2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3315-8618-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/EMBC58623.2025.11253543

Fine-Tuning Strategies for Continual Online EEG Motor Imagery Decoding: Insights from a Large-Scale Longitudinal Study

Martin Wimpff[1], Bruno Aristimunha[2,3], Sylvain Chevallier[2], and Bin Yang[1]

***Abstract*— This study investigates continual fine-tuning strategies for deep learning in online longitudinal electroencephalography (EEG) motor imagery (MI) decoding within a causal setting involving a large user group and multiple sessions per participant. We are the first to explore such strategies across a large user group, as longitudinal adaptation is typically studied in the single-subject setting with a single adaptation strategy, which limits the ability to generalize findings. First, we examine the impact of different fine-tuning approaches on decoder performance and stability. Building on this, we integrate online test-time adaptation (OTTA) to adapt the model during deployment, complementing the effects of prior fine-tuning. Our findings demonstrate that fine-tuning that successively builds on prior subject-specific information improves both performance and stability, while OTTA effectively adapts the model to evolving data distributions across consecutive sessions, enabling calibration-free operation. These results offer valuable insights and recommendations for future research in longitudinal online MI decoding and highlight the importance of combining domain adaptation strategies for improving BCI performance in real-world applications.**

***Clinical Relevance*—Our investigation enables more stable and efficient long-term motor imagery decoding, which is critical for neurorehabilitation and assistive technologies.**

I. INTRODUCTION

A brain-computer interface (BCI) measures brain activity and translates it into control commands for computers or other external devices [1]. This provides a direct alternative to natural neural pathways, enabling BCIs to replace, restore, enhance, supplement, or improve the brain's interaction with its external or internal environment [2], [3].

A widely used method for controlling BCIs is the motor imagery (MI) paradigm. In this paradigm, the user imagines the movement of a body part without physically performing the action. This imagined movement engages neural mechanisms similar to those involved in actual execution [4], making MI-BCIs particularly effective for promoting motor recovery in chronic stroke patients [2].

However, novice users often struggle to elicit the correct brain patterns, a challenge known as BCI inefficiency [5]–[7], also referred to as BCI illiteracy. Unlike paradigms such as P300 or steady-state visually evoked potentials, which rely on responses to external stimuli, MI requires users to endogenously modulate their brain rhythms, i.e., actively regulate their neural activity which is known to be more challenging. Potential solutions to this BCI inefficiency fall mostly into two categories: either promoting user learning or improving the decoder [5], [6].

As with almost any skill acquisition process, effective BCI usage depends on practice guided by feedback [8]. Research even indicates that implicit learning, where users develop skills through self-regulation guided by feedback, may be more effective than explicitly guiding or instructing the user [9]–[11]. To facilitate such implicit learning, closed-loop systems that provide real-time feedback are essential.

The other possible solution for successful BCI usage is to increase the quality of the decoder. While this has been the subject of a large research effort for several decades [12], one important – although less investigated issue – is the adaptiveness of the decoder. While users must develop the ability to generate the correct brain patterns, the system must also adjust to the evolving neural activity of the users. Decoder adaptation is essential in this process, ensuring that the BCI remains effective despite evolving brain patterns over time.

Importantly, user learning and decoder adaptation are not independent processes but are tightly interconnected. This dynamic interaction can be conceptualized as a *two-learner problem*, where both the user and the decoder adapt to each other's changing trajectories over time [13]–[15]. Through mutual learning, they try to find an optimal communication strategy. In longitudinal settings, this interdependence becomes particularly significant as the user's neural patterns may shift substantially over time due to factors such as learning, neuroplasticity, or changes in the environment.

Research has demonstrated that decoder adaptation can enhance the user's ability to control the BCI, leading to overall performance improvements [15]–[20]. However, determining the optimal strategy for recalibrating decoders over time remains an open question. Current approaches vary primarily in the frequency of recalibration [17], and the data composition used for recalibration [16], [19].

Within the realm of deep learning, the process of gradually adapting to an incoming stream of data from different domains can be described as domain-incremental continual learning [21], [22]. However, continual learning in MI decoding presents unique challenges compared to its applications in other fields. In traditional settings, continual learning methods prioritize efficient adaptation to new tasks or domains while retaining knowledge of previous ones,

This research was funded by the Quantum Human Machine Interfaces (QHMI) project within the QSens - Quantum Sensors of the Future Cluster grant number 03ZU1110DC. BA and SC is supported by ANR-22-CE33-0015-01 and ANR-17-CONV-0003.

[1]Martin Wimpff and Bin Yang are with the Institute of Signal Processing and System Theory, University of Stuttgart, 70569 Stuttgart, Germany martin.wimpff@iss.uni-stuttgart.de

[2]Bruno Aristimunha and Sylvain Chevallier are with the Inria TAU team, LISN-CNRS, Universite Paris-Saclay 91400 Orsay, France. [3] Federal University of ABC, Santo Andre, Brazil

thereby avoiding catastrophic forgetting.

