EEG2MOTION: SYNTHESIZING DIVERSE HUMAN MOTIONS FROM EEG SIGNALS DURING MOTION OBSERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human motion is governed by a hierarchical motor system where the brain provides high-level intentions and lower-level structures coordinate detailed dynamics. Existing brain-computer interfaces (BCIs) typically oversimplify this into constrained classification or low-dimensional control, failing to capture the richness of natural movement. Bridging this gap to achieve diverse, full-body motion synthesis remains challenging due to the substantial cross-modal divergence between sparse neural signals and high-dimensional kinematics, as well as the lack of large-scale paired EEG-motion datasets. To address this, we introduce EEG2MOTION, the first EEG-motion-text dataset for human motion synthesis from motion-observation EEG, comprising nearly 20,000 paired samples across thousands of motions. Using this dataset, we first demonstrate via multimodal contrastive learning that non-invasive EEG embeddings can be aligned with text, video, and motion representations to decode high-level semantics. We then propose EEG-conditioned Masked Motion Model (EMMM), a generative framework that unites an EEG encoder with a motion decoder to synthesize continuous, full-body human motions directly from brain activity. Experimental results show that EMMM generates coherent and realistic motion sequences from non-invasive brain signals. To the best of our knowledge, this is the first work to generate diverse full-body human motions from non-invasive brain signals, opening a new direction toward generative and open-vocabulary motor BCIs. See our anonymous project page: https://eeg2motion-demo-2026.github.io/EEG2MOTIONdemo_page/.

1 INTRODUCTION

Human movement is governed by a hierarchical motor system in which cortical signals provide high-level motor information while lower-level neural and peripheral structures coordinate detailed movement dynamics (Frank & Earl, 1990; Rizzolatti & Luppino, 2001; Shadmehr & Krakauer, 2008). This biological architecture is particularly critical for the design of Brain-Computer Interfaces (BCIs) (Wolpaw, 2007). Because not all information required to reconstruct natural human motions is directly observable from cortical activity, motor BCIs require an intelligent "copilot" capable of receiving the multi-layered, sparse motor signals transmitted by the cortex and translating them into complete, natural human motions (Lee et al., 2025).

Most existing motor BCIs simplify movement decoding into highly constrained paradigms. Typical examples include classifying a small number of motor imagery (MI) categories or predicting two-dimensional and three-dimensional cursor trajectories (Wang et al., 2012; Wolpaw et al., 1991). While such paradigms have demonstrated practical utility, their limited motion space restricts deployment in complex real-world scenarios and provides only a partial understanding of the intricate brain-body interaction underlying human movement. In contrast to conventional constrained decoding tasks, we investigate, for the first time, the feasibility of open-vocabulary motion semantic retrieval from non-invasive electroencephalography (EEG) and further explore the synthesis of full-body human motions directly from brain activity.

To enable full-body motion synthesis, we adopt an Action Observation (AO) paradigm rather than active execution or motor imagery (Hardwick et al., 2018). This design is motivated by several crit-

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.

ical considerations. Active execution introduces substantial motion artifacts, while motor imagery is difficult to scale to thousands of diverse actions and exhibits considerable inter-subject variability (Altaheri et al., 2023). Action observation can elicit neural components that partially overlap with those involved in motor execution (Hardwick et al., 2018), and previous studies on image and video generation have shown that semantic features can be captured from visual EEG signals (Bai et al., 2023; Liu et al., 2024b). Consequently, we designed an experimental paradigm where subjects wear EEG caps while observing videos of diverse human motions. We term this dataset EEG2MOTION. The motion samples are derived from the HumanML3D dataset (Guo et al., 2022), which contains text-to-motion data with each motion paired with one or more textual descriptions.

Leveraging this dataset, we conduct two core experiments. The first investigates multimodal contrastive learning (Oord et al., 2018), evaluating the feasibility of aligning EEG data with different modal representations. Textual descriptions, video frames, and motion posture sequences serve as the three target modalities.

Second, we propose a generative motion decoding framework based on masked motion model (MMM) (Pinyoanuntapong et al., 2024b). In this framework, EEG features extracted by an EEG encoder serve as conditioning signals for a motion generation model, while the motion decoder reconstructs human motion through masked token prediction. This formulation establishes a prototype of a generative motor decoding system: the brain provides sparse but informative motor cues, whereas the motion generation model acts as a "motion copilot," learning a prior over human movements and translating neural intentions into complete and natural full-body motions.

