Handwriting decoding as a challenging Motor Imagery task for EEG Foundation Models

Srinivas Ravishankar
Cognitive Science
University of California San Diego
San Diego, CA 92093
srravishankar@ucsd.edu

Virginia R. de Sa
Cognitive Science, Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093
desa@ucsd.edu

Abstract

Foundation Models (FMs) for EEG have achieved state-of-the-art performance on multiple Motor Imagery (MI) datasets, indicating their potential to provide robust, generalizable representations across diverse contexts. In this work, we investigate handwriting decoding as a challenging MI task to evaluate the generalizability of FMs for EEG. Recent studies in handwriting decoding have reported higher-than-chance performance in decoding handwritten letters from EEG. However, all prior works have attempted to decode handwriting from EEG during actual motion. Furthermore, they assume that precise movement-onset is known. We introduce a setting closer to real-world use where either movement-onset is not known or movement does not occur at all, fully utilizing motor imagery. Crucially, we find that current FMs for EEG, despite showing SOTA performance in multiple MI datasets (hand vs foot classification) are outperformed by a task-specific EEGNet model in this fine-grained task. In parallel, we also investigate avenues that are most promising for improving decoding performance. In our 4-letter classification task, we show that (a) Knowledge of movement-onset is crucial to reported decoding performance in prior works, with average performance across subjects dropping from 41.4% to 32.4%. (b) Increasing test-time signal quality provides significant performance improvements (45% to 78% in our best subject) compared to scaling training data with single-trial EEG. (c) Fully imagined handwriting can be decoded from EEG with higher-than-chance performance.

1 Introduction

Patients with locked-in syndrome are unable to communicate through traditional means such as speech or writing, since their motor functions are severely impaired [1, 2]. Brain Computer Interfaces (BCIs) that allow these subjects to communicate even in a limited fashion can significantly increase their quality of life [3, 4]. Many existing BCIs utilize motor imagery [5] as a control signal, which involves the subject attempting or imagining some familiar motor movement. The motor imagery paradigm is desirable as it allows for endogenous, asynchronous control, making it an ideal choice for everyday usage. Classical motor-imagery BCIs have used control signals such as left-hand/right-hand clenching, or tongue/hands/feet movement, etc [6]. However, these types of control signals suffer from limited information transfer rates (ITR), and subjects must endure slow and cumbersome communication.

Recently, imagined handwriting has produced significantly higher ITR than prior work, having been explored in the invasive setting [7]. This previous study showed that the motor-imagery signals associated with handwriting can be reliably decoded using data from intracortical microelectrodes implanted in the precentral gyrus. In a single subject, pre-screened for high prior performance, this

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

work demonstrated a real-time character error rate (CER) as low as 5.4%, with a rate of communication approaching average smartphone typing speeds of the participant's demographic.

Inspired by this work, there have been a number of studies attempting to decode handwriting from Electroencephalography (EEG), a non-invasive brain modality. However, we find that many of these studies are affected by various confounds: [8] reported an accuracy of 94% in a classification between the characters of "HELLO, WORLD!". However, their experimental design involved writing the above characters in the same order every trial in specified boxes while looking at the tablet. Independent Components (ICs) representing eye-movements were not discarded from the EEG. In our own analysis, excluding all other predictors, this artifact alone achieves 85.6% accuracy in the classification task, while reported accuracy is 94%. A later study [9] used this dataset and achieved similarly high classification performance $\sim 91\%$. It is likely that this work was similarly confounded by eye-movement artifacts owing to the experimental design.

