TASER: TRIAL-AWARE SHIFT ESTIMATION AND RE-CENTERING FOR CAUSAL EEG TEST-TIME ADAPTATION IN BRAIN-COMPUTER INTERFACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) decoders are often adapted one window at a time. Yet overlapping windows are repeated views of the same trial, so one atypical trial can move the adapter many times. We introduce TASER, a causal and label-free method that adapts once per brain–computer interface (BCI) acquisition unit: a trial for motor imagery and SSVEP, or a stimulus group for ERP/P300. A frozen encoder maps the signal into a latent space, where source prototypes define the class geometry. TASER records the current prediction before adaptation, pools the associated embeddings into one observation, and estimates a shared shift between source and target representations. Window redundancy and within-trial dispersion determine the update strength. The shift then re-centers the prototypes for the next trial, without target labels, pseudo-labels, replay, or backpropagation. Across 13 datasets, 153 participants, and four encoder seeds, TASER reaches 70.00% dataset-balanced accuracy, compared with 68.61% for the closest baseline. It ranks first on eight datasets and second on three, with a median online latency of 1.25 ms per decision. Updating after every window instead of once per trial reduces accuracy by 5.29 points; removing the shift costs 5.53 points. The ranking persists across three temporal resolutions. Therefore, these results indicate that respecting the BCI acquisition unit can make causal EEG adaptation both more accurate and less expensive.

1 INTRODUCTION

Brain–Computer Interfaces (BCI) are systems capable of inferring a user's intent from human brain patterns, measured through non-invasive electroencephalography (EEG), and translating it into commands for an external device (Wolpaw et al., 2002; Lotte et al., 2018). Their applications include neurorehabilitation (Vourvopoulos et al., 2023; Jin et al., 2024), communication (Khan et al., 2025), robotics (Tonin & Millán, 2021), and virtual reality (Wen et al., 2021). However, despite years of research, the translational impact of BCI remains limited by two persistent transfer problems: first, a BCI decoder trained on the data of one or more subjects cannot be directly applied to the data of a new subject due to high inter-subject variability of the EEG signal. Researchers conventionally collect calibration data from the specific subject to train the decoder, which could be time-consuming and frustrating. Second, the performance of the decoder after the training process may decline over time due to the highly non-stationary nature of the EEG, as a consequence of differences in physiology, electrode placement, impedance, and cognitive state (Vidaurre et al., 2011; Wimpff et al., 2023). A useful BCI must therefore adapt quickly to covariate shift of data distribution while continuing to produce accurate predictions.

Test-time adaptation (TTA) addresses this setting by adjusting a source-trained model during deployment from the incoming target stream, without introducing a new supervised calibration stage. Unlike other machine learning problems, BCI paradigms make this setting distinct from an ordinary stream of independent samples. BCI recordings are generally organized into *trials*, each corresponding to one task instance: an imagined movement in motor imagery, attention to a flickering stimulus in steady-state visual evoked potential (SSVEP) decoding, or a stimulus-locked response in an event-related potential (ERP)/P300 system (Pfurtscheller & Neuper, 2001; Tangermann et al., 2012; Lotte et al., 2018). Motor-imagery and SSVEP trials are divided into overlapping windows for

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.

temporally resolved decoding, whereas an ERP/P300 selection groups multiple time-locked stimulus events. To approximate online deployment using existing recordings, we replay each session in its original acquisition order. At trial t, or ERP/P300 selection t, the current signal is available for prediction, whereas the adapter state may depend only on the source data and target observations preceding t. The state can incorporate the current trial only after its prediction has been recorded. This predict-then-update protocol preserves causal order while enabling reproducible evaluation (Carrara & Papadopoulo, 2024). The central question is therefore how to extract a stable update from a complete EEG trial while preserving this causal order. We argue that the update should respect the same unit as the BCI decision: the trial, not each window in isolation.

Transfer learning in EEG-based BCIs aims to reduce subject-specific calibration by reusing information from other participants or sessions. Classical approaches adapt compact feature statistics or linear decision rules. PMean updates the pooled feature mean of an LDA classifier from unlabeled target data (Vidaurre et al., 2011), while MPMLDA extends adaptive discriminant analysis to multiclass BCIs (Llera et al., 2014). EDAPT addresses a complementary supervised setting by continually fine-tuning the decoder from target feedback (Haxel et al., 2025).

Recent EEG-specific TTA methods move adaptation into deep decoders. EEG-OTTA combines online alignment, adaptive normalization, and entropy minimization (Wimpff et al., 2023); T-TIME uses information maximization (Li et al., 2024); PLTA maintains a pseudo-labeled feature bank (Lin et al., 2026); and BFT avoids backpropagation by aggregating transformed views or feature subnetworks (Li et al., 2026). SPDIM models conditional and label shifts in a positive-definite geometry (Li et al., 2025), whereas BTTA-DG uses probabilistic feature projection and inference (Luo et al., 2026). Their protocols differ in supervision, target access, online computation, and retained state, which complicates direct comparison under a causal stream.

TTA research outside EEG provides complementary adaptation mechanisms. Tent updates normalization parameters by minimizing prediction entropy (Wang et al., 2021); NOTE introduces a balanced memory for temporally correlated streams (Gong et al., 2022); RoTTA combines memory with teacher-student updates (Yuan et al., 2023); and PETAL models uncertainty during continual adaptation (Brahma & Rai, 2023). Gradient-free alternatives include T3A, which constructs class templates from pseudo-labels (Iwasawa & Matsuo, 2021), and LAME, which refines predictions through local target geometry (Boudiaf et al., 2022). NEO keeps the model fixed and re-centers its latent representation using cumulative target statistics (Murphy et al., 2026). PCOC/MOCA instead updates class distributions from labeled input-class pairs and therefore belongs to a neighboring supervised setting (Harrison et al., 2020). For BCI deployment, these choices matter: weight updates and pseudo-labels add flexibility but may increase computation and propagate prediction errors, whereas cumulative statistics are inexpensive and stable but respond more slowly to recent shifts.

We introduce **TASER**, a **T**rial-**A**ware **S**hift **E**stimation and **R**e-Centering method designed around the acquisition unit of the BCI itself. TASER keeps fixed both the encoder, trained on pooled EEG data from source participants distinct from the target BCI user, and the class geometry derived from those source data. To predict an incoming trial of a new unseen subject, it embeds all of its windows and jointly compares their evidence with source-derived temporal prototypes. Once the prediction has been recorded, the trial is compressed into a single latent observation. Its mean estimates how the target representation has moved, while its dispersion and effective number of overlapping windows determine how strongly that observation should update a shared translation of the source prototypes. The updated translation is then available to the following trial, thus adapting the BCI decoder over time (Figure 1). TASER consequently adapts the decoder at the level at which EEG decisions are made -i.e., the trial-instead of treating every overlapping window as a separate adaptation step. The encoder and classifier parameters never change; additionally, deployment requires no target labels, pseudo-labels, replay buffer, or backward pass. The only evolving state is one latent shift vector and one uncertainty scalar, so the adaptation cost remains constant as the session grows.

TASER is one part of the contribution. Controlled ablations separately test whether temporal prototypes help, whether the state should be updated once per trial or after every window, whether adaptation should emphasize recent or all previous trials, and whether updating the shared latent shift helps at all. We then evaluate causal EEG adaptation across motor imagery, ERP/P300, and SSVEP, using multiple datasets, and random seeds. The resulting benchmark compares classical EEG adaptation, gradient-free TTA, and optimization-based TTA under the same predict-then-update constraint, and reports predictive performance together with latency and memory.

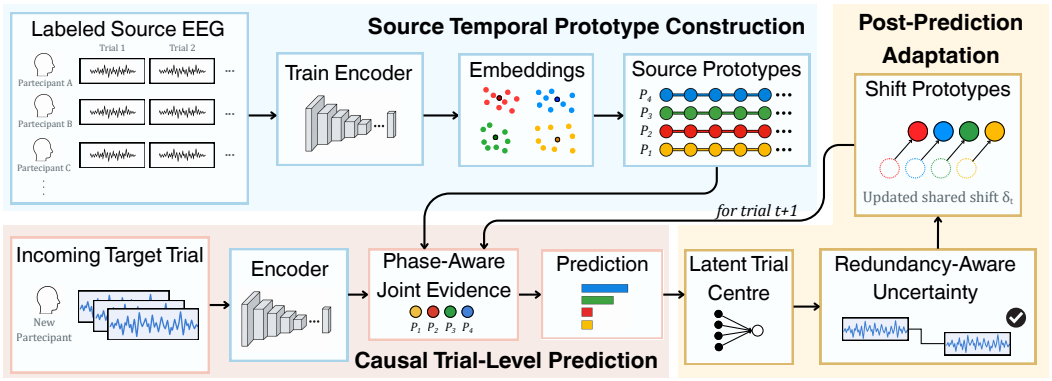


Figure 1: TASER's source and causal target stages. Labeled source trials produce a frozen encoder and ordered class-phase prototypes. At target trial t, phase-matched distances from all overlapping-window embeddings are aggregated into one prediction using only the prior state. After prediction, the trial center estimates a shared shift, and overlap-aware uncertainty controls its update to δ_t. This translation preserves the source geometry and is first used at trial $t+1$.

2 METHOD

We treat a complete trial, rather than an overlapping window, as the unit of prediction and adaptation. Let $\mathcal{S}$ denote the source participants and C the number of decoding classes. Participants are indexed by s, trials by i, windows within a trial by j, and classes by c. An EEG window $x_{i,j} \in \mathbb{R}^{E\times L}$ contains E channels and L temporal samples; its latent representation has dimension D. At deployment, the target trial $\mathcal{T}_t = \{x_{t,j}\}_{j=1}^{M_t}$ contains M_t ordered windows. The protocol is strictly predict–then–update: trials $1,\dots,t-1$ determine the state used to predict $\mathcal{T}_t$, and trial t may affect only later decisions. TASER follows three stages: source temporal prototype construction, causal trial-level prediction, and post-prediction adaptation.