In summary, the primary contributions of this work are three-fold:

- **We introduce EEG2MOTION**, the first EEG dataset explicitly tailored for human motion synthesis, comprising nearly 20,000 paired EEG-motion-text samples and covering thousands of distinct human motions.
- **We benchmark multimodal contrastive learning on this dataset**, evaluating the alignment performance of EEG data against text, video, and motion features and achieving performance that significantly outperforms randomized baselines.
- **We propose EEG-conditioned Masked Motion Model (EMMM)**. By jointly optimizing an EEG encoder and a motion decoder, we achieve the first end-to-end synthesis of complete, natural human motions directly from non-invasive EEG signals.

2 Related Work

2.1 Text-to-Motion Generation

Benefiting from the emergence of large-scale annotated benchmarks such as KIT (Plappert et al., 2016) and HumanML3D (Guo et al., 2022), the field of cross-modal motion synthesis has witnessed rapid advancement. Contemporary methodologies primarily bifurcate into continuous Diffusion-based models (Tevet et al., 2022; Karunratanakul et al., 2023; Zhang et al., 2024) and discrete Autoregressive or Masked Transformers (Pinyoanuntapong et al., 2024b;a; Jiang et al., 2023). Beyond text-conditioned generation, recent work has explored fine-grained motion editing and complex conditioning, including joint trajectory control (Pinyoanuntapong et al., 2025) and dynamic scene interactions (Tevet et al., 2025). These advances enrich motion representations and provide useful priors for integrating human motion synthesis with brain-computer interfaces, by allowing neural signals to specify high-level semantics while learned motion priors complete fine-grained details.

2.2 Generative Brain Decoding

In recent years, neural decoding has shifted from constrained classification and regression toward Generative Brain Decoding (Shukla et al., 2025). In this paradigm, neural features are aligned with target modality representations to form a shared latent space, while pretrained generative foundation models provide modality priors that enable the synthesis of natural content from sparse neural cues.

Recent work has extended generative brain decoding across multiple modalities, such as in image synthesis (Bai et al., 2023; Kavasidis et al., 2017), video reconstruction (Liu et al., 2024b), speech

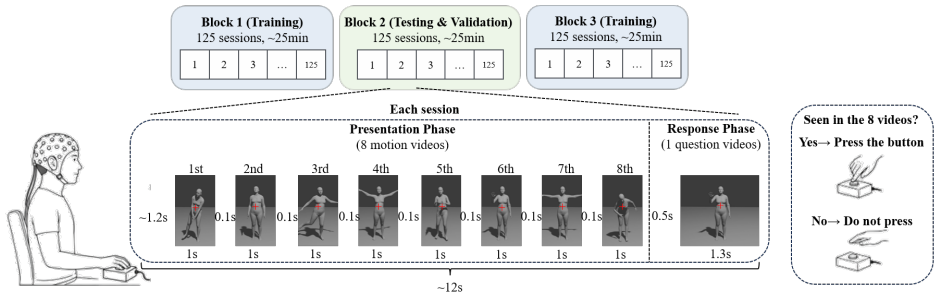

Figure 1: Illustration of the EEG2MOTION data collection paradigm. Participants wearing an EEG cap observed human motion videos rendered from HumanML3D motion sequences while EEG signals were continuously recorded. The experiment consisted of three blocks, each containing 125 sessions. The second block was reserved for validation and testing, whereas the remaining blocks were used for training. Each session contained 8 motion clips of one second each. Participants were instructed to maintain visual attention and indicate whether they remained attentive using a button press after each session.

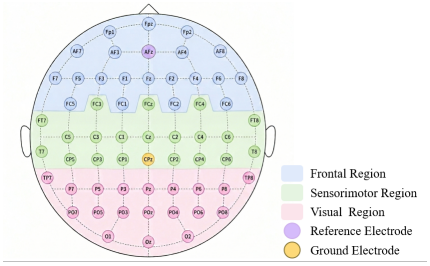

Figure 2: EEG electrode montage and regional partition. The 59-channel EEG setup followed the international 10–10 system. Electrodes were divided into three approximately balanced groups according to their spatial locations: frontal, sensorimotor, and posterior visual regions, corresponding to high-level action understanding, motor representation, and visual motion perception processes, respectively. AFz and CPz were used as the reference and ground electrodes.

synthesis (Littlejohn et al., 2025; Chen et al., 2024; Anumanchipalli et al., 2019), or open-vocabulary text translation (Liu et al., 2024a; Wang & Ji, 2022). Crucially, within the motor decoding domain, a parallel shift toward unconstrained, whole-body generation has recently begun; He et al. (He et al., 2026) decoded unconstrained whole-body kinematics in freely moving monkeys by conditioning an autoregressive generative model on intracranial cortical signals. However, the integration of generative human motion models with non-invasive EEG decoding remains entirely unexplored. Our work bridges this critical gap, extending open-vocabulary generative neural decoding into the realm of complex human kinematics.