Two works [10, 11] collected EEG data for all 26 characters of the English alphabet during index-tracing and handwriting respectively. The reported performance was high, around 45% and 31% respectively on the entire letter set. In the first work, we find that using only electrodes `Fp1`, `Fp2`, `T8`, `TP10`, `P8`, average decoding performance is $\sim 24\%$, well above chance (3.8%), indicating information from eye/shoulder motion is likely contributing to the high decoding performance. Furthermore, in their experimental design, the onset of the letter cue also acts as the writing cue. Thus, the letter is traced right after it appears on the screen, and trials are epoched around this onset. Recent works [12, 13] have shown that complex visual stimuli can be decoded from EEG. Thus, decoding performance may also be affected by visual decoding confounds. In the second work's experimental design, while the letter cue and the writing cue are separated, the subject received visual feedback on-screen as they traced the letter. A natural inclination while writing letters is to follow the pen tip with the eyes, potentially generating sensorimotor signals that are characteristic of each letter and separate from the sensorimotor signals associated with handwriting. The authors show that gamma frequency and frontal electrodes are dominant in decoding performance. Recent work [14] has also shown that complex video stimuli can be decoded from EEG, raising similar visual decoding possibilities as the previous work.

In contrast, while [15] reports modest performance, $\sim 26.2\%$ over a 10 letter classification, their subjects trace the letter inside a box while fixating on the screen, and do not see any visual feedback related to the letter while they write. We emphasize that future work must follow these considerations to preclude confounds.

However, a practical system for locked-in patients will require the ability to work with motor imagery, as in the intracortical study, rather than actual motion. While all prior work has dealt with real writing (on a tablet or in the air), our work investigates the unique challenges associated with decoding handwriting in settings where onset timing is unknown, or movement does not occur at all.

In addition, we conduct novel analyses in the setting employed by prior work, involving real motion and known onset timing. Invasive work [7] has shown that scaling training data is a crucial step in achieving high performance, and we investigate if an EEG-based decoder similarly scales. To answer this question, we collect a large number of trials from a subject, to conduct sample-complexity analysis. In a complementary analysis, we also investigate if the SNR of a single-trial of EEG acts as a bottleneck to decoding performance.

The contributions of our work are as follows:

- We find that knowledge of onset timing is crucial to achieve reported decoding performance in literature; in a realistic setting where onset timing is unknown, performance drastically reduces, motivating future work aimed at addressing this gap.
- We find that scaling training data appears to steadily increase decoding performance in single-trial decoding. While this suggests that Foundation Models might be particularly beneficial to this task, we find that specialist models consistently outperform FMs, challenging the generalization of MI datasets used in pretraining.
- Increasing SNR by averaging trials solely during evaluation improved performance in our best subject from 45% to 78% in a 4-way classification, a 73% increase.
- To the best of our knowledge, this work is the first to show that purely imagined handwriting can be decoded from single-trial EEG data with higher-than-chance performance.

2 Methods

2.1 Dataset

Data was collected from four right-handed participants (2 male and 2 female) using 32 EEG channels. While our cohort size is quite limited compared to prior work [15], we opted instead to collect more trials per subject and over multiple days. Our dataset can thus be used to analyze cross-session stability, with some sessions within a subject separated by more than 1 year. In place of the standard 10-20 montage, a custom montage was created to record more densely from the motor area of the brain, at the expense of fewer electrodes in the occipital and posterior regions of the head. The montage is provided in the supplementary. Cz was used as reference. Participants were instructed to fixate on a screen while they wrote on an Android tablet placed on the desk in front of them, using a digital stylus as shown in Figure 1a in the Motor Execution (ME) paradigm. In the Motor Imagery (MI) paradigm, they placed their hands on their lap instead. Thus, letters were independent of the location on the tablet, and participants' eyes did not track the letter as they were writing it. Four letters were chosen for this study, L, V, O and W. The block diagram of the experiment for a single trial is shown in Figure 1b.

(a) Experimental Setup (b) Block diagram

Figure 1: Experimental design for a single trial. The participant fixates on the monitor in front while writing on the tablet on the desk. One of the four letters is shown for 800 ms, followed by a blank screen for a randomly chosen period between 400-600 ms. Then a fixation cross appears on the screen for 1000 ms, during which the participant writes the letter on the tablet while looking at the cross. This is followed by a blank screen for 500 ms, after which the next trial begins.