In MI decoding, the latter consideration is irrelevant, as the data distribution continuously evolves over time, and revisiting previous distributions is neither necessary nor feasible, considering the non-stationary nature of the data. This differs from applications such as automotive systems, where recurring conditions (e.g., varying weather) require continual learning to handle repeated scenarios. However, it is worth mentioning that a certain level of decoder stability tends to benefit user learning in MI decoding [18].

Another distinctive aspect of continual MI decoding across multiple sessions is the potential absence of calibration data for the upcoming (target) session [16]. As a result, offline adaptation between sessions must rely solely on the most recent data, i.e., the data from previous session(s).

To address the distribution shift between consecutive sessions, i.e., the most recent session and the upcoming target session, online test-time adaptation (OTTA) [23], [24] emerges as a viable approach. OTTA leverages the incoming sample-wise stream of unlabeled target data after deployment to adapt the model dynamically to the evolving unknown target distribution.

While decoder adaptation and continual learning have shown promise, key questions remain about effectively integrating fine-tuning strategies in longitudinal MI decoding across large user groups. The dynamic interplay between user learning and decoder adaptation, alongside EEG's non-stationary nature and limited target data, demands broader studies beyond single-subject or short-term settings.

To address these challenges, we systematically investigate continual learning for online MI decoding in a large-scale longitudinal setting. Our contributions are as follows:

- We are the first to investigate deep learning-based continual learning for MI decoding in a longitudinal setting across a large user group (61 subjects).
- We examine the impact of different fine-tuning strategies on the performance and stability in a realistic causal pseudo-online [25] manner.
- We demonstrate the effectiveness of combining offline fine-tuning together with online test-time adaptation to establish a comprehensive, fully adaptive calibration-free decoding framework. This framework effectively leverages new data as it becomes available to adapt the decoder to users' evolving neural patterns, addressing the domain shifts naturally present in biosignals. It not only ensures sustained performance and stability across sessions and subjects but also enables continuous performance improvements.

II. MATERIALS & METHODS

A. Data

We employ the dataset published by *Stieger et al.* in 2021 [26], which includes data from 61 subjects with 7-11 sessions per user. To our knowledge, this is the only publicly available motor imagery (MI) dataset that captures longitudinal user learning within a large population with online feedback [27]. In contrast, commonly used BCI datasets, such as those from the BCI competitions [28], focus on small user groups (typically around 10 subjects) and neglect user learning, as they include only a few sessions (typically ≤ 2). More recent databases [7], [29] have expanded to include significantly larger subject pools, enabling more robust analyses across users but still offer limited sessions per user. While data recorded for the Cybathlon competitions [17] addresses this limitation by including multiple sessions over extended periods [18], it focuses solely on a single patient, i.e., the pilot competing in the event. The only other dataset comparable to the one used in this study is provided by *Forenzo et al.* [19]. However, it includes significantly fewer subjects and sessions than the Stieger2021 dataset. Moreover, since their study alternates between different decoders, they were not able to observe any user learning over time. While this potentially provides valuable insights into the impact of decoder stability on user learning, it raises questions about the dataset's suitability to investigate (mutual) learning dynamics.

Stieger2021 dataset: The dataset was collected at a sampling rate of 1000 Hz using 64 EEG channels. Data collection spanned 7 to 11 sessions, with an 8-week gap between the first two sessions, followed by sessions recorded every 2-3 days. For our analysis, we selected 24 channels centered around the motor cortex, resampled the data to 250 Hz, and applied a band-pass filter between 8 Hz and 30 Hz to capture the μ and β rhythm. Each session consisted of 450 trials, equally divided among three paradigms: left/right movement, up/down movement, and combined 2D movement. Participants in the *left/right (LR)* movement paradigm imagined opening and closing their right (left) hand to move the cursor to the right (left). In the *up/down (UD)* movement paradigm, they imagined opening and closing both hands to move the cursor upward and voluntary rest to move it downward. The combined 2D movement paradigm required participants to integrate both types of imagery for two-dimensional cursor control. For this study, we focused exclusively on the binary paradigms (LR and UD), as the four-class task typically results in accuracies too low for practical control [26], [30]. This results in 150 trials per session per paradigm.

At the beginning of each trial, the target appeared on the screen for 2 s, indicating the desired direction of cursor movement. During the subsequent feedback phase, participants had up to 6 s to steer the cursor toward the target. The trial ended earlier if any target, i.e., an edge of the screen, was reached. The position of the cursor, serving as visual feedback, was determined using an autoregressive (AR) model of order 16, and was updated every 40 ms. This AR model serves as the baseline for our investigations, representing a widely established online decoding method.

