LEARNING RESIDUAL KINEMATIC CORRECTIONS FOR CONTINUOUS NEURAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Directly improving continuous three-dimensional kinematic decoding from non-invasive EEG remains challenging because of its low signal-to-noise ratio, variability, and non-stationarity. Deep learning architectures such as convolutional neural network–long short-term memory (CNN–LSTM) models can capture spatial and temporal dynamics for continuous kinematic decoding; however, systematic residual errors persist in decoded trajectories. We propose a two-stage decoding framework that applies a residual error learning model to perform kinematic correction on the outputs of a CNN–LSTM decoder. The residual model is trained offline without direct EEG input and instead operates on predicted kinematic trajectories to optimize movement accuracy relative to target trajectories. This design enables targeted correction of systematic decoder errors while preserving the primary neural decoding pipeline. The proposed framework was evaluated offline using data from ten participants across ten sessions of online continuous 3D motor imagery, with feedback provided in both 2D and immersive virtual reality (VR) environments. Decoding performance was quantified using Pearson correlation coefficients (r) and root mean square error (RMSE) along the x, y, and z axes. Compared with the CNN–LSTM applied alone, the CNN–LSTM with residual error correction improved the mean correlation from 0.5076 to 0.7182 ($p < 0.05$) in 2D and from 0.6420 to 0.7783 ($p < 0.05$) in VR, corresponding to relative improvements of 41.5% and 21.2%, respectively. Correspondingly, RMSE was reduced from 0.0564 to 0.0380 in 2D ($p < 0.05$) and from 0.0528 to 0.0352 in VR ($p < 0.05$), representing relative reductions of 32.6% and 33.3%, respectively. These findings demonstrate that this scalable framework enhances continuous 3D MI decoding in BCIs by correcting kinematic errors using an offline residual model without requiring additional neural data, thereby advancing applications in neurorehabilitation, prosthetics, and virtual interaction.

1 INTRODUCTION

Brain-computer interfaces (BCI) have demonstrated remarkable success across a wide range of domains, including cognitive and physical rehabilitation (Millán et al., 2010; Prasad et al., 2010; Cervera et al., 2018), augmentative communication (Willett et al., 2021; 2023; Anumanchipalli et al., 2019), assessing consciousness (Du Bois et al., 2026), entertainment (Marshall et al., 2013; van de Laar et al., 2013), and robotic control (Hochberg et al., 2012; Ajiboye et al., 2017; Edelman et al., 2019), by establishing a direct signaling pathway between neural activity and external devices. In electroencephalography (EEG)-based BCIs, motor imagery (MI) has traditionally been decoded as a discrete classification task, assigning EEG segments to a small set of imagined movements, such as left- or right-hand movement Lotte et al. (2018). Although this approach is well established and supports simple command selection, it suffers from bottlenecks like rigid control dynamics. Continuous MI decoding instead estimates time-varying three-dimensional (3D) kinematics from ongoing EEG, representing movement as a trajectory rather than a sequence of categorical commands.

Artificial intelligence (AI) techniques have proven highly effective in decoding continuous 3D limb motion trajectories and motion imagery from EEG signals (Bradberry et al., 2010; Chen et al., 2022; Wang et al., 2023; Xu et al., 2021; Korik et al., 2025). Recent deep learning approaches, specifically, hybrid convolutional neural network and long short-term memory (CNN–LSTM) architectures

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.

(Mc Shane et al., 2023; McShane et al., 2026), have demonstrated superior decoding accuracy compared to multiple linear regression models (Mc Shane et al., 2022).

Beyond improving the neural decoder itself, another strategy is to correct errors after decoding. Existing error-aware BCI studies have primarily used error-related potentials (ErrPs), transient EEG responses elicited when a user detects a mismatch between the intended and observed outcome of an action, to identify whether an incorrect action has occurred (Spüler & Niethammer, 2015; Lopes-Dias et al., 2019; Mondini et al., 2024; Demchenko et al., 2025). Although these approaches can support self-correcting interaction, they generally formulate error processing as the classification of error occurrence or type rather than the estimation of the direction and magnitude of a continuous kinematic error. Pulferer et al. (2024) moved towards continuous error regression by estimating target–feedback discrepancies from error-related EEG activity and using them to improve trajectory feedback. However, their correction process still requires an additional neural signal reflecting the user's perception of the error and was evaluated only in a simple 2D tracking task.

Residual learning offers a suitable but largely unexplored formulation for EEG-based BCI decoding by preserving a baseline prediction and training a secondary model to estimate the remaining prediction error. This principle has been applied to vehicle modelling, where deep learning based components correct the predictions of physical baseline models (Jiang et al., 2021; Long et al., 2025; Miao et al., 2025). Residual reinforcement learning (Silver et al., 2018; Johannink et al., 2019; Alakuijala et al., 2021; Davchev et al., 2022) adds a learned correction to a baseline controller, but it learns reward-optimised action residuals rather than directly modelling the prediction errors of a frozen decoder.

Directly improving continuous three-dimensional kinematic decoding from non-invasive EEG remains challenging because of its low signal-to-noise ratio, variability, and non-stationarity. Although existing residual-learning approaches similarly combine a baseline prediction with an additive correction, their residual components are not fully decoupled from the information used to construct the baseline prediction. In the BCI setting, such dependence may be undesirable because a correction model that directly processes EEG may therefore inherit many of the difficulties faced by the original neural decoder. To address this problem, we extend the work of McShane et al. (2026) and adopt a different perspective: rather than extracting additional corrective information from EEG, we treat the frozen decoder output as a kinematic sequence and learn its residual errors directly in the output space. Specifically, we propose a two-stage residual learning framework in which a participant-specific CNN–LSTM decoder is first trained and frozen to generate baseline three-dimensional kinematic predictions from EEG. A separate residual model is then trained to predict an additive 3D correction using only the current decoder output, normalised temporal information, and the previous corrected output.

Our key contributions are as follows: (1) we introduce a novel two-stage architecture that preserves a frozen CNN–LSTM decoder for spatiotemporal EEG decoding while enabling the independent residual model to be trained offline without live user interaction; and (2) we enhance decoding robustness and reduce errors through residual kinematic correction that requires neither EEG representations nor target information during inference. By decoupling residual correction from the noisy EEG input, the proposed method achieves relative improvements in mean correlation of up to 41.5% and reductions in mean RMSE of up to 35.7%.

2 Methodology

2.1 Dataset and Baseline BCI Decoder

Ten right-handed adults completed ten sessions on different days, alternating between five 2D screen and five immersive VR real-time feedback sessions. For clarity, sessions within each feedback modality are indexed separately as Session 1 - Session 5. Across sessions, participants engaged in an upper-limb reaching task to 4 targets. Each session comprised eight runs (four executed and four imagined), with two blocks of 16 trials per run. Within each block, four targets (shown in Appendix A) were presented in random order, each appearing four times. This design produces 256 trials per session, 2560 trials per participant across ten sessions, and 25600 trials across the complete dataset. Although the protocol included both executed and imagined runs, this study focuses exclusively on imagined movements; therefore, only imagined-movement runs were used for model

training, validation, and testing. Executed-movement trials were used to construct the reference targets. Specifically, for each participant and session, the executed trials were averaged separately for each of the four target directions, yielding four participant- and session-specific mean trajectories. These averaged executed trajectories served as the references for the corresponding imagined trials. They were used as supervised targets and, during real-time data collection, to generate the predefined trajectory assistance, but were never provided to either model as inference inputs.