Table 1 provides details about the sessions collected over different days, from different subjects, and under two possible timing paradigms. The number of trials were equally distributed among the four letters, presented in a pseudo-random order that ensured (a) Each 4 trials had all letters presented (b) No letter was presented twice in succession. The EEG data were collected at 1000 Hz, notch filtered at 60 Hz, band-pass filtered between 0.3 Hz and 70 Hz, decomposed using ICA, and ICs representing muscle or eye artifacts were rejected by manual inspection. Since this data had a maximum frequency of 40 Hz, it was downsampled to 100Hz for efficient processing without loss of information. We also collected pen-tip trajectories using the tablet while the participant was writing the letters, at 200Hz. These trajectories were re-sampled to match the EEG data. Synchronization was done using photo-diodes, further details are given in supplementary.

2.2 Decoder

For our classification decoders, we use EEGNet [16], as well as foundational models CBraMod[17] and MIRepNet[18] which have reported SOTA results on MI tasks. All models output a letter label corresponding to a snippet of input EEG data, trained using cross-entropy loss. We quantify the final performance of the system by measuring its accuracy in classifying held-out letter instances as one of four letters (L, V, O and W). The chance performance is 25% since the test data set is balanced. All models were subject-specific. Further implementation details are provided in the appendix.

2.2.1 Multi-setting Classification

We perform classification experiments under 3 settings: ME movement-centered, ME cue-centered, and MI cue-centered. In the first setting, we follow prior work and epoch trials around true movement onset. This is meant as a positive control, since prior work has shown that handwriting in this setting can be successfully decoded. In the second setting, we approach a realistic setting where real movement occurs, but precise movement onset is not known to the model. Trials are epoched around the cue to write letters, with timing to actual onset varying across trials. Finally, the third setting is the most realistic setting possible in healthy subjects. Participants were asked to imagine writing the letters while they placed their palms on the lap. Since no actual motion occurred, all trials are cue-centered.

For EEGNet, 1000 ms of data was used as input. This length was chosen to be longer than the longest duration of letter writing in the data, which was ~ 600 ms. In the movement-centered setting, trials were epoched [-200ms, 800ms] around movement onset. In cue-centered settings, trials were epoched [0ms, 1000ms] around the onset of the writing cue (fixation cross). We trained EEGNet models with a four-way classification head and evaluated it using 5-fold cross-validation. For CBraMod, we evaluated windows of 1s (to match the previous setting) as well as 4s (to match its pretraining statistics). We found the latter outperformed the former. Trials were epoched [-1s, 3s] around movement onset, chosen through cross-validation.

2.2.2 Scaling training data vs test-time EEG SNR for performance

We conduct analysis to identify what most improves handwriting decoding from EEG: collecting more training data or increasing test-time SNR? Prior efforts in invasive handwriting decoding achieve significant improvements in performance by scaling training data. However, the inherently low SNR of EEG might present a bottleneck. To evaluate this question, we collected large per-class datasets (e.g., ~ 2400 trials from subject S1, evenly split across 4 letters).

To fairly evaluate the effect of training set size, we hold out a fixed test set (the final 160 trials) and train on progressively larger fractions of the remaining data (10–100%). In parallel, we examine the effect of improving test-time SNR while keeping the training set fixed. Higher effective SNR is simulated by averaging repeated trials of the same class.