3 EEG2MOTION Dataset

3.1 Data Acquisition Paradigm

Our experimental paradigm is shown in Fig. 1. Continuous scalp EEG was collected from 9 subjects using a 64-channel recording system at 1000 Hz, providing 59 channels after excluding channels like EOG (see Fig. 2). Each subject completed 3 recording blocks (one subject had an additional training block); block 2 was used for validation and testing, and the rest for training.

Subjects were seated comfortably and instructed to fixate on a central red cross to minimize eye movements. Visual stimuli consisted of motion data selected from the HumanML3D dataset (Guo et al., 2022), which were then rendered using Blender. Although each motion sequence is paired with textual descriptions in the dataset, only the videos were presented during the experiment. Motions with original durations of 4–6 seconds were selected and standardized to 5 seconds via truncation or

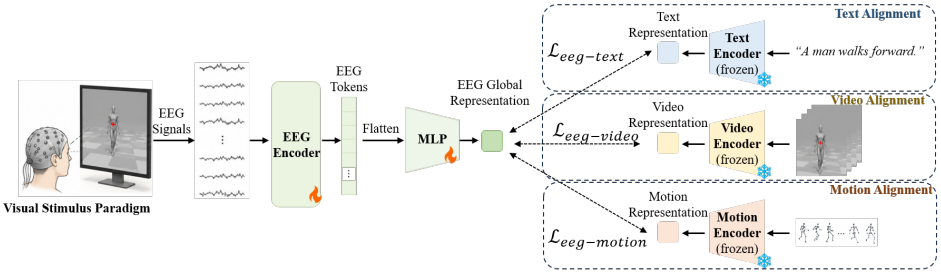

Figure 3: An overview of the proposed cross-modal contrastive learning framework. Continuous EEG signals are encoded into neural tokens via EEGConformer and mapped to the joint latent space using an MLP head. Frozen pre-trained target encoders (CLIP, VideoMAE, and Motion Encoder) extract target features. A symmetric InfoNCE loss is employed to align EEG features and target features

last-frame delay, and played at $5\times$ speed, yielding 1 second per motion clip. The training, validation, and test samples were drawn exclusively from corresponding HumanML3D splits to ensure test and validation motions were unseen during training.

Each block comprised 125 sessions. Each session lasted approximately 12 seconds, containing 8 sequential motion clips, followed by a 1.8-second response phase where subjects identified a query motion. This task was used solely to maintain visual attention, and data from the response phase was excluded from analysis. Consequently, each block yielded 1,000 paired EEG-motion-text samples (block 2 provided 250 validation and 750 test samples per subject).

3.2 Data Preprocessing

3.2.1 EEG Preprocessing.

Each 1-second EEG epoch was band-pass filtered from 0.5 to 45 Hz and downsampled to 250 Hz. Channel-wise z-score normalization was applied using parameters from each subject's training set. The 59 electrodes were partitioned into three functionally balanced regions (see Fig. 2): frontal, sensorimotor, and visual regions. The final EEG tensor shape is $C \times 250$ ($C = 59$).

3.2.2 Motion Preprocessing.

Human motion followed the standard 263-dimensional representation in HumanML3D Guo et al. (2022), encapsulating positions, velocities of 23 joints, and foot-ground contacts. Played back at $5\times$ speed from the original 20 FPS, each motion sequence spans a maximum of 100 frames.

3.2.3 Subject Screening and Dataset Compilation.

A preliminary experiment was conducted to assess each subject's EEG-motion contrastive alignment (see Supplementary Material). We excluded three subjects whose EEG representations failed to achieve significant contrastive alignment with motion features compared to randomized noise inputs. Data from the remaining subjects were merged into a unified dataset encompassing 13,000 training, 1,500 validation, and 4,500 test samples.

4 Method

4.1 Multimodal Contrastive Learning

To investigate whether EEG signals can be aligned with diverse modalities in an open-vocabulary scenario, we propose a cross-modal contrastive learning framework (Fig. 3).

We sequentially adopt text, video, and human motion as the target modality. Frozen pre-trained feature extractors—specifically the CLIP text encoder (Radford et al., 2021), VideoMAE (Tong