The experimental sequence followed a trial-based structure, where each block was self-initiated and preceded by a 30-second countdown. Individual trials, spanning 7.6 s, consisted of four sequential functional stages: a 2.5 s rest phase for baseline fixation and neural stability, followed by a 1.6 s Indication Phase where a directional pre-cue line pointed toward the 3D target. The subsequent 2.5 s target phase was initiated by a 440 Hz auditory tone (64 ms) and featured a 1.5× target expansion alongside animated trajectory guidance to facilitate motor imagery. Each trial concluded with a 1.0 s Reset Phase, signaled by a secondary auditory cue, which prompted the return of the virtual hand to its initial starting position. The trial flowchart is provided in Appendix A. The movement-onset segment was extracted solely from the runs within the target phase to train both the BCI decoder and the residual correction model. The segments were temporally aligned to the trajectory from the resting position toward the target hold, excluding the return movement during the reset phase, and maintained a fixed duration of 2 s across all trials. Further details of the gaze protocol and the controls used to minimise and assess potential ocular artefacts are provided in Appendix B.

EEG neural activity was captured using a wireless g.Nautilus system (g.tec) with 32 g.SCARABEO active electrodes. While a comprehensive 32-channel montage was recorded (including C3, C4, AF3, T7, etc.), online decoding and decoder training were streamlined using a targeted 17-channel subset (F3, FZ, F4, FC5, FC1, FC2, FC6, C3, CZ, C4, CP5, CP1, CP2, CP6, P3, PZ, and P4) centered over the sensorimotor cortex. Concurrently, kinematic trajectories were sampled at 60 Hz via a wrist-mounted Vive tracker. This system provided continuous three-dimensional Cartesian coordinates (x, y, z), which were subsequently subjected to numerical differentiation to derive instantaneous velocity estimates for each time step.

The BCI decoder was developed with an CNN–LSTM architecture based on Event-Related Spectral Perturbation (ERSP) (Makeig, 1993) to decode continuous 3D limb velocity from EEG signals. ERSP provided time–frequency representations of motor-related neural activity by quantifying spectral power changes relative to a baseline. For each of the 17 EEG channels, 40×40 ERSP representations spanning 0–40 Hz were generated from 640 ms before movement onset through execution at 16 ms intervals, then stacked across channels as the model input. The CNN component, consisting of three convolutional layers with 3×3 kernels and 32 filters in the first layer, with max-pooling and dropout to reduce overfitting, learned spatial features from these ERSP representations. Extracted features were then fed into two stacked LSTM layers, with 50 hidden units in the first LSTM layer and a dropout rate of 0.25 applied between recurrent layers, to capture temporal dependencies across sequential windows and maintain contextual information. The network produced continuous predictions of x-, y-, and z-axis velocities through a dense output layer with tanh activations in hidden layers. The decoder architecture is identical to that described in McShane et al. (2026).

2.2 Residual Learning Formulation

Let $\hat{\mathbf{v}}_{i,t}^{\mathrm{BCI}} \in \mathbb{R}^3$ denote the velocity produced by the frozen CNN–LSTM decoder for trial i at time step t, and let $\mathbf{v}_{i,t}^{*} \in \mathbb{R}^3$ denote the corresponding target velocity. The residual model does not replace the baseline decoder. It learns a correction in the same three-dimensional kinematic space and adds this correction to the baseline output.

For a trajectory of length T, the input to the residual model is

$$\mathbf{x}_{i,t} = \left[\hat{\mathbf{v}}_{i,t}^{\mathrm{BCI}},\ \tau_t,\ \tilde{\mathbf{v}}_{i,t-1}\right] \in \mathbb{R}^7, \qquad \tau_t = \frac{t}{T}, \tag{1}$$

where τ_t specifies the normalised time of the cued movement and $\tilde{\mathbf{v}}_{i,t-1}$ is the corrected velocity from the preceding time step. At the beginning of a trajectory, the recurrent context is initialized as $\tilde{\mathbf{v}}_{i,-1} = \hat{\mathbf{v}}_{i,0}^{\mathrm{BCI}}$. A residual network $g_{\boldsymbol{\theta}}$ predicts a bounded correction

$$\Delta\mathbf{v}_{i,t} = \mathbf{s} \odot \tanh(g_{\boldsymbol{\theta}}(\mathbf{x}_{i,t}))\,, \qquad \mathbf{s} = [s_x, s_y, s_z]^{\top}, \tag{2}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{s}$ contains dimension-specific residual scales. The final corrected output is

$$\tilde{\mathbf{v}}_{i,t} = \hat{\mathbf{v}}_{i,t}^{\mathrm{BCI}} + \Delta\mathbf{v}_{i,t}. \tag{3}$$

The model parameters are optimized by minimizing the point-wise mean squared error over N trajectories between corrected and target velocities:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{3NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left\| \tilde{\mathbf{v}}_{i,t} - \mathbf{v}_{i,t}^{*} \right\|_2^2 . \tag{4}$$

Because the baseline decoder is fixed, this objective is equivalently

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{3NT} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \left\| \Delta\mathbf{v}_{i,t} - \left(\mathbf{v}_{i,t}^{*} - \hat{\mathbf{v}}_{i,t}^{\mathrm{BCI}}\right) \right\|_2^2 , \tag{5}$$

which makes explicit that the network learns the residual error left by the baseline decoder. Target velocities are used only to calculate the training loss and are never included in the model input.

2.3 Two-Stage Framework

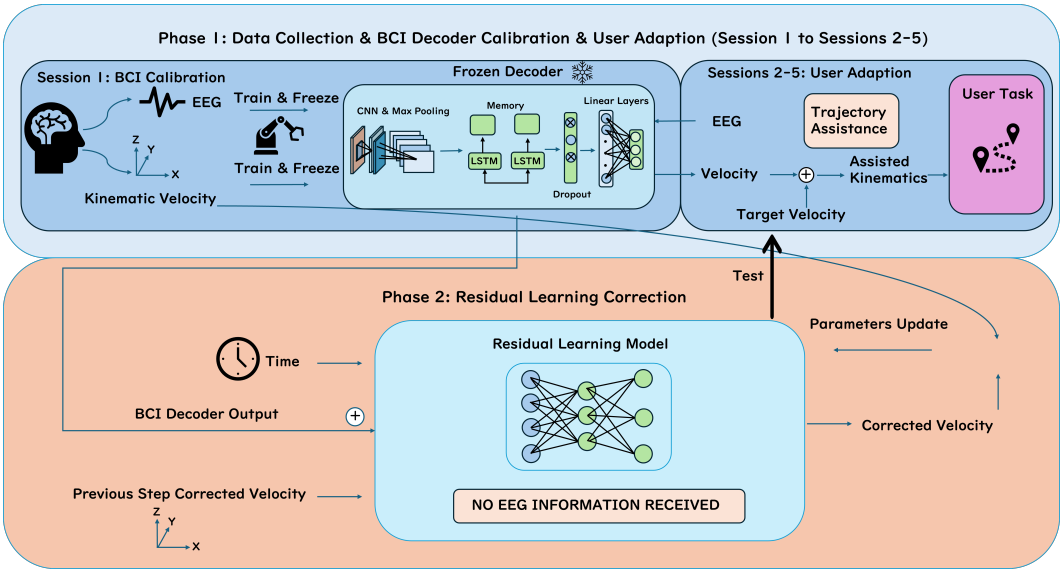


Figure 1: The architecture of the proposed two-stage framework.