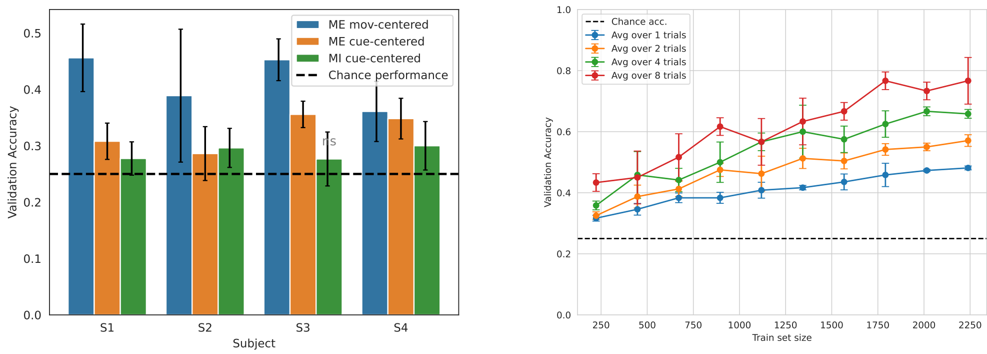
3 Results

3.1 Multi-setting Classification

Despite MIRepNet achieving state-of-the-art performance in multiple hand vs feet MI datasets, it failed to achieve performance above chance on handwriting decoding. Furthermore, we found that EEGNet consistently outperformed a finetuned CBraNet in handwriting decoding, achieving a mean performance across subjects of 41.463 ± 4.1 vs 38.14 ± 3.62 in the ME-movement-centered setting. This suggests handwriting decoding can be a challenging MI benchmark for evaluating future Foundation Models. We show the performance of EEGNet across the three settings and four subjects in Figure 2a. In line with prior work, all subjects achieve higher than chance accuracy in the ME movement-centered setting, and an overall average of 41.463 ± 4.1. However, we report a significant drop in performance to 32.453 ± 2.861, when the trials are ME cue-centered rather than ME movement-centered. Notably, data augmentation strategies such as random shifts of the input to make the model temporally invariant did not improve performance. Finally, we show for the first time that imagined handwriting (MI) can be decoded with greater than chance performance, at 28.758 ± 1.064 (in all subjects except S3). While performance in this setting is also lower than the movement-centered ME setting, part of this drop is likely explained by what we observe in the cue-centered ME setting.

3.2 Single-trial EEG SNR is the performance bottleneck

We evaluated the effect of training dataset size versus test-time SNR. Test-time SNR was increased by averaging multiple trials of the same letter (aligned to movement onset) during evaluation. As shown in Figure 2b, scaling the training set yields modest single-trial accuracy gains, but boosting test-time SNR produces far larger improvements. In S1, accuracy rose from 45% on single trials to 78% when averaging eight trials.



(a) Decoding performance

(b) Sample Complexity vs SNR Boosting

Figure 2: (a) Subject-wise decoding performance on 3 different settings. Performance drops significantly when motor activity onset is unknown (blue vs orange), indicating a key issue to overcome. Imagined handwriting (green) can be decoded with higher-than-chance performance. (b) Decoding performance on S1's last 160 trials, when scaling training data and averaging over various number of trials of the same letter. Mean and standard deviation over 3 random seeds. While there is some performance improvement when scaling training data, boosting test-time SNR significantly improves performance even when training with single-trial EEG;

4 Discussion

Our results support recent work that motor information associated with handwriting can be decoded from EEG. Average single-trial decoding performance over subjects was 41.463% in a 4-way letter classification. Importantly, we show that the precise timing of movement onset is crucial to achieving this performance. With real motion but unknown movement onset, performance dropped significantly. Crucially, this was not improved with data augmentation strategies that attempted to make the model temporally invariant. This may suggest that the motion-onset EEG activity is not clearly detectable.

Additionally, this work demonstrated that decoding handwriting from purely imagined handwriting is possible, but exhibits lower performance compared to movement-centered decoding with real movement. Notably however, the decoding performance is on par with the cue-centered ME trials, suggesting that solutions to the temporal invariance problem might result in performance gains.

Our sample complexity results show a promising trend of improved performance with increased training data, suggesting that foundation models might be beneficial for this task. However, test-time SNR remains a key bottleneck, with significant improvements obtained from averaging time-locked evaluation trials. Finally, current FMs do not appear to improve performance on handwriting decoding compared to specialist models. In particular, we hypothesize that MIRepNet's chance performance on this dataset is due to the strict filtering (8-30 Hz) in their pretraining. Prior work [15] observed useful signals for decoding handwriting in the 0.3-3Hz range, suggesting that FMs should employ minimal preprocessing during pretraining. Overall, more diverse datasets might also be necessary during pretraining to produce generalizable representations useful for a wide variety of tasks.

