.0/3.50	57.0/3.02	22.6/0.849	52.3/3.53	71.3/2.65	55.4/5.58	27.5/3.28	61.5/0.235	120/2.69
DY	46.8/2.57	50.8/2.40	38.5/5.14	41.7/2.47	88.0/4.95	74.8/5.59	19.9/3.01	76.9/4.01	149/5.28
ST	2.86/0.635	5.84/0.429	77.0/4.68	3.03/0.665	36.6/3.67	62.7/4.76	9.48/0.602	67.7/0.263	77.7/7.01
MTRT	23.4/1.57	21.2/1.52	30.7/4.95	25.3/1.90	36.9/5.16	22.9/5.75	13.3/1.59	21.7/3.36	104/5.03
Ours	**1.24**/1.74	**1.70**/2.22	**12.0**/4.66	**0.867**/1.33	**2.86**/7.69	**5.05**/6.58	**1.90**/2.15	**1.92**/3.57	**9.67**/9.33

D.7 Inference Cost and Timing Scope

We profile the default models offline on a single NVIDIA L4 GPU using batch-1 FP32 PyTorch eager execution, with 100 warm-up runs and five repeats of 1,000 measurements. Tables 15 and 16 report the number of parameters, serialized checkpoint size, MACs per input window, peak allocated GPU memory, mean/P95 latency, and throughput. Model latency measures GPU forward execution only. Post-window latency starts from a preprocessed EEG window in CPU memory and ends with the reconstructed motion returned to CPU memory, including host–device transfers.

Table 13: Full subject-level reconstruction results on FR. Each entry reports PCC/R^2/RMSE. Bold and underlined values denote the best and second-best results, respectively, determined using the original unrounded values.

Method	s05	s07	s13	s15	s22
DCN	0.708/0.438/0.205	0.614/ − 0.583/0.274	0.658/0.374/0.284	0.146/ − 0.434/0.311	0.406/0.132/0.244
EC	0.727/0.441/0.206	0.493/ − 0.852/0.298	0.635/0.302/0.300	0.120/ − 0.652/0.333	0.397/ − 0.009/0.262
ContraWR	0.683/0.259/0.239	0.712/ − 0.103/**0.234**	0.646/0.351/0.289	−0.022/ − 0.807/0.342	0.548/0.270/0.225
DY	0.740/0.527/0.189	0.683/ − 0.264/0.248	0.749/0.507/0.252	0.390/−0.174/**0.278**	0.476/0.212/0.232
ST	0.714/0.425/0.209	0.490/ − 0.468/0.276	0.628/0.312/0.298	0.152/ − 0.790/0.340	0.459/0.190/0.237
MTRT	0.721/0.463/0.203	0.552/−**0.036**/0.237	0.626/0.366/0.286	0.093/ − 0.449/0.313	0.533/0.210/0.232
Ours	**0.778**/**0.562**/**0.182**	**0.724**/−0.057/0.235	**0.773**/**0.581**/**0.232**	**0.421**/−**0.146**/0.287	**0.592**/**0.312**/**0.217**

Table 14: Full subject-level reconstruction results on TM. Each entry reports PCC/R^2/RMSE. Bold and underlined values denote the best and second-best results, respectively, determined using the original unrounded values.

Method	s01	s02	s05	s06	s07	s08
DCN	0.384/0.084/0.449	0.126/ − 0.237/0.535	0.111/ − 0.295/0.466	**0.447**/**0.165**/**0.423**	0.633/0.111/0.372	**0.632**/0.314/0.413
EC	0.005/ − 0.051/0.484	0.340/ − 0.084/0.500	0.088/ − 0.047/0.420	0.419/0.122/0.434	0.598/0.004/0.389	0.548/0.251/0.431
ContraWR	−0.030/ − 0.004/0.479	0.004/ − 0.028/0.487	0.012/ − 0.045/0.420	−0.003/ − 0.028/0.471	−0.017/ − 0.237/0.464	0.000/ − 0.006/0.501
DY	0.391/0.029/0.459	0.385/0.080/0.460	0.206/−**0.018**/0.412	0.447/0.164/0.423	0.653/0.142/0.359	0.617/0.363/0.398
ST	0.072/ − 0.032/0.479	0.300/ − 0.057/0.494	**0.218**/−0.019/**0.411**	0.427/0.139/0.430	0.703/0.089/0.364	0.545/0.260/0.427
MTRT	0.000/ − 0.024/0.478	0.001/ − 0.010/0.483	0.005/ − 0.037/0.419	0.026/ − 0.046/0.475	0.007/ − 0.267/0.467	0.384/0.038/0.489
Ours	**0.456**/**0.167**/**0.430**	**0.444**/**0.169**/**0.438**	0.189/ − 0.023/0.415	0.388/0.089/0.443	**0.803**/**0.323**/**0.305**	0.632/**0.379**/**0.392**

Table 15: Default-model size and inference resources on one NVIDIA L4 (batch 1, FP32). MACs are per EEG window; memory is peak allocated GPU memory.