2.1 SOURCE TEMPORAL PROTOTYPE CONSTRUCTION

For each target participant, the remaining participants form the source pool. We divide their data into two partitions. Source-fitting trials are used to train the supervised classifier and compute the class prototypes. Source-selection trials preserve complete ordered sessions and are used to select the encoder checkpoint, construct the class–segment prototypes, and choose K. Neither partition contains data from the target participant.

Frozen encoder and class prototypes. We first train a supervised classifier $h_{\theta,W,b}(x) = Wf_\theta(x) + b$ on the source-fitting trials. Here, $f_\theta : \mathbb{R}^{E\times L} \to \mathbb{R}^D$ maps an EEG window to its latent representation, while W and b define the final linear classification head. After selecting the source checkpoint, we remove this head and freeze f_θ. TASER therefore operates on the embeddings $z_{i,j} = f_\theta(x_{i,j}) \in \mathbb{R}^D$ without modifying the encoder during deployment. Throughout this section, *prototype* means the centroid of the source embeddings assigned to a class. It is a point in latent space, not a trainable parameter. If $\mathcal{I}_c$ contains the source-fitting trial–window pairs from class c, the class prototype $p_c \in \mathbb{R}^D$ and balanced prototype center $\bar{p} \in \mathbb{R}^D$ are

$$p_c = \frac{1}{|\mathcal{I}_c|}\sum_{(i,j)\in\mathcal{I}_c} z_{i,j}, \quad \bar{p} = \frac{1}{C}\sum_{c=1}^{C} p_c. \tag{1}$$

We use $\bar{p}$, rather than the sample-weighted source mean, so that class frequency does not set the origin of the online correction.

Class–segment prototypes. A class prototype averages all source embeddings assigned to that class, regardless of their temporal position within a trial. This removes systematic differences between the beginning, middle, and end of the trial. TASER retains this information by dividing each

source-selection trial into K equal relative temporal segments and computing one prototype per class and segment. The index $k \in \{1, \ldots, K\}$ identifies the normalized temporal segment. For a trial with M_i windows, the normalized location of window j determines its segment:

$$q_{i,j} = \frac{j-1}{M_i} \in [0, 1), \quad a_K(j; M_i) = 1 + \lfloor K q_{i,j} \rfloor . \tag{2}$$

Because the segments cover equal relative portions, trials need not contain the same number of windows. Let $\mathcal{E}$ denote all source-selection trials, let y_i be the known class of trial i, and define $\mathcal{E}_c = \{i \in \mathcal{E} : y_i = c\}$. The set $\mathcal{J}_{i,k}^{(K)} = \{j : a_K(j; M_i) = k\}$ contains the windows of trial i assigned to segment k. Averaging first within a trial segment and only then across trials gives

$$m_{i,k}^{(K)} = \frac{1}{|\mathcal{J}_{i,k}^{(K)}|} \sum_{j \in \mathcal{J}_{i,k}^{(K)}} z_{i,j}, \quad \widetilde{p}_{c,k}^{(K)} = \frac{1}{|\mathcal{E}_c|} \sum_{i \in \mathcal{E}_c} m_{i,k}^{(K)}. \tag{3}$$

This order gives each trial the same influence, regardless of its window count. The resulting ordered sequence of segment prototypes is then re-centered on the class prototype from Equation 1. First, we compute its center,

$$\widetilde{\mu}_c^{(K)} = \frac{1}{K} \sum_{h=1}^{K} \widetilde{p}_{c,h}^{(K)}. \tag{4}$$

Each segment prototype is then translated by the difference between this center and p_c:

$$p_{c,k}^{(K)} = p_c + \left(\widetilde{p}_{c,k}^{(K)} - \widetilde{\mu}_c^{(K)} \right). \tag{5}$$

The term in parentheses retains the segment-specific deviation, while adding p_c restores the class center. Consequently, $K^{-1} \sum_{k=1}^{K} p_{c,k}^{(K)} = p_c$. Thus, K controls how finely within-trial changes are represented without changing the class prototype.

Source selection of temporal resolution. Let $\mathcal{K}$ be the source-fixed set of candidate segment counts specified in Section 3. We discard values larger than the shortest source-selection trial so that every segment contains at least one window. We then hold out each source participant s in turn. Its source-fitting trials are excluded when rebuilding the class prototypes, and its source-selection trials are excluded when constructing the class–segment prototypes. The source-trained encoder remains fixed and is not retrained inside this inner loop. We replay the complete selection session of participant s in acquisition order. Each trial is classified using the causal trial-level procedure described below and is assimilated only after its prediction has been recorded. Source labels are used afterward to compute the trial-level balanced accuracy $B_s(K)$. We average across source participants before selecting K:

$$\bar{B}(K) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} B_s(K), \quad K^* = \min \left(\arg\max_{K \in \mathcal{K}} \bar{B}(K) \right). \tag{6}$$

Once K^* has been selected, we rebuild $p_{c,k}^{(K^*)}$ from all source participants and freeze them. Only source data and labels are used in this construction; no data or labels from the target participant are employed.

2.2 Causal Trial-Level Prediction

After source construction, the encoder, class–segment prototypes, and selected K^* are frozen. TASER can then decode the unseen target participant. Before the first target trial, the shared translation is initialized as $\delta_0 = 0$. For $t > 1$, $\delta_{t-1} \in \mathbb{R}^D$ depends only on trials that have already been predicted. This translation is applied to every class–segment prototype:

$$p_{c,k}^{(t)} = p_{c,k}^{(K^*)} + \delta_{t-1}. \tag{7}$$

Translation leaves the learned geometry intact: neither class–segment relationships nor pairwise distances change.

All windows in the target trial $\mathcal{T}_t$ contribute to one decision. For each class, window j is matched to its relative segment. The resulting trial-to-class dissimilarity and prediction are

$$D_t(c) = \frac{1}{M_t} \sum_{j=1}^{M_t} \left\| z_{t,j} - p^{(t)}_{c,a_{K^*}(j;M_t)} \right\|_2^2, \quad \widehat{y}_t = \arg\min_c D_t(c). \tag{8}$$

Here, $D_t(c)$ measures how closely the complete target trial matches class c. All windows contribute to the same class decision; TASER does not classify them separately and then apply majority voting. When $K^* = 1$, the rule reduces to the shifted source prototype nearest to the trial mean.

2.3 Post-Prediction Adaptation

After prediction, the completed unlabeled trial becomes one observation of the target translation.

One observation per trial. The latent trial center $m_t \in \mathbb{R}^D$ and its shift observation $r_t \in \mathbb{R}^D$ are

$$m_t = \frac{1}{M_t} \sum_{j=1}^{M_t} z_{t,j}, \quad r_t = m_t - \bar{p}. \tag{9}$$

We call r_t the class-agnostic shift observation: it estimates how far the target trial center lies from the balanced source center without assigning the trial to a class. Neither the true nor the predicted class is needed to compute r_t. This label-free construction has a specific ambiguity, visible under the working model $z_{t,j} = p_{y_t} + \delta^\star_t + \epsilon_{t,j}$, where y_t is the unknown true class, $\delta^\star_t$ is the unknown target shift, and $\epsilon_{t,j}$ is residual representation noise. With $\bar{\epsilon}_t = M_t^{-1} \sum_j \epsilon_{t,j}$,

$$r_t = \delta^\star_t + (p_{y_t} - \bar{p}) + \bar{\epsilon}_t. \tag{10}$$

We cannot remove the unknown class term from a single trial. It averages out only when classes recur with approximately balanced local frequencies. More generally, if an update unit contains class proportions $\pi_t(c)$, where $\pi_t(c)$ is the fraction of its trials from class c, its bias is $\sum_c \pi_t(c) p_c - \bar{p}$.

Effective number of overlapping windows. Overlapping EEG windows reuse samples, so counting all M_t windows as independent would make the trial center appear too precise. We instead estimate an integrated autocorrelation time in latent space. After removing the trial center, the residual embedding is

$$u_{t,j} = z_{t,j} - m_t. \tag{11}$$

For temporal lag ℓ, the pooled similarity of valid window pairs is normalized by its zero-lag value:

$$\gamma_t(\ell) = \frac{1}{M_t} \sum_{j=1}^{M_t - \ell} u_{t,j}^\top u_{t,j+\ell}, \quad \rho_t(\ell) = \frac{\gamma_t(\ell)}{\gamma_t(0)}. \tag{12}$$

For $M_t \geq 3$, we evaluate $\rho_t(\ell)$ up to the fixed limit $L_t = \min(L_{\max}, M_t - 2)$. The initial-positive-sequence rule groups adjacent lags,

$$b_{t,h} = \rho_t(2h-1) + \rho_t(2h). \tag{13}$$

and retains consecutive pairs from $h = 1$ until the first non-positive value. If H_t pairs remain, the integrated correlation time is

$$\tau_t = \mathrm{clip}\left(1 + 2 \sum_{h=1}^{H_t} b_{t,h}, 1, M_t \right). \tag{14}$$

Here $\mathrm{clip}(a, l, u) = \min\{u, \max\{l, a\}\}$. We set $\tau_t = 1$ when a trial has fewer than three windows and $\tau_t = M_t$ when a longer latent sequence has zero variance. The effective number of independent windows is then

$$M_{\mathrm{eff},t} = \frac{M_t}{\tau_t}. \tag{15}$$

Independent windows yield $M_{\mathrm{eff},t} \approx M_t$; near-duplicates drive it toward one. A separate statistic measures how widely the embeddings are dispersed around their trial center:

$$V_t = \frac{1}{M_t} \sum_{j=1}^{M_t} \|z_{t,j} - m_t\|_2^2. \tag{16}$$

Combining redundancy and dispersion gives the uncertainty of the trial center:

$$R_t = \frac{V_t}{M_{\mathrm{eff},t}} = \frac{\tau_t}{M_t} V_t. \tag{17}$$

This is uncertainty about the shift observation, not the predicted class. For a fixed window count, greater dispersion or stronger correlation raises R_t, so repeated or inconsistent windows cannot make the update artificially confident.