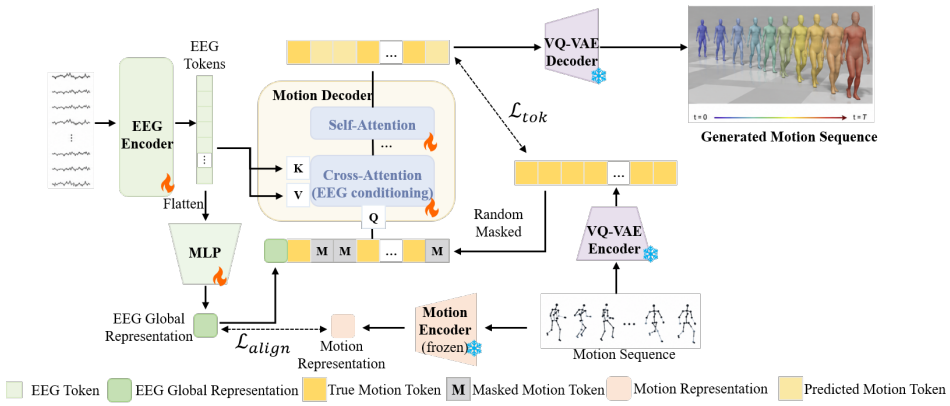

Figure 4: Overview of the EEG-conditioned Masked Motion Model framework. The architecture consists of an EEG encoder, a pre-trained VQ-VAE, and a Transformer-based motion decoder. Inside the decoder, global EEG features are prepended to motion tokens as a prompt prefix, while fine-grained EEG tokens provide neural guidance via cross-attention layers. During training, the alignment loss $\mathcal{L}_{\text{align}}$ is computed between the global EEG features and the target motion features. Concurrently, the ground-truth motion tokens are randomly masked and fed alongside the EEG inputs into the motion decoder, which is optimized via $\mathcal{L}_{\text{tok}}$ to predict the codebook indices corresponding to the masked positions.

et al., 2022), and a motion encoder pre-trained via contrastive learning with text (Guo et al., 2022)—are leveraged to extract target features $\mathbf{Z}_{\text{target}} \in \mathbb{R}^{B\times D}$ (target $\in \{\text{text}, \text{video}, \text{motion}\}$), where B and D denote the batch size and feature dimension, respectively. The size of D depends on the output feature dimension of the corresponding encoder. For VideoMAE, the features are obtained by averaging the output spatiotemporal tokens.

For the EEG modality, we employ the EEGConformer as the feature extractor (Song et al., 2022), where the batch normalization layers are replaced with subject-specific batch normalization to alleviate inter-subject variability. The raw EEG input $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$ (C: channels, T: time points) is transformed via a convolutional embedding layer and processed by Transformer blocks to yield the EEG tokens $\mathbf{E}_{\text{eeg}} \in \mathbb{R}^{B\times K\times N}$ (K: token number, N: token dimension). These tokens are then flattened and forwarded through an MLP projection head to produce the final EEG global features $\mathbf{Z}_{\text{eeg}} = \text{MLP}(\text{Flatten}(\mathbf{E}_{\text{eeg}})) \in \mathbb{R}^{B\times D}$.

To achieve cross-modal alignment, we apply a symmetric InfoNCE loss on the L_2-normalized features $\hat{\mathbf{Z}}_{\text{eeg}}$ and $\hat{\mathbf{Z}}_{\text{target}}$ (Radford et al., 2021):

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B}\sum_{i=1}^{B}\left(\log\frac{\exp(\mathbf{S}_{i,i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{S}_{i,j}/\tau)} + \log\frac{\exp(\mathbf{S}_{i,i}/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{S}_{j,i}/\tau)}\right) \tag{1}$$

where $\mathbf{S} = \hat{\mathbf{Z}}_{\text{eeg}}\hat{\mathbf{Z}}_{\text{target}}^{\top} \in \mathbb{R}^{B\times B}$ represents the cosine similarity matrix, and τ is a temperature parameter (set to 0.07 empirically) to scale the logits.

4.2 EEG-conditioned Masked Motion Model

To synthesize realistic human motions directly from neural representations, we propose an EEG-conditioned Masked Motion Model framework (Fig. 4). We integrate the EEG encoder with the motion decoder from MMM (Pinyoanuntapong et al., 2024b), a text-to-motion generation model, and train them together to enable EEG-conditioned human motion synthesis.

4.2.1 MODEL ARCHITECTURE.

The architecture comprises an EEG encoder, a pre-trained VQ-VAE for motion tokenization, and a Transformer-based motion decoder. The EEG encoder adopts the identical configuration as the contrastive learning section but is trained from scratch. Given the raw EEG input $\mathbf{X}$, it extracts the fine-grained EEG tokens $\mathbf{E}_{\text{eeg}} \in \mathbb{R}^{B\times K\times N}$ and the global EEG representation $\mathbf{Z}_{\text{eeg}} \in \mathbb{R}^{B\times D}$.

Following (Pinyoanuntapong et al., 2024b), the pre-trained VQ-VAE encoder compresses every 4 frames of continuous human motion into a single discrete token, yielding a target sequence $\mathbf{Y} = \{y_i\}_{i=1}^{L}$ of length L, while its decoder reconstructs continuous motions from these tokens. To effectively inject neural guidance, the global EEG feature $\mathbf{Z}_{\text{eeg}}$ is first prepended to the motion token sequence as a prompt prefix. This combined sequence is then fed into the Transformer-based motion decoder.