B. Model

We employ the BaseNet [31] architecture, which can be considered a modern evolution of the shallow convolutional neural networks ShallowNet [32] and EEGNet [33]. To make the architecture suitable for online decoding, we use

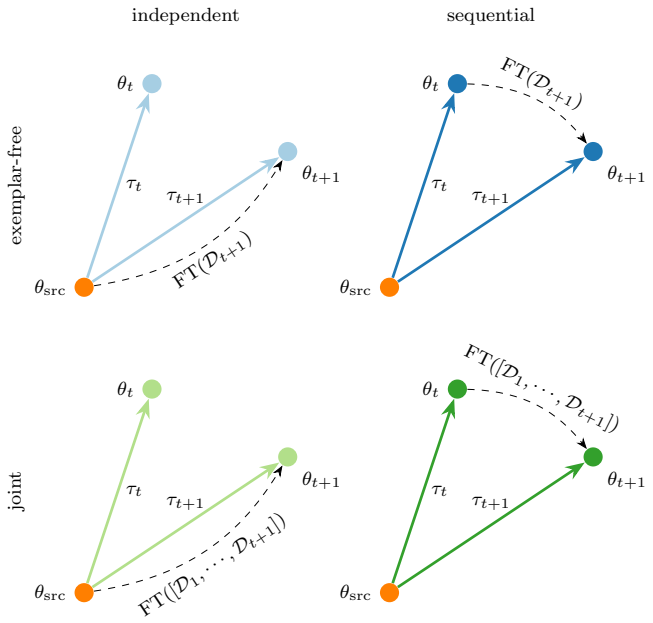


Fig. 1: Fine-tuning process and task vectors across different settings in a two-dimensional weight space. Solid arrows represent task vectors, dashed lines illustrate the fine-tuning trajectory.

real-time adaptive pooling (RAP) [34], which modifies the pooling layers to enable the decoding of sliding windows. We select sliding windows of length 1 s (as done in, e.g., [19], [30], [34]) and an update frequency of 25 Hz to match the original online setting. This leads to the RAP parameters $k_1 = s_1 = 5, k_2 = 50, s_2 = 2$. The complete source code is available on GitHub [1].

C. Training strategies

Each subject participates in up to 11 sessions, with each session containing $N_t = 150$ trials X_i per paradigm paired with corresponding labels y_i. The dataset for a single session is represented as $\mathcal{D}^{\text{subject}}_{\text{session}} = \{(X_i, y_i)\}_{i=1}^{N_t}$. During supervised pre-training, we employ a cross-subject leave-one-subject-out strategy to learn subject-invariant representations as done in [35], providing an effective initialization for subsequent subject-specific fine-tuning. The source dataset is constructed by aggregating data from the first session of the remaining $N-1 = 60$ subjects, denoted as $\mathcal{D}^{\text{i}}_{\text{source}} = \bigcup_{j\in\{1,\ldots,N\}\setminus\{i\}} \mathcal{D}^{j}_{1}$. This approach enables calibration-free online decoding for unseen subjects by leveraging data from the other participants. Moreover, it reflects a practical scenario in which only limited data from multiple subjects is available at the initial stage of source model training. Subsequent subject-specific applications can then be highly personalized.

After pre-training, the model undergoes supervised fine-tuning using subject-specific data under a causal constraint, ensuring that only data recorded prior to the test session is used for fine-tuning. This fine-tuning step is essential as it enables the model to adapt to subject-specific patterns, which can vary widely across users. By refining the pre-trained, subject-invariant model with personalized data, the model becomes better suited to each user's unique characteristics, resulting in improved decoding performance.

We investigate two distinct data settings typically examined in continual learning: *exemplar-free* and *joint*. In the exemplar-free setting, only the data from the prior session is used for fine-tuning, whereas the joint setting incrementally incorporates data from all prior sessions.

Additionally, we evaluate two fine-tuning strategies: *independent* and *sequential*. The independent strategy reinitializes fine-tuning from the pre-trained source model for each new session, while the sequential strategy builds upon the most recently fine-tuned model of the target subject. Combining both, there are four different settings in total: **exemplar-free independent**, **exemplar-free sequential**, **joint independent** and **joint sequential**, which are compared to the **baseline** from [26] and the non-adapted, subject-invariant **source** model.

D. Task vector notation

A widely used representation for describing the variations in fine-tuned models is the task vector notation [36]. For simplicity, we will omit the target subject index in the following explanation. Here, θ_{src} represents the model weights after pre-training on the source data, while θ_t denotes the weights following the t-th fine-tuning iteration. The task vector is defined as $\tau_t = \theta_t - \theta_{\text{src}}$, which specifies a direction within the weight space. The task vector notation of the four different settings together with the corresponding fine-tuning trajectory is visualized in Fig. 1.

Following the causal constraint of only using the previous sessions for fine-tuning, the fine-tuned weights θ_t are evaluated using the datset $\mathcal{D}_{t+1}$ from the session $t+1$.