Bounded recursive update. The shift starts at $\delta_0 = 0$, as specified above. Its scalar uncertainty $P_t \geq 0$ represents the uncertainty associated with the current shift estimate. We initialize its scale from the spread of the class prototypes:

$$s = \|\mathrm{std}_c(p_c)\|_2 + \varepsilon, \quad P_0 = s^2, \quad Q = (q_{\mathrm{scale}} s)^2. \tag{18}$$

Here $\mathrm{std}_c(p_c) \in \mathbb{R}^D$ is computed coordinate-wise across the C prototypes, $\varepsilon > 0$ stabilizes the scale, and the source-fixed multiplier q_{scale} determines the process variance $Q \geq 0$. Once a trial is complete, its innovation measures disagreement with the previous shift. The innovation, nominal gain, and bounded update rate are

$$e_t = r_t - \delta_{t-1}, \quad g_t = \frac{P_{t-1}}{P_{t-1} + \max(R_t, \varepsilon_R)}, \quad w_t = \mathrm{clip}(g_t, w_{\min}, w_{\max}). \tag{19}$$

We then update the shared translation and propagate its uncertainty:

$$\delta_t = \delta_{t-1} + w_t e_t, \quad P_t = (1 - w_t) P_{t-1} + Q. \tag{20}$$

Thus, a reliable trial moves an uncertain estimate farther, whereas a noisy trial has less influence. The constant $\varepsilon_R > 0$ prevents a zero denominator. The upper bound on w_t stops one trial from replacing the accumulated state; the lower bound keeps the tracker from becoming permanently inert.

3 Experimental Setup

We study calibration-free transfer to an unseen participant. For each dataset and seed, every participant serves once as the target in a leave-one-subject-out (LOSO) fold. The remaining participants form the source pool. Their fitting sessions train the supervised classifier and define the class prototypes, whereas their separate selection sessions choose the checkpoint and hyperparameters. The evaluation session of the target participant is used only as the test stream and is replayed once in its recorded order. We replay each target session causally: a decision is recorded before the corresponding observation can modify the adapter. For motor imagery and SSVEP, the trial is both the prediction unit and the subsequent update unit. Its overlapping windows therefore produce one trial prediction and, only afterward, one adaptation step. ERP/P300 has a different acquisition structure. TASER predicts each stimulus-locked event separately, but waits until all events in the same selection have been classified before updating. Their embeddings are then pooled into one adaptation step; their predicted labels are not combined, and performance is measured for each event.

3.1 Compared Methods

The common-backbone comparison begins with two fixed controls. Frozen classifies each target-window embedding with the source-trained linear head $Wz+b$ and obtains the trial label by majority vote. Proto-Frozen replaces that head with Euclidean distance to one source centroid per class, then applies the same majority vote over the window predictions. Neither control learns from the target stream. Against these controls, we compare TASER with seven label-free adapters: NEO-Causal (Murphy et al., 2026), T3A-Causal (Iwasawa & Matsuo, 2021), RoTTA (Yuan et al., 2023), T-TIME-Single (Li et al., 2024), BFT-A and BFT-D (Li et al., 2026), and PLTA (Lin et al., 2026). Within a fold, all ten methods start from the same source checkpoint and receive the same ordered observations.

Table 1: Datasets and retained participants included in the evaluation.

Paradigm	Dataset	Subjects	Classes	Channels
MI	BNCI2014-001 (BCI Comp. IV-2a) (Tangermann et al., 2012)	9	4	22
	BNCI2014-002 (Scherer et al., 2013)	14	2	15
	BNCI2014-004 (BCI Comp. IV-2b) (Tangermann et al., 2012)	9	2	3
	BNCI2015-001 (Faller et al., 2012)	12	2	13
	Zhou2016 (Zhou et al., 2016)	4	3	14
ERP/P300	BNCI2014-008 (Riccio et al., 2013)	8	2	8
	BNCI2014-009 (Aric'o et al., 2014)	10	2	16
	EPFL P300 (Hoffmann et al., 2008)	8	2	32
	Sosulski2019 (Sosulski et al., 2021)	13	2	31
SSVEP	Nakanishi2015 (Nakanishi et al., 2015)	9	12	8
	Kalunga2016 (Kalunga et al., 2016)	12	4	8
	MAMEM3 (Oikonomou et al., 2016)	11	5	14
	Wang2016 (Wang et al., 2017)	34	40	64

3.2 Task, Datasets and Evaluation

The benchmark spans motor imagery (MI), event-related potential (ERP)/P300 detection, and steady-state visual evoked potential (SSVEP) decoding. Its 13 datasets contain 153 retained participants, 2-40 classes, and 3-64 EEG channels (Table 1). We access all datasets through MOABB (Jayaram & Barachant, 2018); participant counts therefore follow the folds exposed by the benchmark loaders. The common-backbone experiments use EEGNet (Lawhern et al., 2018) as the primary encoder. For MI and SSVEP, the main protocol extracts 1-s windows at 20 Hz; robustness is also tested with 0.5-s windows at 20 Hz and 1-s windows at 10 Hz. ERP/P300 retains the complete 0.8-s event epoch. Appendix A specifies preprocessing, encoder training, checkpoint selection, hyperparameters, random seeds, and hardware.

Balanced accuracy at the decision unit, a trial for MI and SSVEP, an event for ERP/P300, is the primary metric. Macro-F1 and Cohen's κ account for class imbalance and differing chance levels. We also measure balanced accuracy over the first, middle, and final thirds of each stream to expose when adaptation helps or fails. We report the median and 95th-percentile online latency per decision unit, together with peak accelerator memory and whether the method executes backpropagation during deployment. Latency includes prediction and subsequent adaptation. Source training is excluded, and all methods are profiled with the same device and software stack.

3.3 Ablation Tests

Every ablation retains the source checkpoint, target order, and predict-then-update protocol. A0 is complete TASER; A1 collapses its temporal geometry to $K = 1$. A2 preserves the trial prediction but assimilates the windows as separate observations, directly testing the update unit. The remaining arms concern memory and adaptation strength: A3 fixes the gain to its mean on source-only pseudo-target streams, A4 uses the cumulative estimate $\delta_t = t^{-1} \sum_{i=1}^{t} r_i$, and A5 suppresses adaptation by holding $\delta_t = 0$. Each arm is evaluated with the primary temporal protocol and four encoder seeds on two datasets from each paradigm; the six dataset means receive equal weight.

4 Results and Discussion

Table 2 compares the common-backbone methods on the recorded target streams. Balanced accuracy is computed at the BCI decision unit. Each dataset receives equal weight in the cross-dataset summaries, so datasets with more participants do not dominate the ranking. The supplementary material reports macro-F1, Cohen's κ, participant-level results, and alternative temporal resolutions.

TASER obtains the highest dataset-balanced mean, 70.00%, followed by NEO at 68.61%, Proto-frozen at 64.39%, and Frozen at 64.21%. It ranks first on eight of the thirteen datasets and second on three. The margin over NEO is 1.34 points for motor imagery, 2.08 for ERP/P300, and 0.77 for SSVEP. Relative to Frozen, the corresponding gains are 7.65, 7.65, and 1.61 points. The smaller and less consistent SSVEP margin is compatible with a limitation of the shared-shift model: frequency-specific representations may not move as a single translation.

Table 2: Pseudo-online balanced accuracy (%): participant mean ± standard deviation over four seeds. Bold and underline mark the best and second-best result; ERP/P300 is scored per event.