5 Conclusion and Future Work

In this work, we demonstrated that prior studies on handwriting decoding from EEG systematically overestimate the performance of a realistic BCI relying on imagined handwriting. Building on these studies, we identified broader trends in decoding both executed and imagined handwriting, showing that motor imagery can be decoded successfully even in paradigms without overt movement. Our results show encouraging performance gains with increased data, suggesting that foundation models could be particularly beneficial for this task. However, current foundation models, while achieving state-of-the-art results on other motor imagery benchmarks, do not yet surpass specialist models here. Taken together, this work underscores the challenges of building a high-performing EEG-based handwriting BCI and establishes imagined handwriting decoding as a challenging benchmark for evaluating foundation models for EEG.

Acknowledgments Supported by NSF IIS-1817226, NSF IIS-CRCNS-2208362, and hardware funding from Sony.

References

[1] Marie-Christine Rousseau, Stephane Pietra, Mohammed Nadji, and Thierry Billette de Villemeur. Evaluation of quality of life in complete locked-in syndrome patients. *Journal of palliative medicine*, 16(11):1455–1458, 2013.

[2] Steven Laureys, Frédéric Pellas, Philippe Van Eeckhout, Sofiane Ghorbel, Caroline Schnakers, Fabien Perrin, Jacques Berre, Marie-Elisabeth Faymonville, Karl-Heinz Pantke, Francois Damas, et al. The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? *Progress in brain research*, 150:495–611, 2005.

[3] Tomislav Milekovic, Anish A Sarma, Daniel Bacher, John D Simeral, Jad Saab, Chethan Pandarinath, Brittany L Sorice, Christine Blabe, Erin M Oakley, Kathryn R Tringale, et al. Stable long-term BCI-enabled communication in ALS and locked-in syndrome using LFP signals. *Journal of neurophysiology*, 120(7):343–360, 2018.

[4] Andrea Kübler. The history of BCI: From a vision for the future to real support for personhood in people with locked-in syndrome. *Neuroethics*, 13(2):163–180, 2020.

[5] Jean Decety. The neurophysiological basis of motor imagery. *Behavioural brain research*, 77(1-2):45–52, 1996.

[6] Eltaf Abdalsalam Mohamed, Mohd Zuki Yusoff, Aamir Saeed Malik, Mohammad Rida Bahloul, Dalia Mahmoud Adam, and Ibrahim Khalil Adam. Comparison of EEG signal decomposition methods in classification of motor-imagery BCI. *Multimedia Tools and Applications*, 77:21305–21327, 2018.

[7] 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.

[8] Leisi Pei and Guang Ouyang. Online recognition of handwritten characters from scalp-recorded brain activities during handwriting. *Journal of Neural Engineering*, 18(4):046070, 2021.

[9] Jun-Young Kim, Deok-Seon Kim, and Seo-Hyun Lee. Towards scalable handwriting communication via EEG decoding and latent embedding integration. *arXiv preprint arXiv:2411.09170*, 2024.

[10] Ayush Tripathi, Aryan Gupta, AP Prathosh, Suriya Prakash Muthukrishnan, and Lalan Kumar. NeuroAiR: Deep learning framework for airwriting recognition from scalp-recorded neural signals. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[11] Xiaowei Jiang, Charles Zhou, Yiqun Duan, Ziyi Zhao, Thomas Do, and Chin-Teng Lin. Neural spelling: A spell-based BCI system for language neural decoding. *arXiv preprint arXiv:2501.17489*, 2025.

[12] Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. Seeing through the brain: Image reconstruction of visual perception from human brain signals, 2023.

[13] Teng Fei, Abhinav Uppal, Ian Jackson, Srinivas Ravishankar, David Wang, and Virginia R de Sa. Perceptogram: Reconstructing visual percepts from EEG. *arXiv preprint arXiv:2404.01250*, 2024.

[14] 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, 2025.

[15] Markus R Crell and Gernot R Müller-Putz. Handwritten character classification from EEG through continuous kinematic decoding. *Computers in Biology and Medicine*, 182:109132, 2024.