E. Test-time adaptation

As rebiasing the decoder between different domains [34], [37] is a very important step in MI but recording additional calibration data for each new domain is costly [16], we perform online test-time adaptation (OTTA) [23], [24]. Specifically, we use Euclidean alignment (EA) [38], [39] in an online fashion [34] to mitigate the domain shift of the input data between sessions. Additionally, we use online Adaptive batch norm (AdaBN) [40], [41] to account for the shifts in the batch normalization statistics in the intermediate layers of the model. Both adaptation processes are carried out using only the current (sliding) window of the target session, making this a single-sample OTTA approach. The use of OTTA makes it possible to dispense domain-specific calibration data without losing performance. Another advantage is that, since as we only account for the overall distribution shift, the decision boundary does not change within one session, ensuring stability and thus facilitating user learning [18].

[1] https://github.com/martinwimpff/eeg-continual

F. Metrics

To compare our different approaches, we use the trial-wise accuracy [29] as our primary metric. This means that a trial is considered to be successful if more than 50 % of all windows of that trial are correctly classified. For simplicity, we will refer to this as accuracy in the following.

Reported single values with standard deviations correspond to the mean and standard deviation of individual subject performances. This is achieved by first averaging sessions per subject, ensuring equal representation regardless of the number of sessions completed by each subject.

To compare the similarity between task vectors, we employ the cosine distance $d(\tau_i, \tau_j) = 1 - \frac{\tau_i \cdot \tau_j}{||\tau_i|| \cdot ||\tau_j||} \in [0, 1]$.

III. RESULTS & DISCUSSION

Figures 2a and 2b present the session-wise test accuracy for the different approaches corresponding to the LR and UD paradigms, respectively. For the baseline and our source model, which remain constant (apart from rebiasing) across the sessions for each subject, the performance increase over time can only be attributed to the user exhibiting more discriminative patterns over time. It is, however, worth noting that as our experiments are carried out in a pseudo-online fashion [25], potential user adaptation to our decoders can not be explicitly examined.

Both figures clearly demonstrate that incorporating decoder adaptation enhances the average test accuracy and increases the extent of performance improvement over time. This trend is indicative of successful decoder adaptation. We speculate that in an online experiment, the improvement over time could be even greater, as users would have the opportunity to adjust to the adapting decoder.

The average performance, together with the pairwise p-values (paired two-sided t-test against baseline and source model), are displayed in Fig. 3. For the LR paradigm, all our models outperform the baseline, and all fine-tuning approaches outperform the source model with $p < 0.001$. For the UD paradigm, the source and baseline are not statistically different ($p = 0.748$), and the difference between the baseline and the exemplar-free independent setting is smaller ($p = 0.0124$).

These findings confirm the decoder adaptation's effectiveness while highlighting key differences between the approaches. Notably, leveraging previously acquired subject-specific knowledge, whether explicitly through joint fine-tuning or implicitly by using the previously fine-tuned decoder, leads to improvements in overall performance.

The joint sequential setting achieves the highest accuracy, as it benefits from cumulative progress by reusing the previously fine-tuned model, in contrast to the independent approach. Furthermore, this setting ensures stable fine-tuning by incorporating all previously recorded subject-specific data, unlike the exemplar-free approach.

To further understand the relationship between the fine-tuning approaches, we visualize the session-wise test accuracy for each fine-tuned model in Fig. 4. In each matrix, the first column represents the source performance, while the diagonal entries on the right correspond to the fine-tuning results reported in Fig. 2. Since we adhere to a strictly causal data setting, sessions are not evaluated using models fine-tuned with data recorded in later sessions. Thus, these entries are marked with an X.

For three out of four settings, we observe a clear trend: performance improves with both the progression of sessions (user learning) and increased fine-tuning steps (decoder adaptation). When comparing rows, selecting the most recently fine-tuned model (i.e., the field on the right) exhibits the highest test accuracy.

In the exemplar-free independent scenario, these trends are still apparent but show a diminished degree of improvement across sessions and fine-tuning steps. We attribute this to reduced stability, as fine-tuning relies only on data from the most recent session, which can vary significantly due to factors such as user concentration, motivation, or environmental influences. Consequently, a single session may not adequately represent the current state of the user. Moreover, in this scenario, the decoder cannot leverage previously acquired subject-specific knowledge and must begin adapting to the user from scratch for each new session, limiting its ability to achieve cumulative progress.