Motor imagery										
Dataset	TASER	NEO	Proto	Frozen	RoTTA	PLTA	BFT-D	T-TIME	BFT-A	T3A
BNCI14-001	**50.30**± 17.85	48.45± 17.65	42.05± 17.07	44.23± 17.57	44.97± 13.82	41.52± 11.88	41.09± 9.23	46.09± 13.87	41.28± 9.39	38.60± 12.49
BNCI14-002	**77.11**± 14.44	73.48± 15.00	67.65± 16.42	68.69± 16.51	68.69± 13.95	74.97± 13.92	70.60± 11.56	71.85± 12.77	70.83± 11.22	69.38± 15.85
BNCI14-004	**78.35**± 14.13	77.48± 13.75	72.93± 12.85	74.86± 12.41	71.80± 14.69	75.21± 13.05	77.07± 13.27	75.95± 13.23	77.04± 13.53	71.83± 14.60
BNCI15-001	**74.95**± 15.44	72.55± 15.03	60.09± 13.25	62.15± 12.04	64.35± 10.84	68.79± 15.68	68.82± 13.14	69.84± 13.46	69.12± 13.09	63.33± 12.04
Zhou16	79.17± 3.59	**81.25**± 2.13	67.67± 7.16	71.71± 3.69	72.75± 3.27	75.83± 3.79	73.83± 4.92	77.83± 3.83	73.50± 4.75	62.04± 12.55
ERP/P300										
Dataset	TASER	NEO	Proto	Frozen	RoTTA	PLTA	BFT-D	T-TIME	BFT-A	T3A
BNCI14-008	**72.74**± 3.13	70.91± 3.58	72.68± 3.10	63.46± 5.86	61.62± 5.12	72.47± 3.12	64.72± 4.72	66.52± 4.41	63.26± 4.60	71.76± 3.00
BNCI14-009	79.54± 5.97	79.51± 5.06	79.44± 6.06	73.70± 8.54	72.85± 6.92	**79.71**± 6.95	64.21± 5.02	64.70± 5.14	62.83± 5.35	78.42± 5.54
EPFL	61.98± 6.08	**62.18**± 6.57	61.72± 6.17	56.54± 4.86	54.77± 4.54	59.47± 6.61	54.79± 3.85	55.27± 4.46	54.21± 3.43	59.45± 5.15
Sosulski19	**70.60**± 6.36	63.96± 6.22	69.71± 5.21	60.59± 8.59	59.72± 8.77	67.72± 6.99	63.38± 6.26	66.21± 6.20	62.40± 6.33	68.73± 4.95
SSVEP										
Dataset	TASER	NEO	Proto	Frozen	RoTTA	PLTA	BFT-D	T-TIME	BFT-A	T3A
Nakanishi15	88.03± 21.42	88.99± 20.43	82.44± 25.22	88.99± 20.13	**90.21**± 18.50	83.99± 18.03	83.93± 20.68	82.31± 21.16	83.43± 20.81	76.92± 20.76
Kalunga16	**62.40**± 14.23	57.19± 12.28	57.26± 12.15	58.27± 11.41	55.70± 16.97	60.42± 12.45	49.32± 8.42	44.34± 9.82	49.80± 8.57	51.30± 16.67
MAMEM3	47.76± 19.95	48.99± 21.05	45.45± 18.02	47.15± 20.53	45.16± 19.03	39.83± 16.65	**50.69**± 17.88	44.80± 16.91	50.13± 17.96	35.77± 14.11
Wang16	**67.06**± 17.25	67.00± 16.36	57.98± 19.91	64.40± 18.92	66.89± 16.35	18.36± 9.44	52.19± 17.73	47.87± 13.53	50.78± 17.17	37.99± 11.92

Across the nine MI and SSVEP datasets, TASER records the highest mean macro-F1 (68.8%) and Cohen's κ (55.8), compared with 67.1% and 53.9 for NEO. Participant variability remains large, especially for SSVEP, as shown by the standard deviations in Table 2 and the participant-level distributions in the appendix.

Two ablations materially change the result. Averaged over two datasets from each paradigm, updating after every window rather than once per trial lowers balanced accuracy by 5.29 points; suppressing the online shift lowers it by 5.53 points. The remaining differences are below one point: setting $K = 1$ costs 0.13, replacing the adaptive gain with its source-matched fixed value costs 0.17, and averaging all past shift observations costs 0.81. Within this experiment, the useful core is an evolving shift updated once per trial.

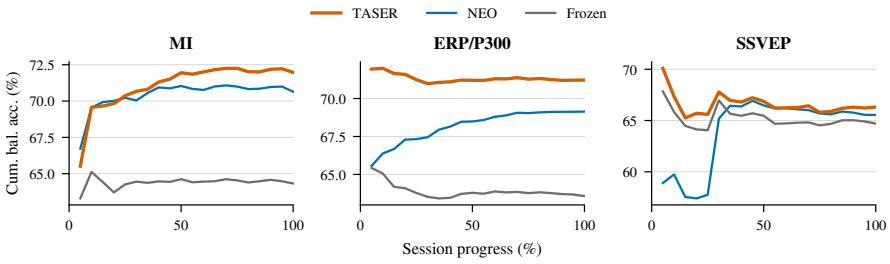
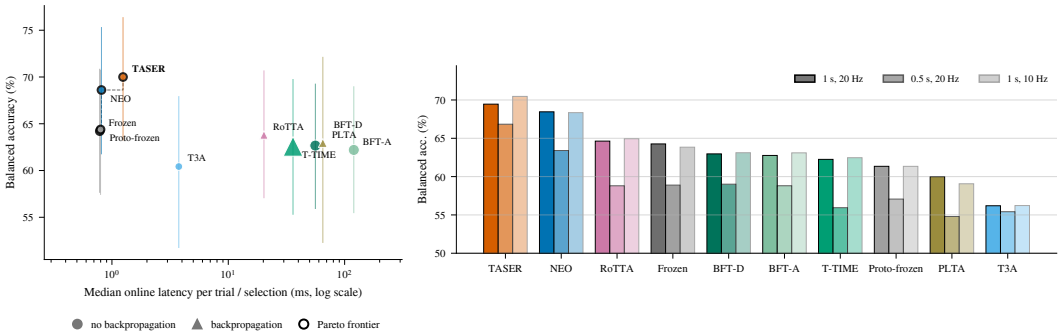


Figure 2: Cumulative balanced accuracy over the natural target stream. Lines are dataset-balanced means over the same encoders and observations; their endpoints give the task means in Table 2. Non-cumulative and per-dataset curves are reported in Appendix Figures 7-10.

The cumulative curves in Figure 2 reveal different behavior across the three paradigms. On motor imagery, TASER begins slightly below NEO after 5% of the session (65.5% versus 66.7%), overtakes it by the middle (72.0% versus 71.0%), and finishes at 72.0% versus 70.6%; Frozen ends at 64.3%. ERP/P300 behaves differently. TASER is already at 71.9% after the first 5% and remains near that level, ending at 71.2%, while NEO rises from 65.6% to 69.1% and Frozen ends at 63.6%. The early separation cannot be explained by accumulated target adaptation alone and is consistent with a contribution from TASER's prototype-based decision rule.

The SSVEP curve is less stable. TASER moves from 70.1% at 5% of the stream to 66.3% at its end, while NEO moves from 58.9% to 65.5% and Frozen finishes at 64.7%. The initial TASER-NEO gap therefore contracts from 11.2 to 0.8 points. Their shape also reflects which classes and participants

contribute at each normalized position. The non-cumulative blocks and dataset-specific curves in the appendix expose this variation directly.



(a) Accuracy-latency trade-off. (b) Temporal-resolution robustness.

Figure 3: Efficiency and temporal robustness. (a) Dataset-balanced accuracy versus median online latency; marker area encodes peak memory and bars show 95% hierarchical bootstrap intervals. (b) Accuracy under three temporal configurations on the same folds; full values are reported in Table 3.

Figure 3a places this accuracy gain in its deployment context. TASER requires a median 1.246 ms per decision and 18.3 MB of peak accelerator memory, with no backward pass. NEO is faster at 0.813 ms but is 1.39 accuracy points lower; Frozen takes 0.786 ms and is 5.79 points lower. TASER, NEO, and the two fixed controls form the measured accuracy-latency frontier. The remaining adapters take between 3.756 and 119.91 ms per decision without exceeding TASER's accuracy. This comparison does not include source training, but it captures the cost paid while the BCI is operating.

TASER also remains first across the temporal protocols in Figure 3b. TASER averages 69.5% with 1-s windows at 20 Hz, 66.8% with 0.5-s windows at 20 Hz, and 70.5% with 1-s windows at 10 Hz. NEO obtains 68.5%, 63.4%, and 68.3%, leaving TASER margins of 1.0, 3.4, and 2.2 points. The ranking therefore persists across the tested window durations and extraction rates. The larger margin for 0.5-s windows is consistent with trial-level pooling becoming more useful

5 Conclusion

TASER adapts a frozen EEG decoder through one shared latent shift and one update per acquisition unit. Across thirteen datasets, it attains the strongest dataset-balanced result among the common-backbone methods, without pseudo-labels, replay, or backpropagation. The ablation identifies the evolving shift and trial-level update as the two consequential choices; the temporal partition and adaptive gain have much smaller isolated effects. TASER is also inexpensive and retains its ranking under the three tested temporal protocols. Distribution shift in a BCI is not solely an undesired variation. During closed-loop use, participants may learn to produce neural patterns that are more stable or easier to distinguish. Longitudinal studies have associated such changes in EEG feature geometry with improved BCI control and persistent user learning (Tortora et al., 2022; Ceradini et al., 2025). TASER is well matched to this setting: its shared translation can keep the decoder aligned with the common movement of the target representation while preserving the source encoder and class geometry. It can therefore accommodate gradual neural change without relying on pseudo-labels or repeated parameter updates that might destabilize the decoder. The present benchmark contains no interactive feedback, so it cannot determine whether a shift reflects user learning, physiology, electrode placement, or another source of non-stationarity. It nonetheless identifies TASER as a practical basis for studying such mutual adaptation. Future work should therefore extend the shared-shift model with a task- and participant-specific decoder that can separate undesired drift from learning-related changes without sacrificing causal operation. Longitudinal human-in-the-loop studies are needed to test whether this adaptation supports, rather than masks, user learning. A subsequent evaluation in closed-loop robotic control would move the analysis beyond offline decoding accuracy and measure task success, responsiveness, and stability while the user and decoder adapt to one another.

AI USE STATEMENT

Generative AI tools supported initial drafting and language editing. The authors verified all sources, claims, and generated text and take full responsibility for the manuscript. AI was not used to generate data or establish mathematical results.

ETHICS STATEMENT

This work uses public, de-identified EEG datasets and involves no new data collection or participant recruitment. Data use follows the terms of the original providers. Given the observed participant variability, the results do not establish readiness for clinical or high-stakes deployment.

REFERENCES

P. Aric'o, F. Aloise, F. Schettini, S. Salinari, D. Mattia, and F. Cincotti. Influence of P300 latency jitter on event related potential-based brain–computer interface performance. *Journal of Neural Engineering*, 11(3):035008, 2014. doi: 10.1088/1741-2560/11/3/035008.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3582–3591, 2023.