The proposed architecture consists of two stages, shown in Figure 1. In the first stage, for each participant and feedback modality, a CNN–LSTM decoder was trained offline using EEG data from the corresponding Session 1 (S1), with separate models for the 2D and VR feedback. After calibration, the decoder was frozen and deployed online throughout Sessions 2–5 (S2–S5), generating kinematic predictions from incoming EEG and providing immediate visual feedback to the participant. Here, "online" refers to real-time interaction and feedback rather than online model learning or parameter updating. During the imagined-movement runs, trajectory assistance was applied to support user adaptation through positively biased feedback by blending the real-time decoded trajectory with the corresponding predefined target trajectory. In S1, 100% trajectory assistance was used because the decoder had not yet been trained and therefore no decoder-based feedback was provided. The assistance level was set to 65% in S2 and progressively reduced to 60%, 50%, and 40% in S3, S4, and S5, respectively. The percentage indicates how strongly the decoded trajectory was shifted towards the target trajectory. Importantly, trajectory assistance affected only the visual feedback presented to the participant; it did not modify the underlying decoder predictions or update the decoder parameters. Thus, the fixed decoder provided a stable EEG-to-kinematics mapping across sessions, while the assisted kinematics drove the visual effector during online interaction. In the second stage, a separate

residual correction model was trained subsequently and entirely offline using the Session 1 decoder outputs. The residual model was not deployed during online data collection and was subsequently evaluated offline, with its parameters frozen, on the decoder outputs collected from Sessions 2–5. Separate residual models were trained for each participant and feedback modality.

3 EXPERIMENTS

3.1 HYPERPARAMETERS OPTIMISATION AND DATA PROCESSING

To ensure a fair comparison, the hyperparameters of both the baseline CNN–LSTM decoder and the residual correction network were tuned, with an identical number of 200 search iterations allocated to each model. For the residual correction network, the search objective was to maximise the mean correlation between the corrected and target trajectories. The residual learning correction networks are implemented using lightweight fully connected layers. The tuned parameters, which include learning rate, hidden layer dimension, batch size, axis-specific residual correction scales, number of network layers, and their range are shown in Appendix C. We used 5-fold cross-validation, and hyperparameters were selected based on the mean validation performance across the five folds; once the best parameters were selected, the model was retrained on the combined training and validation sets and assessed against the four completely unseen test sessions (S2-S5), ensuring no data leakage issues in the evaluation. The data (e.g. BCI decoder outputs) were normalised to [0, 0.3] using axis-wise global min–max scaling, with separate minimum and maximum values calculated from all training samples for each axis. These training-derived parameters were applied unchanged to the corresponding axes of the test data during inference, so no test data informed normalisation. Computational resources used are described in detail in Appendix F.

3.2 EVALUATION METRIC

The correction accuracy is evaluated using Pearson correlation coefficient and root mean square error (RMSE). For the i-th trial and spatial dimension d, the Pearson correlation coefficient ($r_{i,d}$) between the decoded trajectory and the target trajectory is defined as

$$r_{i,d} = \frac{\sum_{t=1}^{T_i}(X_{t,d,i} - \bar{X}_{d,i})(Y_{t,d,i} - \bar{Y}_{d,i})}{\sqrt{\sum_{t=1}^{T_i}(X_{t,d,i} - \bar{X}_{d,i})^2}\sqrt{\sum_{t=1}^{T_i}(Y_{t,d,i} - \bar{Y}_{d,i})^2} + \epsilon}, \tag{6}$$

where $X_{t,d,i}$ and $Y_{t,d,i}$ represent the decoded and target points, respectively, while $\bar{X}_{d,i}$ and $\bar{Y}_{d,i}$ denote their temporal means, t indexes time steps, T_i is the number of samples in trial i, and ϵ is a small constant added to avoid division by zero. The individual correlations in Equation 6 is averaged across trials to compute $\bar{r}_d = \frac{1}{N}\sum_{i=1}^{N} r_{i,d}$, and then across spatial dimensions to compute $r = \frac{1}{D}\sum_{d=1}^{D} \bar{r}_d$, where N is the total number of trials, D is the number of spatial dimensions, i indexes trials ($i = 1, \ldots, N$), and d indexes spatial dimensions ($d = 1, \ldots, D$).

For the i-th trial and spatial dimension d, RMSE between the decoded trajectory and the target trajectory is defined as:

$$\text{RMSE}_{i,d} = \sqrt{\frac{1}{T_i}\sum_{t=1}^{T_i}(X_{t,d,i} - Y_{t,d,i})^2}. \tag{7}$$

This RMSE is then averaged across trials and then across spatial dimensions as $\overline{\text{RMSE}}_d = \frac{1}{N}\sum_{i=1}^{N}\text{RMSE}_{i,d}$ and $\overline{\text{RMSE}}_{\text{overall}} = \frac{1}{D}\sum_{d=1}^{D}\overline{\text{RMSE}}_d$.

3.3 EVALUATING STRATEGIES

Fixed Decoder Generalisation (FDG), Sequential Adaptive Training (SAT), and Within-Session Reconstruction (WSR) are three distinct training and evaluation strategies applied to the standard supervised CNN–LSTM decoder, following McShane et al. (2026) exactly; they are not variants or modifications of the proposed residual correction method. FDG represents true online BCI performance. The decoder was trained offline using Session 1 data and its parameters were then

frozen during online deployment in Sessions 2–5, as described in Section 2.3. Performance variations under FDG therefore reflect user learning and adaptation. SAT is an offline evaluation strategy in which the decoder is trained using one session and tested on the subsequent session, for example, S1→S2 and S2→S3. SAT therefore captures the combined effects of participant adaptation and periodic decoder recalibration. WSR trains and evaluates the decoder within the same session and provides an upper-bound estimate of session-specific decoder capability. Analogous to FDG, Fixed Residual Learning (FRL) trains the residual correction model using only Session 1 data, freezes its parameters, and evaluates it across Sessions 2–5. The residual model was neither deployed during real-time BCI interaction nor used to provide feedback to participants. Instead, it was only trained and evaluated using the decoder outputs collected during the corresponding sessions, as described in Section 2.3.

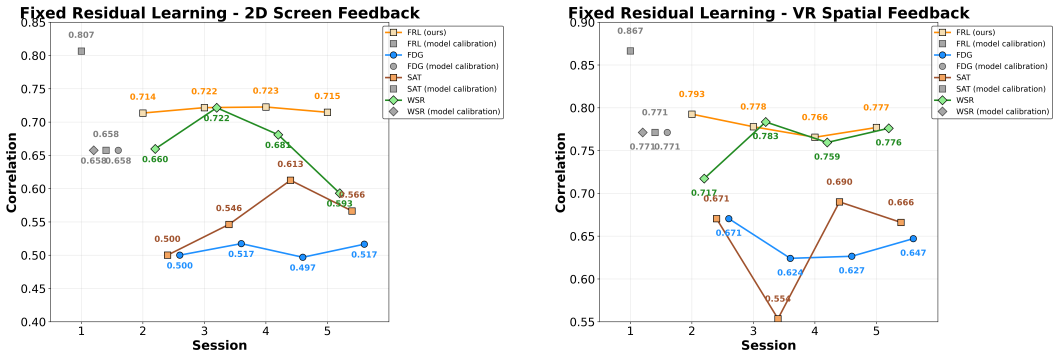
3.4 Statistical Analysis

For group-level analyses of Pearson correlation and RMSE, performance was averaged within each of the 10 participants across trials, spatial axes, and test sessions S2–S5, yielding one value per strategy, metric, and feedback condition. The 2D and VR conditions were analysed separately; no feedback-by-strategy interaction was tested. One-way repeated-measures ANOVA (Girden, 1992) was used for correlation in both conditions and for VR RMSE, whereas the Friedman test (Friedman, 1937) was used for 2D RMSE. Significant omnibus effects were followed by planned comparisons between FRL and each alternative strategy using paired t-tests, except for 2D RMSE, for which Wilcoxon signed-rank tests (Wilcoxon, 1992) were used. Statistical significance was assessed using an alpha level of $\alpha = 0.05$. Cohen's d (Cohen, 2013) and the standardized Wilcoxon effect size r_z were used to quantify the parametric and non-parametric pairwise effects, respectively. To account for the three planned pairwise comparisons within each metric and feedback condition, the Holm–Bonferroni correction (Holm, 1979) was applied to the raw p-values. The prerequisite analyses used to determine the final selection of parametric or non-parametric tests are detailed in Appendix D. Visualisations of corrections for some plots are shown in Appendix E

4 Results

We compare FRL with FDG, SAT, and WSR under both 2D and VR feedback, reporting Pearson correlation followed by RMSE. The figures show performance across sessions, while the overall comparisons focus on the four test sessions (S2–S5).