Dataset	Params (M)	CKPT (MiB)	MACs (G)	Peak (MiB)
FBM	80.217	306.17	1.430	325.4
TM	20.156	77.05	0.373	150.1
FR	21.008	80.29	0.214	126.5

Table 16: Offline inference timing on one NVIDIA L4: 100 warm-ups and five repeats of 1,000 measurements. Model latency is GPU forward time; post-window latency includes CPU–GPU transfers from preprocessed CPU EEG to reconstructed CPU motion.

Dataset	Model mean/P95 (ms)	Post-window mean/P95 (ms)	Windows/s
FBM	17.750/20.071	18.082/21.610	56.84
TM	17.474/18.286	17.572/18.420	57.91
FR	17.102/17.770	17.289/18.143	57.80

Across the three default models, the largest post-window P95 latency is 21.610 ms on FBM, with peak allocated GPU memory of 325.4 MiB. These measurements characterize only the model-side computation after a preprocessed 2-s EEG window becomes available; they exclude EEG acquisition, preprocessing, feedback, and control. No complete online closed-loop BCI experiment has been conducted. End-to-end acquisition-to-action latency, feedback stability, and online calibration therefore remain to be evaluated.

Table 17: EEG encoder architecture. C denotes the number of EEG channels, N_e denotes the number of EEG tokens, and d_e denotes the EEG embedding dimension. These values are dataset-dependent.

Component	Architecture
Input reshape	(0): Unsqueeze channel dimension $[B, C, T] \to [B, 1, C, T]$
Temporal patch embedding	(1): Conv2d$(1, 40,$ kernel $= (1, 10),$ stride $= (1, 1),$ pad $= (0, 0))$ $[B, 1, C, T] \to [B, 40, C, T - 9]$
Channel fusion	(2): Conv2d$(40, 40,$ kernel $= (C, 1),$ stride $= (1, 1),$ pad $= (0, 0))$ $[B, 40, C, T - 9] \to [B, 40, 1, T - 9]$
Normalization + activation	(3): BatchNorm2d$(40) \to$ ELU $[B, 40, 1, T - 9] \to [B, 40, 1, T - 9]$
Adaptive pooling	(4): AdaptiveAvgPool2d$((1, N_e))$ $[B, 40, 1, T - 9] \to [B, 40, 1, N_e]$
Patch dropout	(5): Dropout$(p{=}0.5)$
Projection to model dimension	(6): Conv2d$(40, d_e,$ kernel $= (1, 1),$ stride $= (1, 1))$ $[B, 40, 1, N_e] \to [B, d_e, 1, N_e]$
Rearrange to token sequence	(7): Rearrange "$b\ e\ 1\ w \to b\ w\ e$" $[B, d_e, 1, N_e] \to [B, N_e, d_e]$
Learned positional embedding	(8): Pos $\in \mathbb{R}^{1 \times N_e \times d_e}$, added to EEG tokens $[B, N_e, d_e] \to [B, N_e, d_e]$
Transformer encoder blocks $\times 6$	(9–14): Repeat for $l{=}1 \ldots 6$: LayerNorm(d_e) MultiHeadAttention(dim $= d_e,$ heads $= 8,$ attn_drop $= 0.5)$ Residual + Dropout$(p{=}0.5)$ LayerNorm(d_e) FFN: Linear$(d_e, 4d_e) \to$ GELU $\to$ Dropout$(0.5) \to$ Linear$(4d_e, d_e)$ Residual

E Model Architecture Details

E.1 EEG Encoder Architecture

Input/Output. The EEG encoder maps a raw EEG window $\mathbf{X} \in \mathbb{R}^{B \times C \times T}$ to a sequence of EEG tokens $\mathbf{E} \in \mathbb{R}^{B \times N_e \times d_e}$, where the number of EEG channels C and hidden dimension d_e are dataset-dependent. Unless otherwise stated, we use dropout $p{=}0.5$ throughout the encoder.

E.2 E2M Transformer Architecture

Input/Output. The E2M module performs parallel motion-token prediction conditioned on EEG tokens. Given EEG features $\mathbf{E} \in \mathbb{R}^{B \times N_e \times d_e}$, we first project them to the E2M hidden dimension d_m. We then concatenate the projected EEG tokens with a set of learnable target slots $\mathbf{Q} \in \mathbb{R}^{N_m \times d_m}$, expanded along the batch dimension. The resulting sequence length is $L = N_e + N_m$. A single Transformer forward pass outputs logits over the motion codebook, and the predicted motion tokens are obtained from the final N_m target-slot positions. The additive attention mask on the concatenated sequence is

$$M_{ij} = \begin{cases} 0, & j \le i, \\ -\infty, & j > i. \end{cases}$$

Thus each target slot attends to the full EEG prefix and preceding/current slots. Predictions are not fed back, and all target logits are computed in one pass. The EEG encoder itself uses unmasked

Table 18: Tensor notation. Mathematical code IDs are one-based. We configure $N_e = N_m = N$ and $d_e = d_c$, while keeping the symbols distinct to identify their roles.