Igor Carrara and Théodore Papadopoulo. Pseudo-online framework for BCI evaluation: A MOABB perspective using various MI and SSVEP datasets. *Journal of Neural Engineering*, 21(1):016003, 2024.

Matteo Ceradini, Stefano Tortora, Silvestro Micera, and Luca Tonin. The effect of user learning for online EEG decoding of upper-limb movement intention. *IEEE Transactions on Medical Robotics and Bionics*, 7(2):633–641, 2025. doi: 10.1109/TMRB.2025.3537663.

Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(3):313–319, 2012. doi: 10.1109/TNSRE.2012.2189584.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Lisa Haxel, Jaivardhan Kapoor, Ulf Ziemann, and Jakob H. Macke. EDAPT: Towards calibration-free BCIs with continual online adaptation. *arXiv preprint arXiv:2508.10474*, 2025.

Ulrich Hoffmann, Jean-Marc Vesin, Touradj Ebrahimi, and Karin Diserens. An efficient P300-based brain–computer interface for disabled subjects. *Journal of Neuroscience Methods*, 167(1):115–125, 2008. doi: 10.1016/j.jneumeth.2007.03.005.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Vinay Jayaram and Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018.

Wenjie Jin, Xinxin Zhu, Lifeng Qian, Cunshu Wu, Fan Yang, Daowei Zhan, Zhaoyin Kang, Kaitao Luo, Dianhuai Meng, and Guangxu Xu. Electroencephalogram-based adaptive closed-loop brain–computer interface in neurorehabilitation: A review. *Frontiers in Computational Neuroscience*, 18:1431815, 2024. doi: 10.3389/fncom.2024.1431815.

Emmanuel K. Kalunga, Sylvain Chevallier, Quentin Barthélemy, Karim Djouani, Eric Monacelli, and Yskandar Hamam. Online SSVEP-based BCI using riemannian geometry. *Neurocomputing*, 191:55–68, 2016. doi: 10.1016/j.neucom.2016.01.007.

Shujhat Khan, Leonie Kallis, Harry Mee, Salim El Hadwe, Damiano Barone, Peter Hutchinson, and Angelos Kolias. Invasive brain–computer interface for communication: A scoping review. *Brain Sciences*, 15(4):336, 2025. doi: 10.3390/brainsci15040336.

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.

Shanglin Li, Motoaki Kawanabe, and Reinmar J. Kobler. SPDIM: Source-free unsupervised conditional and label shift adaptation in EEG. In *International Conference on Learning Representations*, 2025.

Siyang Li, Ziwei Wang, Hanbin Luo, Lieyun Ding, and Dongrui Wu. T-TIME: Test-time information maximization ensemble for plug-and-play BCIs. *arXiv preprint arXiv:2412.07228*, 2024.

Siyang Li, Jiayi Ouyang, Zhenyao Cui, Ziwei Wang, Tianwang Jia, Feng Wan, and Dongrui Wu. Backpropagation-free test-time adaptation for lightweight EEG-based brain–computer interfaces. *arXiv preprint arXiv:2601.07556*, 2026.

Zhanhui Lin, Yonglin Wu, Chenyun Dai, and Yao Guo. Enhancing real-time cross-subject EEG decoding via pseudo-label tuning and adaptation. *Applied Soft Computing*, pp. 115150, 2026. doi: 10.1016/j.asoc.2026.115150.

Alberto Llera, Vicenç Gómez, and Hilbert J. Kappen. Adaptive multiclass classification for brain computer interfaces. *Neural Computation*, 26(6):1108–1127, 2014. doi: 10.1162/NECO_a_00592.

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.

Huan Luo, Na Lu, Xiaopeng Wang, and Xu Niu. Bayesian test-time adaptation via dirichlet feature projection and GMM-driven inference for motor imagery EEG decoding. In *International Conference on Learning Representations*, 2026.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. NEO: No-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations*, 2026.

Masaki Nakanishi, Yijun Wang, Yu-Te Wang, and Tzyy-Ping Jung. A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PLOS ONE*, 10(10):e0140703, 2015. doi: 10.1371/journal.pone.0140703.

Vangelis P. Oikonomou, Georgios Liaros, Kostantinos Georgiadis, Elisavet Chatzilari, Katerina Adam, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Comparative evaluation of state-of-the-art algorithms for SSVEP-based BCIs. *arXiv preprint arXiv:1602.00904*, 2016.

Gert Pfurtscheller and Christa Neuper. Motor imagery and direct brain–computer communication. *Proceedings of the IEEE*, 89(7):1123–1134, 2001.

Angela Riccio, Luca Simione, Francesca Schettini, Alessia Pizzimenti, Maurizio Inghilleri, Marta Olivetti Belardinelli, Donatella Mattia, and Febo Cincotti. Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis. *Frontiers in Human Neuroscience*, 7:732, 2013. doi: 10.3389/fnhum.2013.00732.

Reinhold Scherer, Josef Faller, David Balderas, Elisabeth V. C. Friedrich, Markus Pr"oll, Brendan Allison, and Gernot R. M"uller-Putz. Brain–computer interfacing: More than the sum of its parts. *Soft Computing*, 17(2):317–331, 2013. doi: 10.1007/s00500-012-0895-4.

Jan Sosulski, David H"ubner, Aaron Klein, and Michael Tangermann. Online optimization of stimulation speed in an auditory brain–computer interface under time constraints. *arXiv preprint arXiv:2109.06011*, 2021.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Mueller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Luca Tonin and José del R. Millán. Noninvasive brain–machine interfaces for robotic devices. *Annual Review of Control, Robotics, and Autonomous Systems*, 4:191–214, 2021. doi: 10.1146/annurev-control-012720-093904.

Stefano Tortora, Gloria Beraldo, Francesco Bettella, Emanuela Formaggio, Maria Rubega, Alessandra Del Felice, Stefano Masiero, Ruggero Carli, Nicola Petrone, Emanuele Menegatti, and Luca Tonin. Neural correlates of user learning during long-term BCI training for the Cybathlon competition. *Journal of NeuroEngineering and Rehabilitation*, 19:69, 2022. doi: 10.1186/s12984-022-01047-x.

Carmen Vidaurre, Motoaki Kawanabe, Paul von Bünau, Benjamin Blankertz, and Klaus-Robert Müller. Toward unsupervised adaptation of LDA for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 58(3):587–597, 2011. doi: 10.1109/TBME.2010.2093133.

Athanasios Vourvopoulos, Mathis Fleury, Luca Tonin, and Serafeim Perdikis. Editorial: Neurotechnologies and brain–computer interaction for neurorehabilitation. *Frontiers in Neuroergonomics*, 4:1203934, 2023. doi: 10.3389/fnrgo.2023.1203934.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Yijun Wang, Xiaogang Chen, Xiaorong Gao, and Shangkai Gao. A benchmark dataset for SSVEP-based brain–computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(10):1746–1752, 2017. doi: 10.1109/TNSRE.2016.2627556.

Dong Wen, Yali Fan, Sheng-Hsiou Hsu, Jian Xu, Yanhong Zhou, Jianxin Tao, Xifa Lan, and Fengnian Li. Combining brain–computer interface and virtual reality for rehabilitation in neurological diseases: A narrative review. *Annals of Physical and Rehabilitation Medicine*, 64(1):101404, 2021. doi: 10.1016/j.rehab.2020.03.015.

Martin Wimpff, Mario Döbler, and Bin Yang. Calibration-free online test-time adaptation for electroencephalography motor imagery decoding. *arXiv preprint arXiv:2311.18520*, 2023.

Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan. Brain–computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Bangyan Zhou, Xiaopei Wu, Zhao Lv, Lei Zhang, and Xiaojin Guo. A fully automated trial selection method for optimization of motor imagery based brain–computer interface. *PLOS ONE*, 11(9): e0162657, 2016. doi: 10.1371/journal.pone.0162657.

A Implementation Details

Motor-imagery and SSVEP signals are band-pass filtered at 8-35 and 5-45 Hz, respectively, and resampled to 200 Hz. ERP/P300 events use a 1-20 Hz passband and a sampling rate of 128 Hz. In every fold, channel-wise normalization is fitted on the source-fitting data and then held fixed for source selection and target evaluation.

We train one EEGNet source encoder for each dataset, LOSO target participant, seed, and temporal protocol defined in Section 3.2. Methods in the common-backbone comparison share the resulting checkpoint and cached embeddings within a protocol. Checkpoints are not reused across temporal protocols.

Source encoders minimize cross-entropy with AdamW for at most 100 epochs (10^{-3} learning rate, 10^{-4} weight decay, batch size 16). Balanced accuracy on the source-selection sessions selects the checkpoint, with early stopping after 15 consecutive epochs without an improvement above 10^{-4}. We retain every source window and repeat each fold with seeds $\{0, 1, 2, 3\}$. Published baseline hyperparameters are kept when compatible with causal evaluation.

TASER sets $q_{\text{scale}} = 0.05$, bounds $w_t \in [0.0606, 0.1176]$ (exponential-memory spans of approximately 32 and 16 trials), and considers $\mathcal{K} = \{1, 2, 3, 4, 6\}$. Source participants select the feasible segment count before target replay as described in Section 2.1. Predictive experiments were run as one-GPU jobs on NVIDIA A40 accelerators.

B Temporal Robustness and Additional Metrics

Table 3: Robustness to the window length and the update rate (EEGNet). Balanced accuracy (%) per dataset on the same 274 folds; small numbers are the standard deviation over participants (mean column: over datasets). Bold marks the best method of each dataset and configuration.