4.1 Pearson Correlation Coefficient



(a) 2D feedback

(b) VR feedback

Figure 2: Pearson correlation coefficient of FRL comparison with three training strategies of BCI decoder in both 2D and VR feedback across sessions

Figure 2 compares the trajectory correlation across five sessions using the proposed method FRL and three BCI decoder strategies (FDG, SAT, and WSR) under both 2D screen feedback and VR feedback. Each point represents the mean correlation per session; grey markers indicate performance when

Table 1: Comparison of Overall Pearson Correlation Coefficient: FRL vs. BCI Decoder Strategies under 2D and VR Feedback

2D Feedback

Strategy	Correlation Mean (SD)				Δ	t-test		$\lvert d\rvert$	ANOVA	
	X	Y	Z	Overall		$\lvert t(9)\rvert$	p		F	p
FDG	0.1209 (0.6885)	0.7469 (0.3208)	0.6551 (0.3678)	0.5076 (0.3206)	+0.2106	5.6503	0.0006	1.7868	21.2653	< 0.0001
SAT	0.1647 (0.6945)	0.7894 (0.2719)	0.7149 (0.3180)	0.5563 (0.3029)	+0.1619	6.9975	0.0002	2.2128		
WSR	0.3845 (0.6765)	0.8261 (0.2583)	0.7809 (0.2818)	0.6638 (0.2984)	+0.0544	1.4739	0.1746	0.4661		
FRL	**0.2865 (0.6264)**	**0.9829 (0.0180)**	**0.8850 (0.1031)**	**0.7182 (0.2167)**	—	—	—	—		

VR Feedback

Strategy	Correlation Mean (SD)				Δ	t-test		$\lvert d\rvert$	ANOVA	
	X	Y	Z	Overall		$\lvert t(9)\rvert$	p		F	p
FDG	0.2492 (0.6854)	0.8455 (0.2243)	0.8316 (0.2366)	0.6420 (0.2870)	+0.1363	8.6477	<0.0001	2.7346	48.2159	< 0.0001
SAT	0.3019 (0.6727)	0.8245 (0.2647)	0.8090 (0.2689)	0.6451 (0.2977)	+0.1332	8.8209	<0.0001	2.7894		
WSR	0.5297 (0.6149)	0.8797 (0.1829)	0.8679 (0.1930)	0.7591 (0.2557)	+0.0192	2.4447	0.0370	0.7731		
FRL	**0.3819 (0.6595)**	**0.9786 (0.0362)**	**0.9744 (0.0419)**	**0.7783 (0.2218)**	—	—	—	—		

* Note: Δ denotes the overall improvement of FRL, negative values indicate an RMSE reduction; p denotes Holm–Bonferroni-adjusted p-values. Pairwise p-values are Holm–Bonferroni adjusted. F denotes the statistic from the one-way repeated-measures ANOVA. Q denotes the Friedman test statistic. $|t(9)|$ denotes absolute t-statistic; W is the Wilcoxon statistic; $|d|$ and r_z indicate Cohen's d and standardized Wilcoxon effect size, respectively.

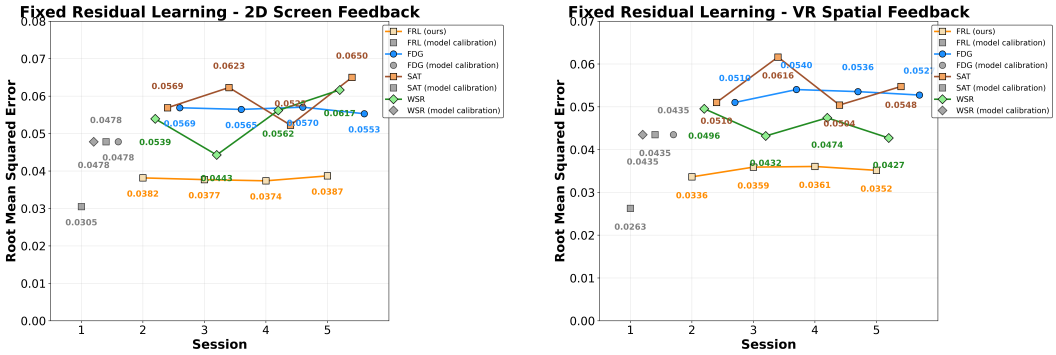
evaluated on the same session used for training (for FRL, FDG, and SAT). Under 2D feedback, FRL consistently achieves the highest correlation across the test sessions, ranging from 0.714 to 0.723, with stable performance. This represents a substantial improvement over FDG (relative improvement of 38.3–45.5%) and a clear advantage over SAT (17.9–42.8%). FRL matches WSR in Session 3 (0.722 for both strategies) and outperforms it in the other three test sessions, with relative improvements ranging from 6.2% to 20.6%. Under VR spatial feedback, FRL achieves correlations ranging from 0.766 to 0.793 across the test sessions, corresponding to improvements of 18.2–24.7% over FDG and 11.0–40.4% over SAT. WSR performs closest to FRL and slightly outperforms it in Session 3 (0.783 versus 0.778), whereas FRL maintains an advantage of approximately 0.1–10.6% across the remaining sessions.

Table 1 illustrates that FRL achieves higher overall trajectory correlation than all compared strategies across both 2D and VR feedback conditions. Under 2D feedback, one-way repeated-measures ANOVA confirmed a significant main effect of decoder strategy. FRL significantly outperforms FDG, with a relative improvement of 41.5%. A similar significant improvement is observed over SAT (29.1%), whereas the relative gain over WSR (8.2%) does not reach statistical significance. Under VR feedback, one-way repeated-measures ANOVA also revealed a significant main effect of decoder strategy. FRL significantly improves correlation over FDG (+ 21.2%) and over SAT (+ 20.6%). The smaller improvement over WSR (2.5%) is also statistically significant after Holm correction ($p = 0.0370$).

4.2 Root Mean square Error

Figure 3 illustrates the RMSE across five sessions for FRL and three BCI decoder strategies under both 2D and VR feedback. Session 1 shows same-session model-calibration performance, whereas the quantitative comparisons below focus on the four test sessions. Lower values indicate better performance. Under 2D feedback, FRL consistently achieves the lowest RMSE (0.0374–0.0387), outperforming FDG (0.0553–0.0570, relative reduction of 30.0–34.4%), SAT (0.0523–0.0650, 28.4–40.5%), and WSR (0.0443–0.0617, 14.9–37.3%). Under VR feedback (Fig. 3b), FRL also consistently achieves the lowest RMSE (0.0336–0.0361), improving over FDG (0.0510–0.0540, 32.6–34.1%), SAT (0.0504–0.0616, 28.4–41.7%), and WSR (0.0427–0.0496, 16.9–32.3%).