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

[17] Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for EEG decoding. *arXiv preprint arXiv:2412.07236*, 2024.

[18] Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, and Dongrui Wu. MIRepNet: A pipeline and foundation model for EEG-based motor imagery classification. *arXiv preprint arXiv:2507.20254*, 2025.

A

A.1 Montage

Data was collected from 32 EEG channels. Instead of the standard 10-20 montage, a custom montage was designed to record more densely from the mid-line area, as shown in Figure 3. This was done to better capture signals from the motor cortex, at the expense of fewer electrodes in the occipital and posterior regions of the head. Cz was used as reference.

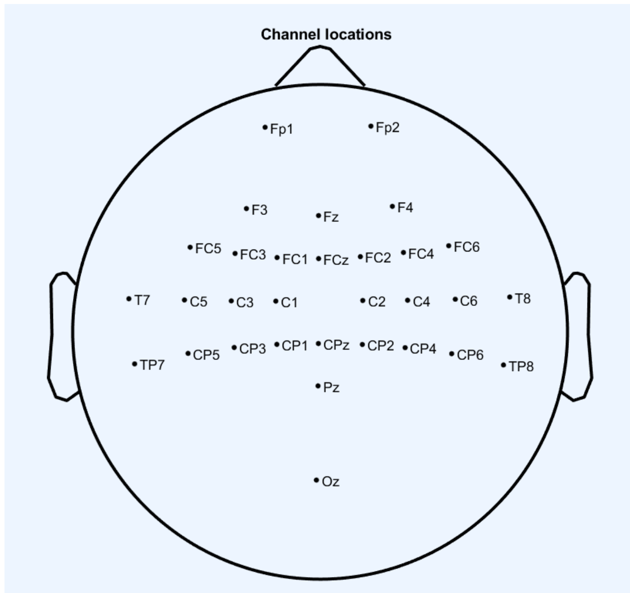


Figure 3: A custom montage designed to record more densely from the midline area, over the motor cortex. Cz as reference.

A.2 Synchronization

Three streams of data needed to be synchronized: The EEG stream from the amplifier, task cues from the recording computer, and the pen trajectories from the tablet. The LabStreamingLayer (LSL) networking ecosystem was used to send and receive the data between the devices. However, network latency can lead to streams being offset.

To ensure precise synchronization, we connected two photo-diodes to the amplifier, attaching one to a section of the tablet screen, and one to a section of the experimental monitor. The task program and the trajectory recording program were designed to flash the photo-diode region whenever a task cue was being shown, or a trajectory was drawn. These photodiode spikes were perfectly synchronized to the EEG since they fed into the same amplifier. The data/timestamps that were recorded with some latency by the LSL protocol were then aligned to the EEG stream using the closest photodiode spike in it. We found network latencies and offsets of upto 80 ms that were resolved using the photodiode-based synchronization.

A.3 Implementation details

All models were trained using the Adam optimizer with a learning rate of 1e-3. The EEGNet architecture employed 8 temporal and 8 spatial filters. CBraMod was fine-tuned in two configurations: full-model fine-tuning and a frozen backbone with a trainable linear head, with the latter achieving higher performance (reported in the results). All experiments were conducted on a single NVIDIA A6000 GPU.

A.4 Data collection

Subject	Session date	N Trials
S1	23 Mar 2023	315
S1	26 Apr 2023	696
S1	4 Mar 2024	1268
S1	1 Jan 2025	1200
S1 MI	1 Jan 2025	1240
S1	11 Jan 2025	120
S1 MI	11 Jan 2025	400
S2	25 Feb 2024	339
S2	27 Feb 2024	336
S2	20 Oct 2024	778
S2 MI	29 Dec 2024	1198
S2	20 Jan 2025	319
S3	3 Nov 2024	474
S3 MI	3 Nov 2024	300
S4	7 Nov 2024	400
S4 MI	7 Nov 2024	400

Table 1: Details of data collection schedule