Method	BNCI 2014-001	BNCI 2014-002	BNCI 2014-004	BNCI 2015-001	Zhou 2016	Kalunga 2016	MAMEM3	Nakanishi 2015	Wang 2016	Mean
1 s, 20 Hz (primary)										
TASER	**50.5** ±17.3	**77.6** ±13.9	**78.5** ±13.3	**75.1** ±14.5	79.1 ±3.7	**62.4** ±13.6	47.8 ±19.0	87.6 ±20.7	66.5 ±17.4	**69.5** ±12.9
NEO	48.3 ±17.5	74.4 ±13.7	77.5 ±13.1	73.1 ±14.1	**81.3** ±2.2	57.2 ±11.8	49.0 ±20.1	88.7 ±19.6	66.7 ±16.2	68.5 ±13.5
Frozen	44.3 ±17.2	68.9 ±15.2	74.6 ±12.3	62.1 ±12.4	71.3 ±3.1	58.3 ±10.9	47.1 ±19.6	88.2 ±19.6	63.6 ±18.7	64.3 ±12.8
Proto	42.8 ±17.5	68.0 ±15.4	73.5 ±12.5	59.9 ±12.3	66.8 ±6.5	57.3 ±11.6	45.4 ±17.2	81.0 ±25.8	57.4 ±19.6	61.3 ±11.7
T3A	38.9 ±13.0	69.2 ±15.7	72.3 ±14.0	62.9 ±10.7	61.7 ±10.9	51.3 ±16.0	35.8 ±13.5	76.3 ±19.2	37.5 ±12.2	56.2 ±14.9
RoTTA	45.5 ±14.1	70.2 ±13.8	71.7 ±14.2	65.0 ±9.7	71.8 ±3.6	55.7 ±16.2	45.2 ±18.1	**89.9** ±18.2	**66.7** ±16.2	64.6 ±13.3
T-TIME	45.6 ±13.8	72.1 ±11.6	76.4 ±12.4	69.3 ±12.9	77.8 ±3.3	44.3 ±9.4	44.8 ±16.1	82.0 ±19.7	47.8 ±13.0	62.2 ±15.2
BFT-A	41.1 ±9.4	70.8 ±11.6	77.0 ±12.8	68.8 ±13.1	72.8 ±4.2	49.8 ±8.2	50.1 ±17.1	83.5 ±18.9	51.0 ±17.0	62.8 ±14.0
BFT-D	41.0 ±9.3	70.8 ±11.9	77.3 ±12.4	68.3 ±13.3	72.8 ±4.4	49.3 ±8.1	**50.7** ±17.1	84.1 ±18.8	52.5 ±17.5	63.0 ±14.0
PLTA	41.7 ±10.8	76.1 ±12.9	75.5 ±12.2	69.3 ±14.8	75.2 ±4.0	60.4 ±11.9	39.8 ±15.9	83.9 ±18.7	18.0 ±9.6	60.0 ±20.8
0.5 s, 20 Hz										
TASER	**51.8** ±18.4	**75.0** ±14.2	**78.9** ±13.0	**76.6** ±16.3	79.5 ±1.9	**63.3** ±14.0	**47.3** ±19.8	**81.7** ±22.9	**47.4** ±13.5	**66.8** ±13.7
NEO	49.1 ±16.4	70.2 ±13.3	77.7 ±12.6	70.5 ±15.7	**83.2** ±1.5	61.9 ±11.3	44.3 ±18.3	76.3 ±17.6	37.5 ±10.7	63.4 ±15.3
Frozen	42.3 ±16.7	66.3 ±15.4	74.0 ±12.3	62.3 ±12.8	73.5 ±5.2	60.6 ±9.6	43.1 ±17.6	72.9 ±17.6	35.0 ±12.5	58.9 ±14.1
Proto	41.6 ±16.0	65.4 ±14.3	74.0 ±11.8	60.5 ±13.3	69.3 ±3.9	56.4 ±11.8	41.8 ±16.9	69.4 ±22.5	35.3 ±12.7	57.1 ±13.4
T3A	38.9 ±11.3	68.3 ±14.6	72.6 ±12.4	64.2 ±11.7	71.2 ±5.4	51.7 ±11.7	39.0 ±16.4	65.0 ±20.4	28.0 ±9.7	55.4 ±15.6
RoTTA	44.6 ±14.2	65.0 ±12.2	72.0 ±12.7	61.2 ±10.2	74.0 ±7.4	56.3 ±19.5	41.1 ±15.1	78.1 ±16.8	36.9 ±10.2	58.8 ±14.2
T-TIME	42.8 ±13.2	70.8 ±12.5	76.3 ±12.3	68.2 ±13.3	76.1 ±4.6	47.5 ±12.3	39.5 ±12.3	56.2 ±14.2	26.2 ±8.6	55.9 ±17.0
BFT-A	42.3 ±10.7	70.1 ±12.1	76.7 ±12.9	70.9 ±13.5	76.2 ±3.4	51.7 ±9.6	45.7 ±15.2	66.8 ±17.2	28.8 ±9.0	58.8 ±16.2
BFT-D	42.0 ±10.3	70.2 ±12.0	76.9 ±13.0	70.8 ±13.7	75.4 ±3.0	52.4 ±9.7	47.1 ±15.2	66.9 ±17.6	29.4 ±9.7	59.0 ±15.9
PLTA	39.9 ±11.0	74.9 ±12.6	76.3 ±12.0	67.9 ±14.5	76.8 ±2.9	61.7 ±16.5	35.0 ±13.8	55.9 ±14.3	4.8 ±0.9	54.8 ±22.8
1 s, 10 Hz										
TASER	**50.1** ±18.1	**76.7** ±13.3	**79.0** ±12.7	**75.1** ±15.6	81.1 ±2.2	**61.9** ±12.6	48.5 ±20.8	88.4 ±18.4	**73.4** ±15.5	**70.5** ±13.1
NEO	48.0 ±15.8	72.6 ±14.9	77.7 ±12.8	72.6 ±15.0	**81.8** ±3.5	58.6 ±9.6	47.2 ±19.5	88.4 ±18.3	68.2 ±16.9	68.3 ±13.6
Frozen	42.4 ±16.5	67.6 ±15.8	75.0 ±11.7	62.0 ±11.2	72.8 ±3.5	56.2 ±10.4	46.5 ±21.0	87.2 ±21.3	64.8 ±18.7	63.8 ±13.3
Proto	42.0 ±17.0	65.1 ±15.3	73.7 ±10.5	61.5 ±13.4	69.8 ±4.0	54.1 ±10.8	45.2 ±18.2	82.5 ±24.3	58.2 ±19.1	61.3 ±12.4
T3A	39.1 ±15.3	68.2 ±14.9	72.9 ±12.6	66.0 ±13.8	61.6 ±12.2	50.1 ±14.6	37.6 ±14.7	72.3 ±21.5	38.3 ±11.7	56.2 ±14.1
RoTTA	44.7 ±13.1	68.0 ±13.1	73.3 ±14.9	65.9 ±10.8	74.2 ±4.2	56.5 ±15.6	44.1 ±19.4	**89.7** ±17.4	67.9 ±16.0	64.9 ±13.8
T-TIME	44.8 ±15.7	70.4 ±12.9	75.9 ±12.8	71.8 ±13.3	77.2 ±3.1	45.3 ±12.3	44.4 ±16.4	82.5 ±19.7	50.0 ±13.9	62.5 ±15.1
BFT-A	40.9 ±10.4	69.0 ±11.6	77.0 ±13.0	71.3 ±11.7	73.9 ±4.7	50.2 ±10.2	51.0 ±19.0	82.4 ±19.6	52.1 ±15.8	63.1 ±13.8
BFT-D	40.6 ±10.9	69.2 ±11.0	77.2 ±13.0	70.7 ±12.0	73.2 ±4.7	50.0 ±10.1	**51.0** ±18.7	82.5 ±20.1	53.5 ±16.6	63.1 ±13.7
PLTA	40.8 ±12.1	74.7 ±13.7	75.1 ±12.8	68.7 ±15.1	75.8 ±2.9	56.2 ±14.4	38.4 ±15.8	82.9 ±17.6	19.0 ±9.2	59.1 ±20.6

Table 4: Cohen's κ and macro-F1 on the natural stream (EEGNet): mean over participants of the mean over four seeds; small numbers are the standard deviation over participants (mean column: over datasets). Bold marks the best method of each dataset and metric.