As shown in Table 2, under 2D feedback, FRL significantly reduces RMSE compared with FDG, decreasing the error from 0.0564 to 0.0380, corresponding to a relative reduction of 32.6%. Similar significant reductions are observed relative to SAT (35.7%) and WSR (29.6%). The Friedman test confirmed a significant main effect of decoder strategy, and the Holm-adjusted Wilcoxon signed-rank comparisons were significant for all three comparisons. Under VR feedback, FRL significantly reduces RMSE relative to FDG from 0.0528 to 0.0352 (33.3%), with corresponding reductions of



(a) 2D feedback

(b) VR feedback

Figure 3: Root mean square error comparison of FRL with three training strategies of BCI decoder in both 2D and VR feedback across sessions

Table 2: Comparison of Overall Root Mean Square Error: FRL vs. BCI Decoder Strategies under 2D and VR Feedback

2D Feedback

Strategy	RMSE Mean (SD)				Δ	Wilcoxon		r_z	Friedman	
	X	Y	Z	Overall		W	p		Q	p
FDG	0.0557 (0.0290)	0.0581 (0.0229)	0.0555 (0.0233)	0.0564 (0.0158)	-0.0184	0.0	0.005859	0.8864		
SAT	0.0596 (0.0318)	0.0597 (0.0271)	0.0580 (0.0250)	0.0591 (0.0191)	-0.0211	0.0	0.005859	0.8864	23.1600	< 0.0001
WSR	0.0506 (0.0364)	0.0572 (0.0387)	0.0543 (0.0355)	0.0540 (0.0312)	-0.0160	0.0	0.005859	0.8864		
FRL	**0.0497 (0.0255)**	**0.0325 (0.0164)**	**0.0318 (0.0167)**	**0.0380 (0.0101)**	—	—	—	—		

VR Feedback

Strategy	RMSE Mean (SD)				Δ	t-test		$\lvert d\rvert$	ANOVA	
	X	Y	Z	Overall		$\lvert t(9)\rvert$	p		F	p
FDG	0.0533 (0.0318)	0.0550 (0.0241)	0.0502 (0.0228)	0.0528 (0.0171)	-0.0176	7.8084	< 0.0001	2.4692		
SAT	0.0546 (0.0341)	0.0574 (0.0257)	0.0515 (0.0239)	0.0545 (0.0197)	-0.0193	9.4806	< 0.0001	2.9980	31.1111	< 0.0001
WSR	0.0431 (0.0334)	0.0495 (0.0245)	0.0446 (0.0239)	0.0457 (0.0201)	-0.0105	4.6189	0.0013	1.4606		
FRL	**0.0491 (0.0295)**	**0.0316 (0.0178)**	**0.0248 (0.0145)**	**0.0352 (0.0133)**	—	—	—	—		

* Note: Δ denotes the overall improvement of FRL, negative values indicate an RMSE reduction; p denotes Holm–Bonferroni-adjusted p-values. Pairwise p-values are Holm–Bonferroni adjusted. F denotes the statistic from the one-way repeated-measures ANOVA. Q denotes the Friedman test statistic. $|t(9)|$ denotes absolute t-statistic; W is the Wilcoxon statistic; $|d|$ and r_z indicate Cohen's d and standardized Wilcoxon effect size, respectively.

35.4% relative to SAT and 23.0% relative to WSR. Repeated-measures ANOVA and subsequent Holm-adjusted paired t-tests confirmed significant differences between FRL and all three comparison strategies.

5 Discussion

5.1 Advantages of Residual Correction and Two-Level Optimisation

The two-stage framework proposed in this paper significantly and substantially improves BCI decoding performance through an offline residual learning correction approach (Tables 1and 2). The primary advantage of this residual architecture is its ability to compensate for systematic errors remaining in fixed decoders, including errors associated with decoding 3D motor imagery from non-stationary EEG and variations introduced by user adaptation across 2D and VR feedback sessions. By design, the correction process operates independently of direct EEG input and focuses exclusively on optimizing the kinematic output of the baseline decoder. Furthermore, the framework's decoupled design separates data collection and decoder calibration from offline residual-model training, without requiring online parameter updates or continuous user interaction. Central to the system's performance is a two-level optimisation strategy that combines global hyperparameter selection with a point-wise

kinematic regression objective. Optuna (Akiba et al., 2019) is used to maximize the validation Pearson correlation coefficient over 200 search trials, encouraging the corrected and target trajectories to exhibit similar temporal morphology. Simultaneously, the residual correction network is trained using mean squared error to minimize the point-wise difference between the corrected and target velocities.

The robustness of our approach is supported by statistical analyses across all 10 participants (the dataset includes 10 sessions per participant and is one of the most comprehensive datasets available for 3D motor imagery with embodied VR feedback). Similar to FDG, FRL evaluates a fixed model across sessions and therefore assesses cross-session generalisation. FRL produced statistically significant correlation improvements over both FDG and the periodic recalibration strategy, SAT, under both feedback conditions, with large effect sizes ($|d|$ ranging from 1.7868 to 2.7894). Compared with WSR, the correlation improvement was statistically significant under VR feedback but did not reach significance under 2D feedback. In terms of RMSE, FRL significantly outperformed WSR under both feedback conditions, with a large non-parametric effect size under 2D feedback ($r_z = 0.8864$) and a large parametric effect size under VR feedback ($|d| = 1.4606$).

These findings indicate that residual learning can improve performance beyond the evaluated standalone CNN–LSTM decoder strategies by correcting systematic errors in their kinematic outputs. Furthermore, FRL consistently exhibits a smaller range of performance fluctuations across sessions, in both correlation and RMSE, than other decoding strategies. This reduced variation indicates improved cross-session stability. Notably, under VR feedback, FRL produced significant RMSE reductions with large effect sizes ($|d|$ ranging from 1.4606 to 2.9980). Axis-specific analysis shows that FRL achieves nearly perfect correlations on the Y axis (0.9829 in 2D and 0.9786 in VR) and on the Z-axis under VR feedback (0.9744). The lowest axis-specific RMSE was 0.0248 on the Z-axis under VR feedback. Nevertheless, the comparatively lower and more variable X-axis correlations indicate that this dimension remains more challenging. Across the parametric and non-parametric analyses, the results consistently support the effectiveness of residual correction even when baseline decoder performance is relatively strong. Given that previous CNN–LSTM architectures (Mc Shane et al., 2023) (McShane et al., 2026) represented the existing performance benchmark for EEG-based 3D motor imagery, the present results establish a strong new benchmark.

5.2 Limitations and Future Work

Because the residual framework applies bounded corrections around the baseline output, its overall effectiveness depends on the quality of the initial decoder. The framework also relies on explicit temporal information, which is available in the present cued, synchronous paradigm. Unlike synchronous BCIs, which operate within predefined cue-aligned time windows, asynchronous BCIs require continuous, self-paced decoding Nicolas-Alonso & Gomez-Gil (2012). This temporal dependency therefore limits direct application to asynchronous settings. Despite the promising results, the current study is limited to a discrete four-target reaching task. Future work will include a broader range of trajectories, denser target directions, and time-invariant decoding. The residual model is also optimised using a single MSE-based objective, which primarily penalises point-wise errors and does not explicitly account for other trajectory properties such as target reaching or directional alignment. In addition, Pearson correlation should be interpreted cautiously in this synchronous paradigm, as temporally aligned trajectories may produce high correlation. Finally, although FRL cross-session generalisation was evaluated offline, real-time closed-loop validation is still required to determine whether the observed improvements translate to online BCI performance and user adaptation.

6 Conclusions

We propose a two-stage framework for continuous 3D motor imagery decoding from non-invasive EEG. A CNN–LSTM decoder is calibrated and used to collect synchronized EEG signals; an offline residual model then corrects its frozen outputs without additional neural input. The residual network minimises point-wise mean squared error, while Optuna selects hyperparameters by maximising Pearson correlation. Experiments and statistical analyses under 2D and VR feedback show improvements over baseline strategies in both correlation and RMSE. These findings support scalable residual correction for continuous 3D BCI control in virtual interaction, neurorehabilitation, and prosthetics.