Symbol	Meaning	Shape/domain
H	EEG encoder output	$B \times N_e \times d_e$
H^{e2m}	Adapted EEG tokens	$B \times N_e \times d_m$
Q^B	Batched target slots	$B \times N_m \times d_m$
Z^{e2m}	E2M hidden states	$B \times (N_e + N_m) \times d_m$
P	Motion-token logits	$B \times N_m \times K$
$\mathcal{C}$	Motion codebook	$K \times d_c$
$s_{b,i}$	Discrete code index	$\{1, \ldots, K\}$
G	One-hot selections	$B \times N_m \times K$

Table 19: E2M Transformer architecture used for parallel motion-token decoding. N_e denotes the number of EEG tokens, N_m denotes the number of motion-token slots, d_e denotes the EEG encoder output dimension, d_m denotes the E2M hidden dimension, and K denotes the motion codebook size. In our implementation, N_m is matched to N_e.

Component	**Architecture**
EEG conditioning projection	(0): Linear(d_e, d_m) Input: $\mathbf{E} \in \mathbb{R}^{B \times N_e \times d_e}$, Output: $\mathbb{R}^{B \times N_e \times d_m}$
Learnable target slots	(1): Unified learnable token embeddings $\mathbf{Q} \in \mathbb{R}^{N_m \times d_m}$ The target slots are used across all subjects and expanded to $\mathbb{R}^{B \times N_m \times d_m}$.
Concatenation	(2): Concatenate projected EEG tokens and target slots along the token dimension: $\mathbb{R}^{B \times N_e \times d_m} \oplus \mathbb{R}^{B \times N_m \times d_m} \to \mathbb{R}^{B \times (N_e + N_m) \times d_m}$
Sine-cosine position encoding	(3): Fixed sin-cos positional encoding for length $L = N_e + N_m$, added to the concatenated sequence.
Dropout	(4): Dropout(p=0.5)
Causal self-attention blocks $\times N_{\mathrm{e2m}}$	(5): Repeat for $l = 1, \ldots, N_{\mathrm{e2m}}$: LayerNorm$(d_m)$ CausalSelfAttention(dim $= d_m$, heads $= h$, attn_drop $= 0.5$) with causal mask Residual + Dropout(p=0.5) LayerNorm(d_m) MLP: Linear$(d_m, 4d_m)$ $\to$ GELU $\to$ Linear$(4d_m, d_m)$ $\to$ Dropout(0.5) Residual
Final normalization	(7): LayerNorm(d_m)
Vocabulary head	(8): Linear(d_m, K) (bias=False), where K is the motion codebook size Output logits: $\mathbb{R}^{B \times (N_e + N_m) \times K}$; the final N_m positions correspond to predicted motion tokens.

attention. Learned EEG positional embeddings fit fixed-length observations; fixed sinusoidal E2M positions order conditioning and target slots without additional learned parameters. This positional choice is pragmatic, not a separate contribution.

F Further Discussion

F.1 Limitations.

First, our experiments are restricted to a small number of subjects, reflecting the current scarcity of synchronized EEG–motion datasets and limiting population-level generalization Brantley et al. (2018); Paulo et al. (2025). Second, our motion interface relies on a VQ-VAE tokenizer: discretizing continuous human motion into a finite codebook provides a strong inductive bias and stabilizes long-horizon generation, but it also introduces quantization error and restricts expressivity to the learned vocabulary, potentially attenuating fine-grained kinematics and stylistic nuances. Encouragingly, recent text-to-motion models have begun to reflect on this limitation and explore alternatives that interface autoregressive backbones with *continuous* motion latents (e.g., VAE latents) and use lightweight diffusion modules for high-fidelity synthesis, mitigating the symbolic–continuous mismatch induced by tokenization Zhu et al. (2026). Adapting such continuous-latent or diffusion-bridged motion priors to EEG-conditioned generation is a promising direction, but requires revisiting token-level alignment objectives and efficient online inference. Finally, TM/FR cross-day evaluation and offline L4 profiling address limited aspects of generalization and efficiency. FBM has no verified day metadata, long-term drift is untested, and a bidirectional encoder requires the available observation window. Streaming acquisition, feedback stability, and online user adaptation remain open challenges.

F.2 Broader implications.

Broadly, this work suggests that successful neural decoding of complex behavior may require representational alignment rather than direct signal-to-trajectory mapping Karpowicz et al. (2025). Our framework offers a general blueprint for linking neural data to structured generative models, potentially extending beyond motion to other high-dimensional behavioral or perceptual domains.