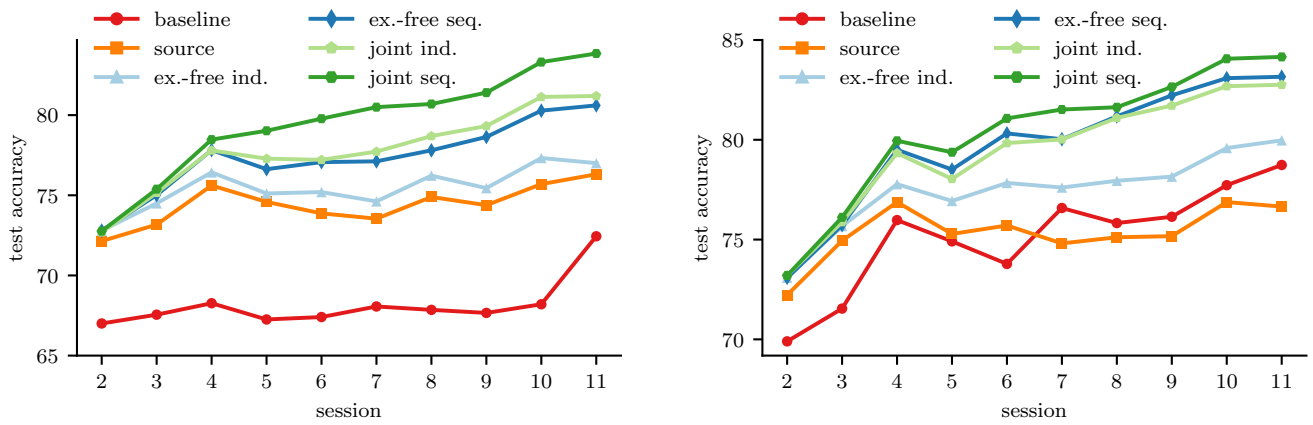
The displayed matrices represent averages across subjects and paradigms, and subject-specific matrices might deviate from the overall trend. Thus, to further assess the validity of selecting the most recently fine-tuned decoder, we conducted a theoretical experiment. For each subject, we identified the best-performing decoder per session to establish a theoretical upper bound. For Fig. 4, this corresponds to picking the highest value per row in each subject-specific matrix.

TABLE I: Theoretical upper bounds for each setting and paradigm, together with the performance of the most recently fine-tuned decoder.

setting	LR		UD	
	most recent	upper bound	most recent	upper bound
ex.-free ind.	75.1 ± 13.6	76.7 ± 13.3	77.1 ± 12.6	78.8 ± 12.4
ex.-free seq.	76.9 ± 13.2	78.0 ± 12.9	79.1 ± 12.3	80.2 ± 12.1
joint ind.	77.2 ± 13.4	77.9 ± 13.2	78.8 ± 12.8	79.9 ± 12.4
joint seq.	78.8 ± 12.6	79.5 ± 12.4	79.8 ± 12.3	80.9 ± 12.0

These upper bounds are presented in Table I for each setting and paradigm, together with the value for picking the most recently fine-tuned decoder. P-values are omitted, as all comparisons are statistically significant with $p < 0.001$. As expected, our strategy performs worse than the theoretical upper bound. However, the overall performance difference is surprisingly small, except for the exemplar-free independent setting. Encouragingly, this suggests that the strategy of selecting the most recently fine-tuned model closely approximates the theoretical optimal decoder selection. In the exemplar-free setting, which still has the lowest upper bound among the settings, adaptation is less stable, making it beneficial to switch between different stages of fine-tuning.

To examine the stability between fine-tuned models more



(a) LR paradigm

(b) UD paradigm

Fig. 2: Accuracy across the sessions for each fine-tuning strategy compared to the source model and the baseline. Each dot represents the test accuracy of a single session, averaged across all subjects. The x-axis denotes session progression, while the y-axis represents the test accuracy (%).

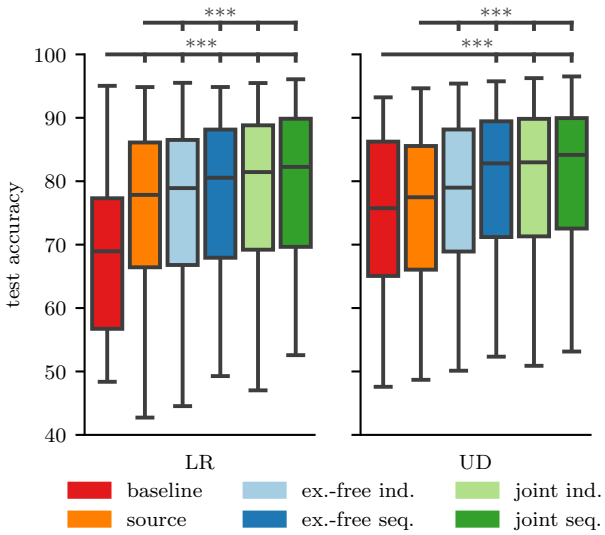


Fig. 3: Average test accuracy over all sessions (2 - 11) per setting. Stars above the brackets indicate a significance level ($p < 0.001$ (***)) when compared to the baseline and source model, respectively.