AI USE STATEMENT

Generative AI tools were used to support literature searching, grammar correction, language refinement, and figure preparation. All research ideas, methodological decisions, experiments, analyses, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This study received ethical approval from the Institutional Research Ethics Committee of the authors' Institutes in accordance with the Declaration of Helsinki and institutional guidelines for research involving human subjects. Participants provided written informed consent prior to participation and were informed of the study procedures, data recording methods (including EEG, kinematic and eye-tracking data), their right to withdraw at any time without penalty, and the intended research use of anonymised data. All data were anonymised prior to analysis and dissemination. No personally identifiable information is included in this publication.

REPRODUCIBILITY STATEMENT

In the paper, the details of implementation have been included. For our models and dataset, a link to an anonymous downloadable source code can be submitted as supplementary materials with the paper; The dataset description has been included in the paper.

REFERENCES

A Bolu Ajiboye, Francis R Willett, Daniel R Young, William D Memberg, Brian A Murphy, Jonathan P Miller, Benjamin L Walter, Jennifer A Sweet, Harry A Hoyen, Michael W Keith, et al. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. *The Lancet*, 389(10081): 1821–1830, 2017.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Minttu Alakuijala, Gabriel Dulac-Arnold, Julien Mairal, Jean Ponce, and Cordelia Schmid. Residual reinforcement learning from demonstrations. *arXiv preprint arXiv:2106.08050*, 2021.

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

Trent J Bradberry, Rodolphe J Gentili, and José L Contreras-Vidal. Reconstructing three-dimensional hand movements from noninvasive electroencephalographic signals. *Journal of neuroscience*, 30 (9):3432–3437, 2010.

María A Cervera, Surjo R Soekadar, Junichi Ushiba, José del R Millán, Meigen Liu, Niels Birbaumer, and Gangadhar Garipelli. Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis. *Annals of clinical and translational neurology*, 5(5):651–663, 2018.

Yi-Feng Chen, Ruiqi Fu, Junde Wu, Jongbin Song, Rui Ma, Yi-Chuan Jiang, and Mingming Zhang. Continuous bimanual trajectory decoding of coordinated movement from eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6012–6023, 2022.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Todor Davchev, Kevin Sebastian Luck, Michael Burke, Franziska Meier, Stefan Schaal, and Subramanian Ramamoorthy. Residual learning from demonstration: Adapting dmps for contact-rich manipulation. *IEEE Robotics and Automation Letters*, 7(2):4488–4495, 2022.

Igor Demchenko, Tamar Shavit, Miri Benyamini, and Miriam Zacksenhouse. Self-correcting brain computer interface based on classification of multiple error-related potentials. *Journal of Neural Engineering*, 22(2):026028, 2025.

Naomi Du Bois, Attila Korik, Stephanie Hodge, Leah Hudson, Ainjila S Elahi, Alain Bigirimana, Natalie Dayan, Jose M Sanchez-Bornot, Alison McCann, Kudret Yelden, et al. Advancing eeg-based assessment of consciousness and cognition in prolonged disorders of consciousness. *Communications medicine*, 2026.

Bradley J Edelman, Jianjun Meng, Daniel Suma, Claire Zurn, Eric Nagarajan, Bryan S Baxter, Christopher C Cline, and BJSR He. Noninvasive neuroimaging enhances continuous neural tracking for robotic device control. *Science robotics*, 4(31):eaaw6844, 2019.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

Ellen R Girden. *ANOVA: Repeated measures*. Number 84. sage, 1992.

Leigh R Hochberg, Daniel Bacher, Beata Jarosiewicz, Nicolas Y Masse, John D Simeral, Joern Vogel, Sami Haddadin, Jie Liu, Sydney S Cash, Patrick Van Der Smagt, et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature*, 485(7398):372–375, 2012.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Shu Jiang, Weiman Lin, Yu Cao, Yu Wang, Jinghao Miao, and Qi Luo. Learning-based vehicle dynamics residual correction model for autonomous driving simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 782–789. IEEE, 2021.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *2019 international conference on robotics and automation (ICRA)*, pp. 6023–6029. IEEE, 2019.

Attila Korik, Naomi Du Bois, Jose Sanchez Bornot, Niall McShane, Christoph Guger, Alessandra Del Felice, Olive Lennon, and Damien Coyle. Decoding the variable velocity of lower-limb stepping movements from eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2025.

Keke Long, Zihao Sheng, Haotian Shi, Xiaopeng Li, Sikai Chen, and Soyoung Ahn. Physical enhanced residual learning (perl) framework for vehicle trajectory prediction. *Communications in Transportation Research*, 5:100166, 2025.

Catarina Lopes-Dias, Andreea I Sburlea, and Gernot R Müller-Putz. Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Scientific reports*, 9(1):17596, 2019.

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

Scott Makeig. Auditory event-related dynamics of the eeg spectrum and effects of exposure to tones. *Electroencephalography and clinical neurophysiology*, 86(4):283–293, 1993.

David Marshall, Damien Coyle, Shane Wilson, and Michael Callaghan. Games, gameplay, and bci: the state of the art. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2): 82–99, 2013.

Niall Mc Shane, Karl McCreadie, Darryl Charles, Attila Korik, and Damien Coyle. Online 3d motion decoder bci for embodied virtual reality upper limb control: A pilot study. In *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (Metroxraine)*, pp. 697–702. IEEE, 2022.

Niall Mc Shane, Karl McCreadie, Darryl Charles, Attila Korik, and Damien Coyle. Decoding motion trajectories in an upper limb bci: Linear regression vs deep learning. In *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pp. 1039–1044. IEEE, 2023.

Niall McShane, Attila Korik, Karl McCreadie, Naomi Du Bois, Darryl Charles, and Damien Coyle. Embodied virtual reality feedback reshapes neural representations to support continuous three-dimensional motor imagery decoding. *arXiv preprint arXiv:2605.29677*, 2026.

Jinyu Miao, Rujun Yan, Bowei Zhang, Tuopu Wen, Jiaqi Li, Zheng Fu, Kun Jiang, Mengmeng Yang, Jin Huang, Zhihua Zhong, et al. Residual learning towards high-fidelity vehicle dynamics modeling with transformer. *IEEE Robotics and Automation Letters*, 10(7):7404–7411, 2025.

J d R Millán, Rüdiger Rupp, Gernot R Müller-Putz, Roderick Murray-Smith, Claudio Giugliemma, Michael Tangermann, Carmen Vidaurre, Febo Cincotti, Andrea Kübler, Robert Leeb, et al. Combining brain–computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in neuroscience*, 4:161, 2010.

Valeria Mondini, Andreea-Ioana Sburlea, and Gernot R Müller-Putz. Towards unlocking motor control in spinal cord injured by applying an online eeg-based framework to decode motor intention, trajectory and error processing. *Scientific Reports*, 14(1):4714, 2024.

Luis Fernando Nicolas-Alonso and Jaime Gomez-Gil. Brain computer interfaces, a review. *sensors*, 12(2):1211–1279, 2012.

Girijesh Prasad, Pawel Herman, Damien Coyle, Suzanne McDonough, and Jacqueline Crosbie. Applying a brain-computer interface to support motor imagery practice in people with stroke for upper limb recovery: a feasibility study. *Journal of neuroengineering and rehabilitation*, 7(1):60, 2010.

Hannah S Pulferer, Kyriaki Kostoglou, and Gernot R Müller-Putz. Improving non-invasive trajectory decoding via neural correlates of continuous erroneous feedback processing. *Journal of Neural Engineering*, 21(5):056010, 2024.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Martin Spüler and Christian Niethammer. Error-related potentials during continuous feedback: using eeg to detect errors of different type and severity. *Frontiers in human neuroscience*, 9:155, 2015.