Method	BNCI 2014-001	BNCI 2014-002	BNCI 2014-004	BNCI 2015-001	Zhou 2016	Kalunga 2016	MAMEM3	Nakanishi 2015	Wang 2016	Mean
Cohen's κ ($\times 100$)										
TASER	**33.7** ±22.4	**54.2** ±27.8	**56.7** ±26.6	**49.9** ±29.6	68.8 ±4.7	**49.9** ±18.2	35.6 ±24.1	86.9 ±22.0	**66.2** ±17.4	**55.8** ±15.7
NEO	31.3 ±22.2	47.0 ±28.9	55.0 ±25.9	45.1 ±28.8	**71.9** ±2.8	42.9 ±15.7	37.9 ±25.6	88.0 ±21.0	66.2 ±16.5	53.9 ±17.2
Frozen	25.6 ±22.1	37.4 ±31.8	49.7 ±23.4	24.3 ±23.1	57.6 ±4.8	44.4 ±14.6	34.8 ±24.4	88.0 ±20.7	63.5 ±19.1	47.3 ±19.1
Proto	22.7 ±21.5	35.3 ±31.6	45.9 ±24.2	20.2 ±25.4	51.5 ±9.3	43.0 ±15.5	33.1 ±21.8	80.8 ±25.9	56.9 ±20.1	43.3 ±17.6
T3A	18.1 ±15.7	38.8 ±30.5	43.7 ±27.5	26.7 ±23.1	43.1 ±16.3	35.1 ±21.3	20.4 ±17.4	74.8 ±21.4	36.4 ±12.0	37.4 ±15.8
RoTTA	26.6 ±17.4	37.4 ±26.9	43.6 ±27.7	28.7 ±20.8	59.1 ±4.2	40.9 ±21.7	33.0 ±23.4	**89.3** ±19.0	66.0 ±16.5	47.2 ±19.4
T-TIME	28.1 ±17.4	43.7 ±24.6	51.9 ±25.0	39.7 ±25.8	66.8 ±5.0	25.8 ±12.5	32.4 ±20.6	80.7 ±21.8	46.5 ±13.7	46.2 ±17.1
BFT-A	21.7 ±11.8	41.7 ±21.6	54.1 ±25.5	38.2 ±25.1	60.2 ±6.2	33.1 ±10.9	39.3 ±21.5	81.9 ±21.4	49.5 ±17.3	46.6 ±16.5
BFT-D	21.5 ±11.6	41.2 ±22.3	54.1 ±25.0	37.6 ±25.2	60.8 ±6.4	32.4 ±10.8	**40.0** ±21.3	82.5 ±21.3	51.0 ±17.9	46.8 ±16.8
PLTA	22.0 ±14.9	49.9 ±26.8	50.4 ±24.6	37.6 ±30.0	63.8 ±4.9	47.2 ±15.9	25.8 ±20.6	82.5 ±18.5	16.3 ±9.5	43.9 ±20.0
Macro-F1 (%)										
TASER	**50.0** ±16.8	**76.7** ±14.3	**78.3** ±13.3	**74.9** ±14.8	78.6 ±3.0	**62.0** ±13.6	46.9 ±19.2	87.4 ±21.5	64.4 ±18.7	**68.8** ±13.1
NEO	47.4 ±16.8	71.8 ±16.7	77.3 ±13.1	71.3 ±15.4	**81.2** ±2.0	54.9 ±12.2	47.1 ±20.9	88.7 ±19.7	**64.5** ±16.9	67.1 ±14.0
Frozen	39.3 ±19.0	62.6 ±21.8	74.2 ±12.1	54.6 ±17.4	70.3 ±3.1	56.3 ±11.7	44.0 ±22.0	88.5 ±19.8	61.1 ±19.7	61.2 ±14.3
Proto	35.5 ±19.0	61.0 ±21.7	71.5 ±13.4	51.1 ±18.3	65.1 ±7.3	55.1 ±12.0	42.6 ±19.3	81.7 ±25.0	54.6 ±21.1	57.6 ±13.4
T3A	35.4 ±12.7	64.2 ±20.2	70.4 ±15.0	59.0 ±13.9	60.5 ±10.7	48.9 ±16.6	32.0 ±15.0	75.7 ±20.1	31.8 ±10.8	53.1 ±15.8
RoTTA	43.5 ±13.0	66.5 ±15.1	69.8 ±16.0	61.2 ±12.8	72.3 ±3.4	54.5 ±15.8	42.8 ±19.8	**90.1** ±17.6	64.4 ±16.6	62.8 ±13.9
T-TIME	45.2 ±12.6	70.9 ±13.7	75.9 ±12.5	69.5 ±13.0	77.7 ±3.5	39.3 ±9.3	42.8 ±16.5	81.9 ±20.2	43.7 ±13.1	60.7 ±16.5
BFT-A	40.8 ±8.9	70.5 ±11.0	77.0 ±12.8	69.0 ±12.6	73.2 ±4.4	44.4 ±7.7	49.1 ±17.8	83.3 ±19.7	47.3 ±16.8	61.6 ±15.1
BFT-D	40.6 ±8.7	70.2 ±11.4	77.0 ±12.5	68.7 ±12.6	73.6 ±4.5	44.1 ±7.6	**49.7** ±17.6	83.8 ±19.5	49.0 ±17.6	61.9 ±15.1
PLTA	38.0 ±10.5	74.7 ±13.6	75.1 ±12.4	68.5 ±15.2	75.3 ±3.4	59.7 ±11.8	35.8 ±16.3	83.6 ±17.3	12.7 ±7.7	58.2 ±22.6

C ACCURACY-LATENCY MEASUREMENTS

Table 5 reports the measurements underlying Figure 3a. Balanced accuracy is averaged equally across the 13 datasets, with a hierarchical bootstrap interval in brackets. Median and p95 are the typical and tail latency per decision unit, respectively; *Peak MB* is the maximum allocated accelerator memory. The *Backprop* column records whether deployment computes gradients, while *Pareto* identifies methods for which no alternative is both faster and more accurate.

Table 5: Accuracy–latency values underlying Figure 3a. Accuracy is the dataset-balanced mean [95% hierarchical bootstrap interval] over 13 datasets.

Method	Bal. acc. (%)	Median (ms)	p95 (ms)	Peak MB	Backprop	Pareto
Frozen	64.21 [57.62, 70.88]	0.786	0.810	18.3	no	yes
Proto-frozen	64.39 [57.39, 70.75]	0.798	0.832	18.3	no	yes
NEO	68.61 [61.73, 75.34]	0.813	0.840	18.3	no	yes
TASER	70.00 [63.21, 76.40]	1.246	2.266	18.3	no	yes
T3A	60.42 [51.72, 67.96]	3.756	7.325	18.3	no	–
RoTTA	63.81 [57.04, 70.70]	20.26	44.58	49.9	yes	–
T-TIME	62.58 [55.27, 69.78]	36.00	47.84	4619.1	yes	–
BFT-D	62.67 [55.89, 69.29]	56.03	66.88	277.2	no	–
PLTA	62.95 [52.26, 72.16]	64.99	111.16	64.9	yes	–
BFT-A	62.20 [55.44, 69.01]	119.91	144.91	277.2	no	–

TASER reaches the highest dataset-balanced accuracy (70.00%) with a median latency of 1.246 ms, 18.3 MB of peak memory, and no online backpropagation. NEO and the two frozen controls are faster but less accurate, so all four remain on the measured Pareto frontier. The other adapters require between 3.756 and 119.91 ms per decision and do not improve accuracy over TASER in this comparison.

D PARTICIPANT-LEVEL RESULTS

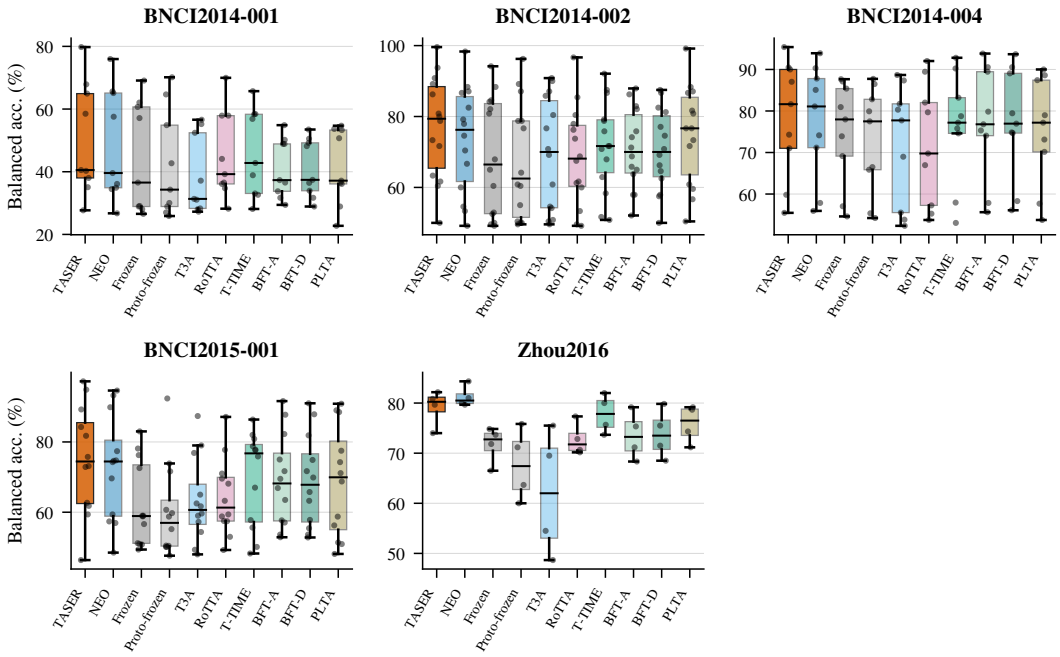


Figure 4: Participant-level balanced accuracy on the five motor-imagery datasets. Each box summarises the participants of a dataset for one method: the center marks the median, the box covers the interquartile range, and the lines show the full participant range. Points are participants, each averaged over four seeds.