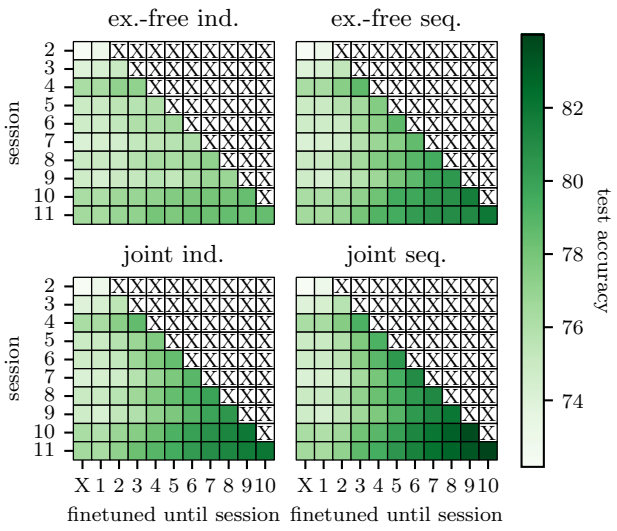


Fig. 4: Average test accuracy (over all subjects and both paradigms) for each fine-tuning setting. The X on the x-axis refers to no fine-tuning, i.e., using the source model.

closely, we calculated the cosine distance between task vectors, as shown in Fig. 5. The distance matrix for the exemplar-free independent scenario supports our earlier conclusions. Since fine-tuning restarts from scratch for each session, the process is less stable, resulting in larger distances between consecutive task vectors compared to the other three settings. Nevertheless, the cosine distance remains well below 1, indicating a degree of relationship and some level of stability between consecutive task vectors. In contrast, for task- or class-incremental fine-tuning, the cosine distance approaches one [36], signifying nearly orthogonal task vectors.

For the other three settings, especially the later task vectors exhibit greater similarity to one another, suggesting increased stability between fine-tuning steps, which tends to benefit user learning [18], [19]. Additionally, across all four settings, the distance to the first task vector is noticeably larger. This may be attributed to greater data discrepancies, given the 8-week gap between the first two sessions compared to only a few days between each of the subsequent ones [26].

The previously presented results demonstrate the advantages of training models in a joint sequential manner. However, since the joint data setting increases the memory and time requirements for fine-tuning, we examined how the

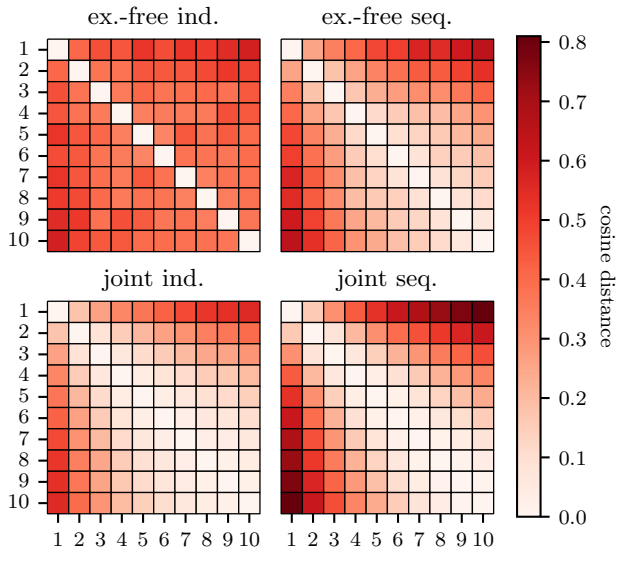


Fig. 5: Cosine distance between task vectors for all strategies.

TABLE II: Ablation study for the number of previous sessions used during sequential fine-tuning.

# previous sessions	LR		UD	
	accuracy	p-value	accuracy	p-value
joint	78.8 ± 12.6	*X*	79.8 ± 12.3	*X*
4	78.7 ± 12.5	0.038	80.0 ± 12.1	0.036
3	78.3 ± 12.7	< 0.001	80.0 ± 12.1	0.231
2	77.7 ± 12.8	< 0.001	79.8 ± 12.1	0.974
exemplar-free	76.9 ± 13.2	< 0.001	79.1 ± 12.3	0.013

number of previous sessions used during sequential fine-tuning influences performance. The results are shown in Table II, with p-values calculated against the joint setting (first row). For the LR paradigm, performance differences are significant across all data settings, though a trend emerges where accuracy approaches the joint setting when incorporating four previous sessions. In contrast, for the UD paradigm, the differences are generally smaller. Interestingly, fine-tuning with only three or four previous sessions slightly outperforms the joint setting. We speculate that this may be due to greater user learning over time (see Figure 2b baseline), where older sessions could hinder rapid adaptation to new data. Nonetheless, the differences are still small, and this hypothesis is speculative as we can not examine the user's behavior in the different settings. Therefore, we still recommend using the joint setting for its stability. However, if memory constraints or a large number of sessions become a concern, this setting could likely be relaxed without a (large) loss in performance.

TABLE III: OTTA ablation study for the source model. P-values (paired two-sided t-test) calculated against the first row.

EA	AdaBN	LR		UD	
		accuracy	p-value	accuracy	p-value
✓	✓	74.0 ± 13.8	*X*	75.2 ± 13.0	*X*
✗	✓	64.3 ± 13.5	< 0.001	64.9 ± 10.5	< 0.001
✓	✗	71.0 ± 13.1	0.061	69.3 ± 15.0	< 0.001
✗	✗	57.9 ± 11.2	< 0.001	59.9 ± 9.90	< 0.001

To eliminate the need for session-specific calibration data while maintaining performance, we utilize OTTA throughout this study, which integrates online EA and online AdaBN. To evaluate the individual contributions of these components, we conducted an ablation study using the source model. The results are presented in Table III, where p-values are calculated against the fully enabled configuration (first row).