In the initial layers of the motion decoder, a cross-attention mechanism aligns motion queries with neural contexts, formulated as $\text{Attention}(\mathbf{Q}_m, \mathbf{K}_e, \mathbf{V}_e) = \text{softmax}(\mathbf{Q}_m\mathbf{K}_e^\top/\sqrt{D})\mathbf{V}_e$, where $\mathbf{Q}_m$ is derived from the motion tokens, and $\mathbf{K}_e, \mathbf{V}_e$ are projected from $\mathbf{E}_{\text{eeg}}$. The subsequent layers utilize self-attention to model temporal motion dynamics. Learnable special tokens, including [MASK], [PAD], and [END], are employed for masking, padding, and indicating the termination of the motion sequence.

4.2.2 TRAINING STRATEGY AND OBJECTIVES.

To regulate the information bottleneck between the EEG encoder and the motion decoder, the symmetric alignment loss $\mathcal{L}_{\text{align}}$ from the contrastive section remains active.

During training, a specific percentage of tokens in $\mathbf{Y}$ are randomly replaced with the [MASK] token, yielding the partially corrupted sequence $\mathbf{Y}_{\bar{\mathcal{M}}}$. The motion decoder is optimized to predict the codebook indices of the masked tokens conditioned on the unmasked context and the EEG signals. The token prediction loss $\mathcal{L}_{\text{tok}}$ is defined as:

$$\mathcal{L}_{\text{tok}} = -\mathbb{E}\left[\sum_{i\in\mathcal{M}} \log p(y_i \mid \mathbf{Y}_{\bar{\mathcal{M}}}, \mathbf{E}_{\text{eeg}}, \mathbf{Z}_{\text{eeg}})\right] \tag{2}$$

where $\mathcal{M}$ denotes the set of masked token indices.

To preserve the prior knowledge of the generative model, the weights of the motion decoder are frozen, while LoRA is deployed to fine-tune the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ projection matrices in the cross-attention layers and the $\mathbf{Q}, \mathbf{V}$ matrices in the self-attention layers (Hu et al., 2022). The EEG encoder is fully trainable.

4.2.3 INFERENCE STAGE.

During inference, the motion sequence length is fixed to $L = 25$ tokens (corresponding to 1 second of motion video) and entirely initialized with [MASK] tokens. Conditioned on the EEG input, a progressive iterative decoding strategy identical to MMM is executed. In each iteration, the model predicts the probability distribution over all masked positions, retains a subset of tokens with the highest confidence, and re-masks the remaining uncertain positions at a specific ratio until the complete sequence is recovered. The finalized token sequence is forwarded to the VQ-VAE decoder to synthesize the final motion sequence.

5 EXPERIMENTS

5.1 IMPLEMENTATION DETAILS.

We evaluate our framework on the EEG2MOTION dataset. The EEG encoder configuration follows EEGConformer (Song et al., 2022), and the motion decoder inherits hyperparameters from MMM (Pinyoanuntapong et al., 2024b) (with the text encoder removed); detailed specifications are provided in the Supplementary Material.

Evaluation Setting	Channels	Top1 ↑	Top2 ↑	Top3 ↑
shuffled baseline	all	$0.033^{\pm.004}$	$0.064^{\pm.004}$	$0.097^{\pm.005}$
noise baseline	all	$0.030^{\pm.003}$	$0.062^{\pm.004}$	$0.094^{\pm.006}$
eeg-text	all	$0.035^{\pm.002}$	$0.074^{\pm.004}$	$0.106^{\pm.005}$
eeg-video	visual only	$\mathbf{0.047}^{\pm.004}$	$\mathbf{0.090}^{\pm.005}$	$\mathbf{0.132}^{\pm.005}$
	all	$0.037^{\pm.004}$	$0.073^{\pm.005}$	$0.109^{\pm.005}$
eeg-motion	frontal only	$0.034^{\pm.002}$	$0.068^{\pm.004}$	$0.103^{\pm.004}$
	sensorimotor only	$0.038^{\pm.003}$	$0.072^{\pm.004}$	$0.102^{\pm.006}$
	visual only	$\underline{0.044}^{\pm.003}$	$0.086^{\pm.005}$	$\underline{0.126}^{\pm.007}$
	all	$0.041^{\pm.003}$	$\underline{0.087}^{\pm.006}$	$0.125^{\pm.006}$

Table 1: Retrieval accuracy of multimodal contrastive learning. Values are reported as mean with 95% confidence interval.