Bram van de Laar, Hayrettin Gürkök, Danny Plass-Oude Bos, Mannes Poel, and Anton Nijholt. Experiencing bci control in a popular computer game. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(2):176–184, 2013.

Pengpai Wang, Xuhao Cao, Yueying Zhou, Peiliang Gong, Muhammad Yousefnezhad, Wei Shao, and Daoqiang Zhang. A comprehensive review on motion trajectory reconstruction for eeg-based brain-computer interface. *Frontiers in neuroscience*, 17:1086472, 2023.

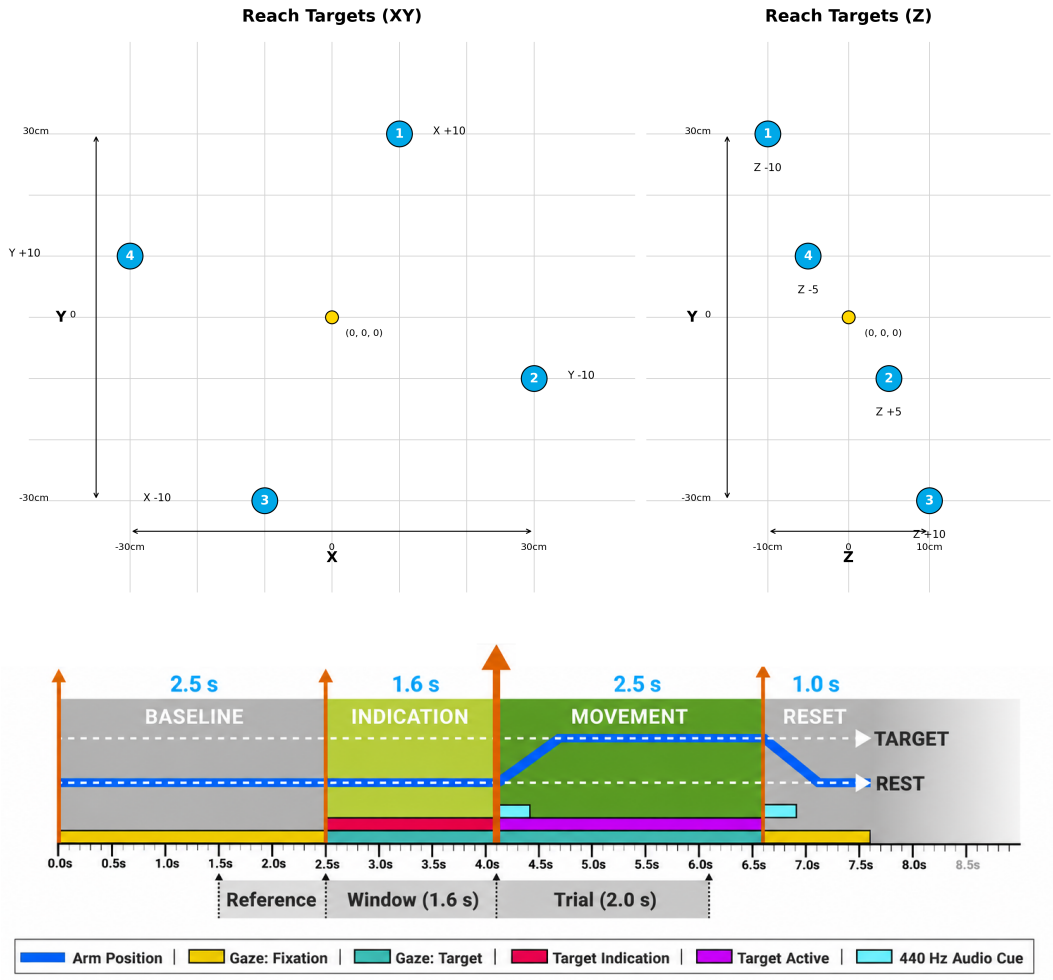
Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pp. 196–202. Springer, 1992.

Francis R Willett, Donald T Avansino, Leigh R Hochberg, Jaimie M Henderson, and Krishna V Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, 2021.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

Baoguo Xu, Yong Wang, Leying Deng, Changcheng Wu, Wenbing Zhang, Huijun Li, and Aiguo Song. Decoding hand movement types and kinematic information from electroencephalogram. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1744–1755, 2021.

APPENDIX A EXPERIMENTS PARADIGM



APPENDIX B EYE-MOVEMENT PROTOCOL AND CONTROL OF OCULAR ARTEFACTS

Eye movements were minimised and participants were instructed to restrict blinking to rest periods to reduce ocular artefacts. Consistent with natural reaching behaviour, participants fixated on the central cross during rest and shifted their gaze to the cued target approximately 1.6 s before movement onset. Because eye movements (Saccades, quick look at target) lasted no more than 200 ms and the decoder input window began 640 ms before movement onset, they occurred outside the analysed EEG window. Eye-tracking signals were not used as inputs to either the CNN–LSTM decoder or the residual correction model.

Several additional measures were used to reduce the possibility that decoding performance was driven by ocular activity. The CNN–LSTM decoder used the targeted 17-channel subset described in Section 2.1, which excluded the most anterior electrodes that are generally most susceptible to electrooculographic contamination. Although electrodes such as F3, Fz, and F4 may retain some residual ocular activity, ocular artefacts attenuate with increasing distance from the eyes. During the development of the experimental paradigm, we also conducted pilot control experiments in which participants viewed the same visual stimuli without performing motor imagery. Under this condition, decoding performance remained at chance level, indicating that visual processing and task-related eye movements alone were insufficient to explain the observed decoding performance. Taken together, the timing of the instructed gaze shift, the restriction of blinking to rest periods, the pilot control

experiments, and the selected electrode montage provide converging evidence that ocular activity was unlikely to have influenced the decoding performance observed in this paradigm.

APPENDIX C HYPERPARAMETERS SEARCHING

Table 3: Hyperparameter search space for the residual correction network

Parameter	Type	Search Range
Learning rate	Log-uniform	$[10^{-5}, 10^{-3}]$
Hidden layer dimension	Categorical	$\{128, 256, 512\}$
Batch size	Categorical	$\{64, 128, 256\}$
Residual scale (x-axis) ($Scale_1$)	Uniform	$[0.1, 1.0]$
Residual scale (y-axis) ($Scale_2$)	Uniform	$[0.1, 1.0]$
Residual scale (z-axis) ($Scale_3$)	Uniform	$[0.1, 1.0]$
Number of hidden layers	Integer	$[2, 4]$

Table 4: Hyperparameter search space for the CNN–LSTM decoder

Parameter	Type	Search Range
Learning rate	Log-uniform	$[10^{-6}, 10^{-2}]$
Batch size	Categorical	$\{6, 12, 24, 48\}$
Convolutional layers	Integer	$[2, 4]$
Kernel size	Categorical	$\{3, 5, 7\}$
Dropout	Uniform	$[0.1, 0.4]$
Activation	Categorical	$\{\mathrm{ReLU}, \tanh\}$
Bias regularisation	Log-uniform	$[10^{-4}, 10^{-3}]$
LSTM layers	Integer	$[1, 2]$
LSTM units	Integer	$[20, 100]$

APPENDIX D PREREQUISITE ANALYSIS FOR STATISTICAL TESTS

As an additional repeated-measures modelling check, linear mixed-effects models were explored separately for the 2D and VR feedback conditions, with decoder strategy specified as a fixed effect and participant specified as a random intercept. However, these models produced singular fits in one or more metric-by-feedback analyses, indicating that the participant-level random-intercept variance was estimated at or near zero and that the random-effects structure was not reliably supported by the available participant-level data. Therefore, LMM-based inference was not retained as the primary statistical framework. Instead, test selection was based on distributional diagnostics conducted separately for each performance metric and feedback condition. Paired t-tests were used when the participant-level paired differences were sufficiently consistent with the normality assumption. Wilcoxon signed-rank tests were used when normality was not consistently supported but the paired differences remained sufficiently symmetric for a signed-rank analysis. The selection of the corresponding omnibus test followed the same metric-level assessment: one-way repeated-measures ANOVA was used for analyses assigned to the parametric framework, whereas the Friedman test was used for the non-parametric analysis. Shapiro–Wilk tests and Quantile–Quantile plots (Q–Q plots) were used to assess normality, while histograms and skewness coefficients were examined to assess symmetry and identify substantial distributional deviations. These diagnostics determined the statistical tests used to compare FRL with the BCI decoder strategies FDG, SAT, and WSR for Pearson correlation and RMSE.