Notably, disabling EA alone contributes to a $\sim 10\%$ loss in performance compared to the source model with both components active. Similarly, disabling AdaBN also decreases performance, though its statistical significance ($p < 0.05$) is only observed for the UD paradigm. This may be due to the generally higher level of user learning observed in the UD paradigm, meaning the impact of OTTA is more pronounced when the distribution shift is greater. Ultimately, while EA has a more substantial effect, both components play a crucial role in ensuring reliable and robust decoding performance.

IV. CONCLUSION

This study explored various fine-tuning strategies for pseudo-online longitudinal EEG MI decoding in a causal setting for a large user group. We investigated the impact of different training strategies to incorporate previously acquired subject-specific knowledge implicitly or explicitly during fine-tuning with various training designs. The results demonstrate that leveraging previously acquired subject-specific information enhances performance and improves stability. Overall, the most effective strategy, *joint sequential* fine-tuning, involved incorporating all previously recorded subject-specific data while continuously building on the last subject-specific fine-tuned model.

Furthermore, the use of OTTA enables calibration-free operation for new sessions or subjects, making our approach well-suited for real-world applications and directly integrable into both applicative and experimental contexts.

Although these results are obtained through offline analysis, we make every effort to replicate a pseudo-online setting that closely approximates real conditions. While our experimental analysis only accounts for one half of the *two-learner* system, we are confident that these results are robust and will hold in online experiments as this study is conducted on a large user group and across several sessions.

REFERENCES

[1] J. Peksa and D. Mamchur, "State-of-the-art on brain-computer interface technology," *Sensors*, vol. 23, p. 6001, 2023.

[2] M. Cervera, S. Soekadar, J. Ushiba, J. Millán, M. Liu, N. Birbaumer, and G. Garipelli, "Brain-computer interfaces for post-stroke motor rehabilitation: a meta-analysis," *Annals of Clinical and Translational Neurology*, vol. 5, pp. 651–663, 2018.

[3] B. Society, "BCI definition," 2024. [Online]. Available: https://bcisociety.org/bci-definition/

[4] J. Decety, "The neurophysiological basis of motor imagery," *Behavioural Brain Research*, vol. 77, pp. 45–52, 1996.

[5] C. Sannelli, C. Vidaurre, K. Müller, and B. Blankertz, "A large scale screening study with a SMR-based BCI: Categorization of BCI users and differences in their SMR activity," *PloS One*, vol. 14, p. e0207351, 2019.

[6] R. Zhang, F. Li, T. Zhang, D. Yao, and P. Xu, "Subject inefficiency phenomenon of motor imagery brain-computer interface: Influence factors and potential solutions," *Brain Science Advances*, vol. 6, pp. 224–241, 2020.

[7] M. Lee, O. Kwon, Y. Kim, H. Kim, Y. Lee, J. Williamson, S. Fazli, and S. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, 2019.

[8] A. Gaume, A. Vialatte, A. Mora-Sánchez, C. Ramdani, and F. Vialatte, "A psychoengineering paradigm for the neurocognitive mechanisms of biofeedback and neurofeedback," *Neuroscience & Biobehavioral Reviews*, vol. 68, pp. 891–910, 2016.

[9] S. Kober, M. Witte, M. Ninaus, C. Neuper, and G. Wood, "Learning to modulate one's own brain activity: the effect of spontaneous mental strategies," *Frontiers in Human Neuroscience*, vol. 7, p. 695, 2013.

[10] C. Jeunet, E. Jahanpour, and F. Lotte, "Why standard brain-computer interface (BCI) training protocols should be changed: an experimental study," *Journal of Neural Engineering*, vol. 13, no. 3, p. 036024, 2016.

[11] M.-C. Corsi, M. Chavez, D. Schwartz, N. George, L. Hugueville, A. E. Kahn, S. Dupont, D. S. Bassett, and F. D. V. Fallani, "Functional disconnection of associative cortical areas predicts performance during BCI training," *NeuroImage*, vol. 209, p. 116500, 2020.

[12] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[13] J. Millán, "Brain-machine interfaces: the perception-action closed loop: a two-learner system," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, pp. 6–8, 2015.

[14] J. Müller, C. Vidaurre, M. Schreuder, F. Meinecke, P. Von Bünau, and K. Müller, "A mathematical model for the two-learners problem," *Journal of Neural Engineering*, vol. 14, p. 036005, 2017.

[15] S. Perdikis and J. Millan, "Brain-machine interfaces: a tale of two learners," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, pp. 12–19, 2020.