For motion synthesis, we compare against two additional strategies. The first is training a motion decoder from scratch, which uses the same architecture but with fewer layers and reduced hidden dimensions to prevent overfitting given data and hardware constraints. The second is a knowledge distillation strategy (Hinton et al., 2015), where the original text-to-motion model MMM serves as the teacher. Specifically, the same masked motion token sequences are fed into both our EEG-to-motion model and MMM, and a KL divergence distillation loss is computed between the output distributions at the corresponding token positions.

All models are trained for $10,000$ iterations on two NVIDIA RTX 3090 GPUs using the AdamW optimizer. The learning rate starts at 2×10^{-4} and decays to 1×10^{-4} after $8,000$ iterations. The optimal checkpoint is selected via validation performance. The default batch size is 512, which is adjusted to 256 during LoRA fine-tuning of the MMM decoder due to memory limits. For motion synthesis, loss coefficients for $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{tok}}$ are empirically set to 1.

5.2 Evaluation Metrics.

Operating in open-vocabulary scenarios, we adopt **R-Precision** as our primary metric. During testing, for each trial, 32 candidate targets (one matched ground-truth and 31 randomly selected distractors) are sampled from the test set. For EEG-to-text/video/motion retrieval, we compute cosine similarity between the EEG query and each candidate target, and rank the candidates accordingly. For motion synthesis, we compute the feature distance between the synthesized motion and 32 candidate motions, and rank them in ascending order. The Top-1, Top-2, and Top-3 accuracies (whether the ground-truth target ranks within the top k) are reported, averaged over multiple runs.

For motion synthesis, we further incorporate three established generative metrics from text-to-motion research (Guo et al., 2022). **Fréchet Inception Distance** (FID) measures the distribution distance between the feature of synthesized motions and real motions. **Multimodal Distance (MM-Dist)** measures the distance in feature space between synthesized motions and their paired ground-truth motions. **Diversity** calculates the average distance in feature space across 300 randomly sampled pairs of synthesized motions.

To address concerns that a powerful decoder might synthesize plausible motions independently, rendering EEG conditioning redundant, we evaluate two randomized control baselines: (1) a **Gaussian Noise Baseline** where input EEG is replaced with standard Gaussian noise, and (2) a **Shuffled Baseline** where EEG inputs are randomly paired with targets.

6 Results

6.1 Multimodal Contrastive Learning Results

The cross-modal retrieval results are summarized in Table 1. All three target modalities achieve retrieval accuracies that significantly outperform the random chance levels (detailed statistical tests

Evaluation Setting	$\mathcal{L}_{\text{align}}$	Channels	Method	Top3 ↑	MM-Dist ↓	FID ↓	Diversity ↑
shuffled baseline	eeg-motion	all	finetuning	$0.091^{\pm.005}$	$9.815^{\pm.079}$	-	-
noise baseline	eeg-motion	all	finetuning	$0.092^{\pm.006}$	$9.638^{\pm.074}$	$8.360^{\pm.601}$	$8.668^{\pm.112}$
motion retrieval	none	all	from scratch	$0.106^{\pm.004}$	$9.392^{\pm.049}$	$\mathbf{2.324}^{\pm.143}$	$9.199^{\pm.085}$
	eeg-video	all	from scratch	$\mathbf{0.123}^{\pm.007}$	$\mathbf{9.119}^{\pm.071}$	$7.130^{\pm.324}$	$8.926^{\pm.147}$
			finetuning	$0.110^{\pm.005}$	$9.447^{\pm.053}$	$5.721^{\pm.279}$	$9.298^{\pm.125}$
	eeg-motion	visual only	from scratch	$0.119^{\pm.005}$	$9.299^{\pm.046}$	$10.641^{\pm.523}$	$8.724^{\pm.136}$
			finetuning	$0.120^{\pm.006}$	$9.324^{\pm.063}$	$6.466^{\pm.368}$	$9.252^{\pm.086}$
		all	from scratch	$\underline{0.121}^{\pm.005}$	$\underline{9.258}^{\pm.053}$	$2.825^{\pm.264}$	$\mathbf{9.471}^{\pm.111}$
			distillation	$0.112^{\pm.005}$	$9.558^{\pm.048}$	$17.354^{\pm.585}$	$8.909^{\pm.126}$
			finetuning	$0.119^{\pm.006}$	$9.275^{\pm.056}$	$\underline{2.743}^{\pm.221}$	$\underline{9.411}^{\pm.097}$

Table 2: Performance of motion synthesis under different evaluation settings. Values are reported as mean with 95% confidence interval.

Evaluation Setting	Encoder Only	+ Decoder	Gain
eeg-video	$0.109^{\pm 0.005}$	$0.144^{\pm 0.005}$	+ 32.1%
eeg-motion	$0.125^{\pm 0.006}$	$0.143^{\pm 0.005}$	+ 14.4%

Table 3: Retrieval Top3 accuracy of EEG global representation and target modality features. Adding a motion decoder improves the alignment.

are in the Supplementary Material). To investigate the spatial contributions of neural dynamics, we conduct a channel ablation study on the video and motion retrieval tasks by comparing different channel configurations.