D.1 WITH 2D FEEDBACK

For Pearson correlation under 2D feedback, the participant-level paired differences between FRL and each comparison strategy were sufficiently consistent with the normality assumption. The corresponding Q–Q plots aligned closely with the diagonal reference lines, the Shapiro–Wilk tests

did not indicate significant departures from normality, and the histograms and skewness coefficients indicated approximately symmetric distributions. Therefore, a one-way repeated-measures ANOVA was used for the omnibus comparison, followed by paired t-tests for the planned comparisons of FRL with FDG, SAT, and WSR. In contrast, the normality assumption was not consistently supported across the 2D RMSE comparisons. To maintain a consistent non-parametric analysis within this metric, the Friedman test was used for the omnibus comparison, followed by Wilcoxon signed-rank tests for the three planned pairwise comparisons. The paired RMSE differences remained sufficiently symmetric to support the use of the signed-rank test.

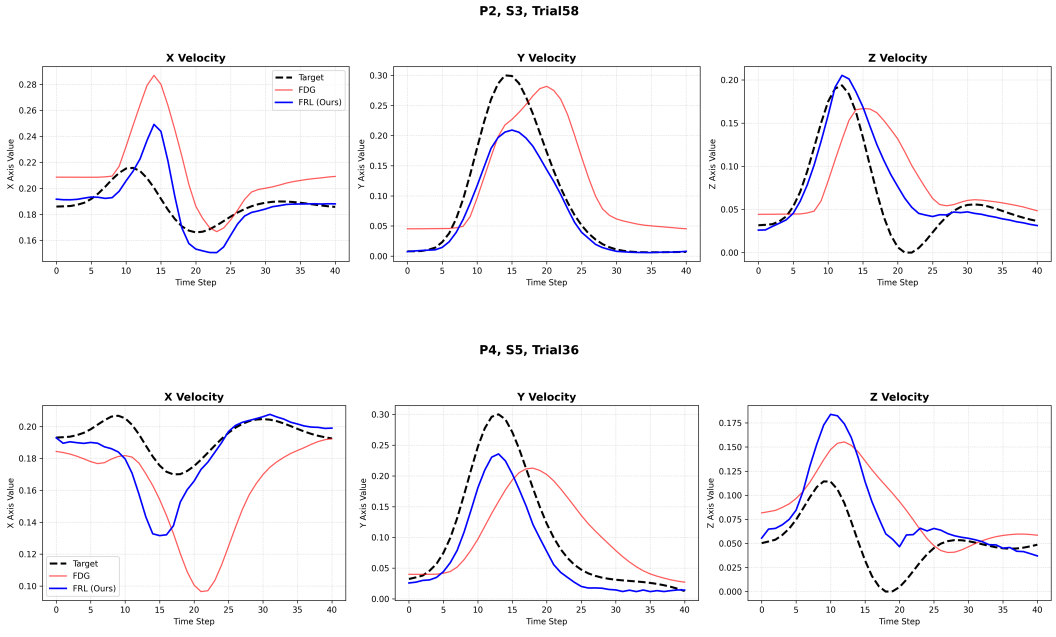
D.2 With VR Feedback

In the VR spatial feedback environment, the diagnostic analyses generally supported the use of parametric comparisons for both Pearson correlation and RMSE. All paired-difference distributions did not show significant departures from normality in the Shapiro–Wilk tests and were approximately aligned with the reference lines in the Q–Q plots. Given this symmetry and normality test, paired t-tests were retained consistently for all three VR correlation comparisons. The RMSE paired differences satisfied the normality assumption across all comparisons. Therefore, one-way repeated-measures ANOVAs were used for the omnibus correlation and RMSE comparisons under VR feedback, followed by paired t-tests comparing FRL with FDG, SAT, and WSR.

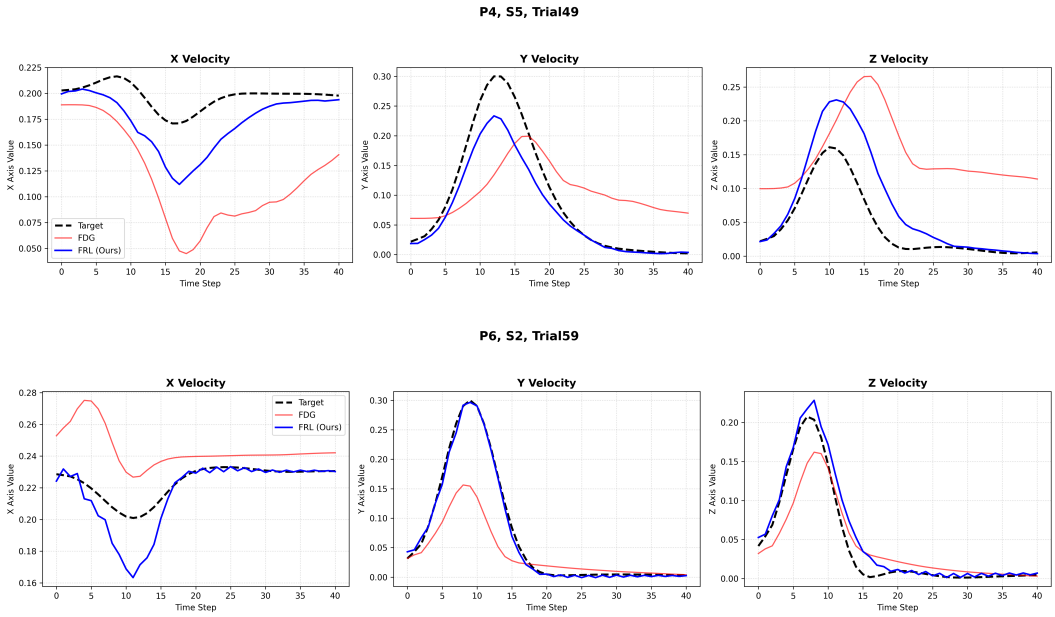
Appendix E Visualisation of Residual Correction

The following figures illustrate the residual correction applied to several individual trials under the 2D and VR feedback conditions.

E.1 2D



E.2 VR



APPENDIX F EXPERIMENTS COMPUTE RESOURCES

All experiments were conducted using NVIDIA RTX 3090 GPUs. The residual models were implemented in PyTorch and trained with CUDA acceleration. Each experiment uses 8 GPUs and requires approximately 6 hours of wall-clock time for hyperparameter optimisation, corresponding to 48 GPU hours per run. A total of 20 experiments (10 for 2D and 10 for VR)were conducted, resulting in approximately 960 GPU hours for hyperparameter optimisation. In addition, the final training of residual correction models for each participant and feedback condition requires approximately 2 GPU hours per run, leading to an additional cost of 40 GPU hours. Overall, the total computational cost is approximately 1000 GPU hours. Experiments were executed in parallel with up to 6 runs simultaneously, resulting in a total wall-clock time of approximately 2 days.