[16] Z. Rao, R. Zhang, S. He, Y. Zhou, Z. Lu, K. Li, and Y. Li, "A once-calibration brain-computer interface to enhance convenience for continuous BCI interventions in stroke patients," *IEEE Sensors Journal*, 2024.

[17] L. Jaeger, R. Baptista, C. Basla, P. Capsi-Morales, Y. Kim, S. Nakajima, C. Piazza, M. Sommerhalder, L. Tonin, G. Valle, and Others, "How the cybathlon competition has advanced assistive technologies," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, pp. 447–476, 2023.

[18] S. Perdikis, L. Tonin, S. Saeedi, C. Schneider, and J. Millán, "The cybathlon BCI race: Successful longitudinal mutual learning with two tetraplegic users," *PLoS Biology*, vol. 16, p. e2003787, 2018.

[19] D. Forenzo, H. Zhu, J. Shanahan, J. Lim, and B. He, "Continuous tracking using deep learning-based decoding for noninvasive brain–computer interface," *PNAS Nexus*, vol. 3, p. 145, 2024.

[20] S. Tortora, G. Beraldo, F. Bettella, E. Formaggio, M. Rubega, A. Del Felice, S. Masiero, R. Carli, N. Petrone, E. Menegatti, and Others, "Neural correlates of user learning during long-term BCI training for the cybathlon competition," *Journal of NeuroEngineering and Rehabilitation*, vol. 19, p. 69, 2022.

[21] G. Ven, T. Tuytelaars, and A. Tolias, "Three types of incremental learning," *Nature Machine Intelligence*, vol. 4, pp. 1185–1197, 2022.

[22] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[23] S. Li, Z. Wang, H. Luo, L. Ding, and D. Wu, "T-TIME: Test-time information maximization ensemble for plug-and-play BCIs," *IEEE Transactions on Biomedical Engineering*, 2023.

[24] M. Wimpff, M. Döbler, and B. Yang, "Calibration-free online test-time adaptation for electroencephalography motor imagery decoding," in *2024 12th International Winter Conference on Brain-Computer Interface (BCI)*, 2024, pp. 1–6.

[25] I. Carrara and T. Papadopoulo, "Pseudo-online framework for BCI evaluation: a MOABB perspective using various MI and SSVEP datasets," *Journal of Neural Engineering*, vol. 21, p. 016003, 2024.

[26] J. Stieger, S. Engel, and B. He, "Continuous sensorimotor rhythm based brain computer interface learning in a large population," *Scientific Data*, vol. 8, p. 98, 2021.

[27] D. Gwon, K. Won, M. Song, C. Nam, S. Jun, and M. Ahn, "Review of public motor imagery and execution datasets in brain-computer interfaces," *Front in Hum Neurosc*, vol. 17, p. 1134869, 2023.

[28] M. Tangermann, K. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Müller-Putz, and Others, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012.

[29] P. Dreyer, A. Roc, L. Pillette, S. Rimbert, and F. Lotte, "A large eeg database with users' profile information for motor imagery brain-computer interface research," *Scientific Data*, vol. 10, p. 580, 2023.

[30] J. Stieger, S. Engel, D. Suma, and B. He, "Benefits of deep learning classification of continuous noninvasive brain–computer interface control," *Journal of Neural Engineering*, vol. 18, p. 046082, 2021.

[31] M. Wimpff, L. Gizzi, J. Zerfowski, and B. Yang, "EEG motor imagery decoding: A framework for comparative analysis with channel attention mechanisms," *Journal of Neural Engineering*, vol. 21, p. 036020, 2024.

[32] R. Schirrmeister, J. Springenberg, L. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, pp. 5391–5420, 2017.

[33] V. Lawhern, A. Solon, N. Waytowich, S. Gordon, C. Hung, and B. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, p. 056013, 2018.

[34] M. Wimpff, J. Zerfowski, and B. Yang, "Tailoring deep learning for real-time brain-computer interfaces: From offline models to calibration-free online decoding," 2024.

[35] C. labs at Reality Labs, D. Sussillo, P. Kaifosh, and T. Reardon, "A generic noninvasive neuromotor interface for human-computer interaction," *bioRxiv*, pp. 2024–02, 2024.

[36] G. Ilharco, M. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," *ArXiv Preprint*, vol. ArXiv:2212.04089, 2022.

[37] S. Kumar, F. Yger, and F. Lotte, "Towards adaptive classification using riemannian geometry approaches in brain-computer interfaces," in *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 2019, pp. 1–6.

[38] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A euclidean space data alignment approach," *IEEE Transactions on Biomedical Engineering*, vol. 67, pp. 399–410, 2019.

[39] B. Junqueira, B. Aristimunha, S. Chevallier, and R. Y. de Camargo, "A systematic evaluation of euclidean alignment with deep learning for EEG decoding," *Journal of Neural Engineering*, vol. 21, no. 3, p. 036038, 2024.

[40] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *ArXiv Preprint*, vol. ArXiv:1603.04779, 2016.

[41] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 11 539–11 551.