The empirical results reveal that channels covering the visual cortex play a mayor role in cross-modal alignment, occasionally surpassing the full-channel configuration. This suggests that under our current paradigm, visual cortical activations encode substantial task-related semantics driven by action observation mechanisms. Conversely, sensorimotor channels yield relatively limited retrieval capabilities, potentially because the rapid, passive video-viewing task inherently prioritizes visuospatial processing over active motor engagement. Nevertheless, these findings demonstrate that the joint latent space successfully bridges neural signals with open-vocabulary motion semantics, providing a solid foundation for subsequent motion synthesis.

6.2 Motion Synthesis Results

The evaluation of the EMMM framework is summarized in Table 2. To facilitate quantitative evaluation, the synthesized and ground-truth motions are projected into a shared latent space using a pre-trained motion encoder, same with text-to-motion generation research (Guo et al., 2022). We comprehensively ablate the effects of the alignment objective $\mathcal{L}_{\text{align}}$, channel combinations, and training methods.

The results reveal that removing $\mathcal{L}_{\text{align}}$ leads to a severe degradation in retrieval performance, despite achieving the lowest FID score. This phenomenon highlights a trade-off between synthesis realism and conditional fidelity; without $\mathcal{L}_{\text{align}}$, the motion decoder can still synthesize plausible human movements, but it loses conditional dependency on EEG, validating that $\mathcal{L}_{\text{align}}$ acts as a crucial information bottleneck to constrain semantic extraction.

Computing $\mathcal{L}_{\text{align}}$ via video features, motion features, or utilizing visual-only channels yields comparable retrieval accuracy. However, combining all channels with motion feature alignment produces a substantially lower FID and higher diversity. This indicates that while visuospatial information provides adequate cross-modal matching cues, achieving realistic and diverse motion synthesis ultimately requires the integration of sensorimotor representations.

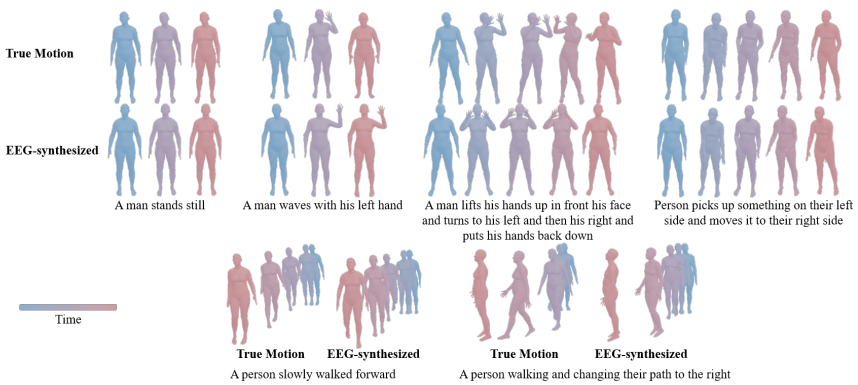

Figure 5: Examples of human motions synthesized from EEG.

Regarding training strategies, LoRA fine-tuning and training from scratch achieve comparable performance, whereas a naive knowledge distillation baseline underperforms, suggesting that cross-modal distillation from established text-to-motion priors demands specialized adaptation frameworks for neural signals.

We also evaluate the retrieval performance between the global EEG representations and the target modality (Table 3). Intriguingly, joint optimization with the downstream motion decoder significantly enhances the retrieval accuracy of the global EEG representation, outperforming all results from the isolated contrastive learning phase. This indicates a cross-level reciprocal optimization: fine-grained motion reconstruction constraints provide complementary low-level details that enrich high-level neural semantic capture, while motion reconstruction concurrently benefits from tighter cross-modal alignment.

Finally, Fig. 5 qualitative visualizes successful retrieval exemplars showcasing the ground-truth and EEG-synthesized motions. Our framework successfully generates realistic, diverse, and semantically aligned human motions guided by non-invasive neural signals. Additional qualitative results are detailed in the Supplementary Material.

7 Discussion and Limitations

While EMMM achieves the first full-body human motion synthesis from non-invasive EEG, several limitations remain. First, our Action Observation paradigm implies that the retrieval performance is largely driven by visual cortical activations rather than the sensorimotor regions, making the task distinct from conventional motion decoding. Second, although our results significantly outperform randomized baselines, a substantial performance gap exists compared to text-to-motion generation, primarily due to the inherent low signal-to-noise ratio of EEG. Third, our framework prioritizes high-level semantic retrieval over absolute joint trajectories, leading to spatial precision deviations. Finally, our methodological innovation in generative modeling is limited, as we do not introduce a generative approach specifically designed for EEG. Nevertheless, this work represents an initial step toward generative motor BCIs.

8 Conclusion

In this work, we introduced EMMM, a generative framework that successfully synthesizes realistic, full-body human motions from non-invasive EEG signals. Supported by our curated EEG2MOTION dataset, we demonstrated that integrating cross-modal alignment with downstream motion modeling balances generative realism and conditional fidelity. By bridging neural signals with high-dimensional kinematics, this framework establishes a competitive baseline and opens a promising frontier for future generative and open-vocabulary motor BCIs.

AI USE STATEMENT

In this work, we used generative AI tools only for language editing and polishing of the manuscript. We did not use generative AI tools for generating research ideas, conducting experiments, analyzing data, generating scientific claims, or producing research artifacts. We have reviewed all AI-assisted text and verified its correctness and consistency with our original work. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

All human participants in this study provided informed consent prior to participation. The study protocol was reviewed and approved by the relevant institutional ethics committee.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, all code used for the experiments is provided in the supplementary materials. We additionally provide execution logs documenting the main experimental results, allowing the reported results to be independently checked and reproduced.

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural computing and applications*, 35(20):14681–14722, 2023.

Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*, 2023.

Xupeng Chen, Ran Wang, Amirhossein Khalilian-Gourtani, Leyao Yu, Patricia Dugan, Daniel Friedman, Werner Doyle, Orrin Devinsky, Yao Wang, and Adeen Flinker. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence*, 6(4): 467–480, 2024.

James S Frank and Marie Earl. Coordination of posture and movement. *Phys Ther*, 70(12):855–863, 1990.

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*, pp. 5152–5161, 2022.

Robert M Hardwick, Svenja Caspers, Simon B Eickhoff, and Stephan P Swinnen. Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. *Neuroscience & Biobehavioral Reviews*, 94:31–44, 2018.

Jieshi He, Puzhe Li, Yanan Sui, and Mu-ming Poo. Neural-behavioral representation of natural whole-body movement in monkeys. *arXiv preprint arXiv:2605.29355*, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2151–2162, 2023.

Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1809–1817, 2017.

Johannes Y Lee, Sangjoon Lee, Abhishek Mishra, Xu Yan, Brandon McMahan, Brent Gaisford, Charles Kobashigawa, Mike Qu, Chang Xie, and Jonathan C Kao. Brain–computer interface control with artificial intelligence copilots. *Nature machine intelligence*, 7(9):1510–1523, 2025.

Kaylo T Littlejohn, Cheol Jun Cho, Jessie R Liu, Alexander B Silva, Bohan Yu, Vanessa R Anderson, Cady M Kurtz-Miott, Samantha Brosler, Anshul P Kashyap, Irina P Hallinan, et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*, 28(4): 902–912, 2025.

Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024a.

Xuan-Hao Liu, Yan-Kai Liu, Yansen Wang, Kan Ren, Hanwen Shi, Zilong Wang, Dongsheng Li, Bao-Liang Lu, and Wei-Long Zheng. Eeg2video: Towards decoding dynamic visual perception from eeg signals. *Advances in Neural Information Processing Systems*, 37:72245–72273, 2024b.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: bidirectional autoregressive motion model. In *European Conference on Computer Vision*, pp. 172–190. Springer, 2024a.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1546–1555, 2024b.

Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9955–9965, 2025.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Giacomo Rizzolatti and Giuseppe Luppino. The cortical motor system. *Neuron*, 31(6):889–901, 2001.

Reza Shadmehr and John W Krakauer. A computational neuroanatomy for motor control. *Experimental brain research*, 185(3):359–381, 2008.

Shreya Shukla, Jose Torres, Akshaj Murhekar, Christina Liu, Abhijit Mishra, Jacek Gwizdka, and Shounak Roychowdhury. A survey on bridging eeg signals and generative ai: from image and text to beyond. *arXiv preprint arXiv:2502.12048*, 2025.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit Bermano, and Michiel Van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations*, volume 2025, pp. 46506–46520, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Deng Wang, Duoqian Miao, and Gunnar Blohm. Multi-class motor imagery eeg decoding for brain-computer interfaces. *Frontiers in neuroscience*, 6:151, 2012.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

Jonathan R Wolpaw. Brain-computer interfaces (bcis) for communication and control. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pp. 1–2, 2007.

Jonathan R Wolpaw, Dennis J McFarland, Gregory W Neat, and Catherine A Forneris. An eeg-based brain-computer interface for cursor control. *Electroencephalography and clinical neurophysiology*, 78(3):252–259, 1991.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.