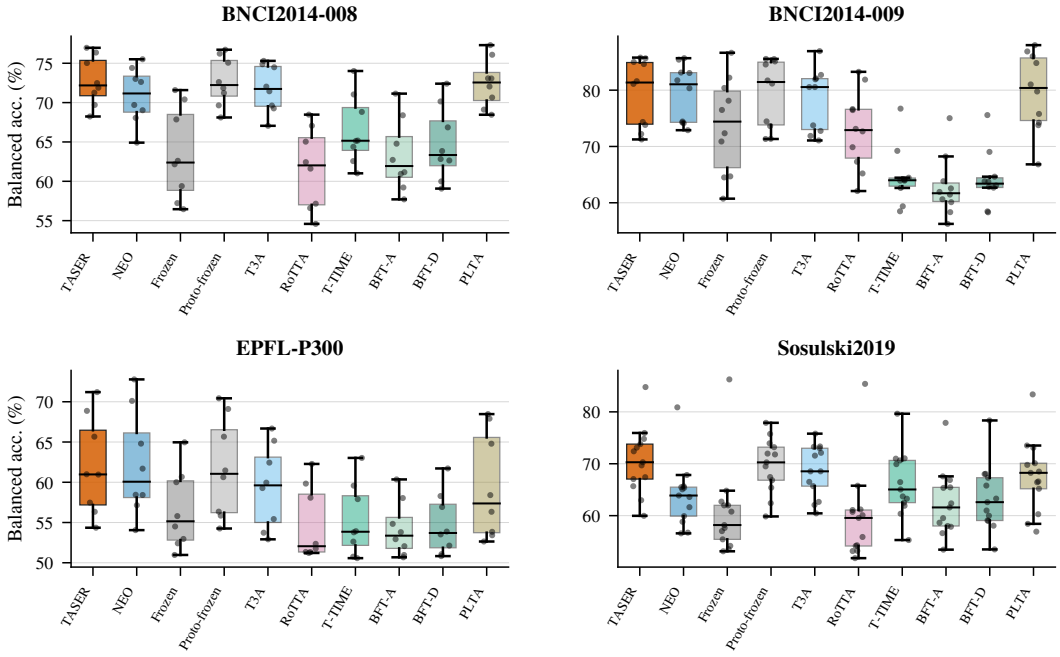


Figure 5: Participant-level balanced accuracy on the four ERP/P300 datasets. Each box summarises the participants of a dataset for one method: the center marks the median, the box covers the interquartile range, and the lines show the full participant range. Points are participants, each averaged over four seeds.

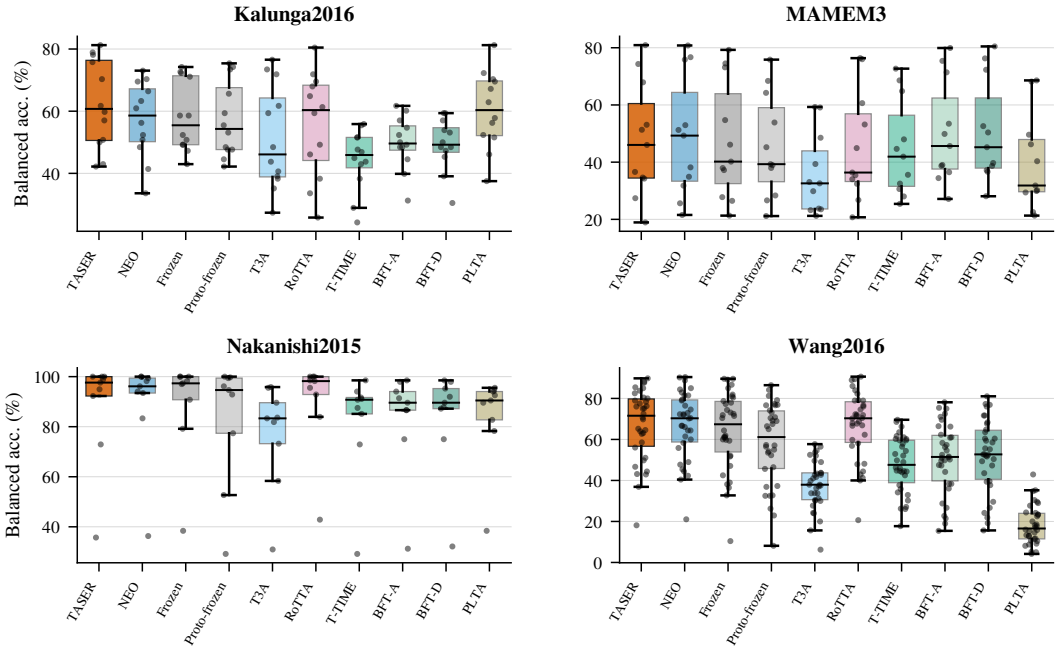


Figure 6: Participant-level balanced accuracy on the four SSVEP datasets. Each box summarises the participants of a dataset for one method: the center marks the median, the box covers the interquartile range, and the lines show the full participant range. Points are participants, each averaged over four seeds.

E SESSION-LEVEL DYNAMICS

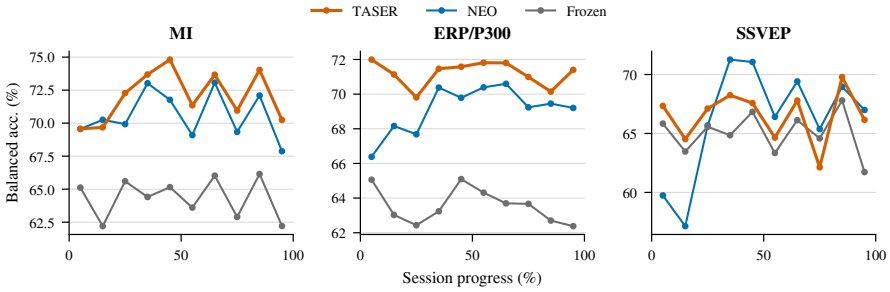


Figure 7: Balanced accuracy in each 10% block of the natural target stream (non-cumulative), for TASER, NEO and Frozen on the same encoder and trials of every fold; lines are dataset-balanced means. Figure 2 shows the cumulative version.

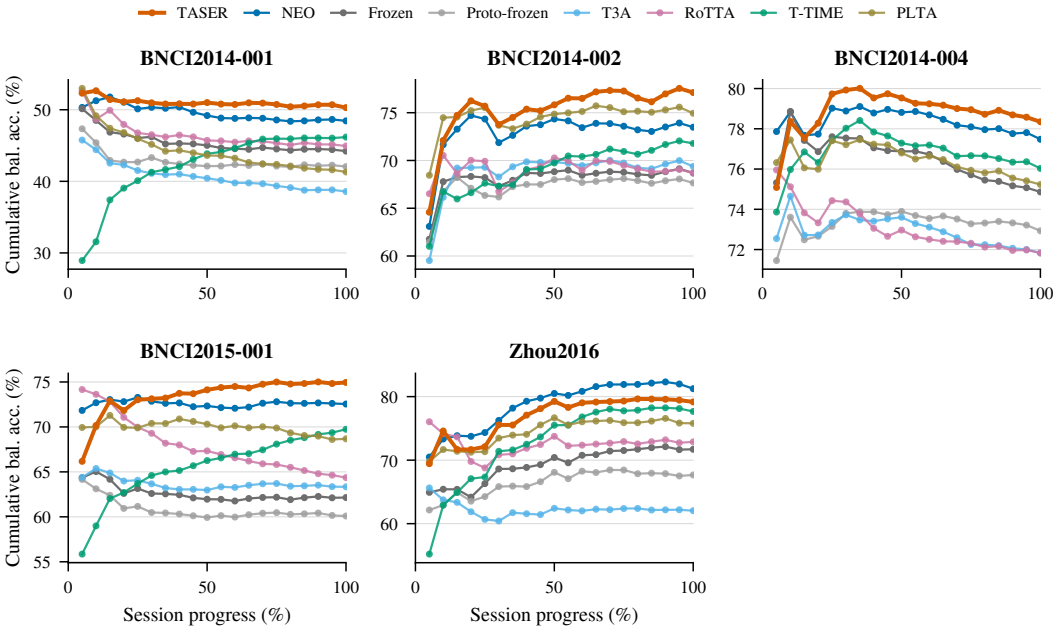


Figure 8: Cumulative balanced accuracy along the natural target stream on the five motor-imagery datasets, mean over participants of the mean over four seeds. All methods share the encoder and the trials of every fold.

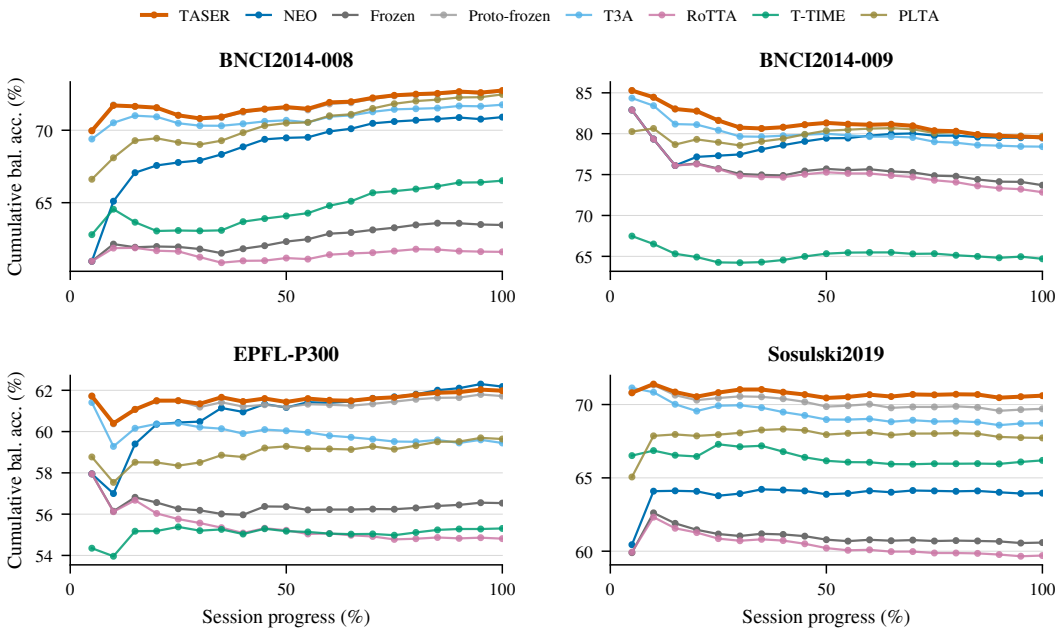


Figure 9: Cumulative balanced accuracy along the natural target stream on the four ERP/P300 datasets, mean over participants of the mean over four seeds. All methods share the encoder and the trials of every fold.

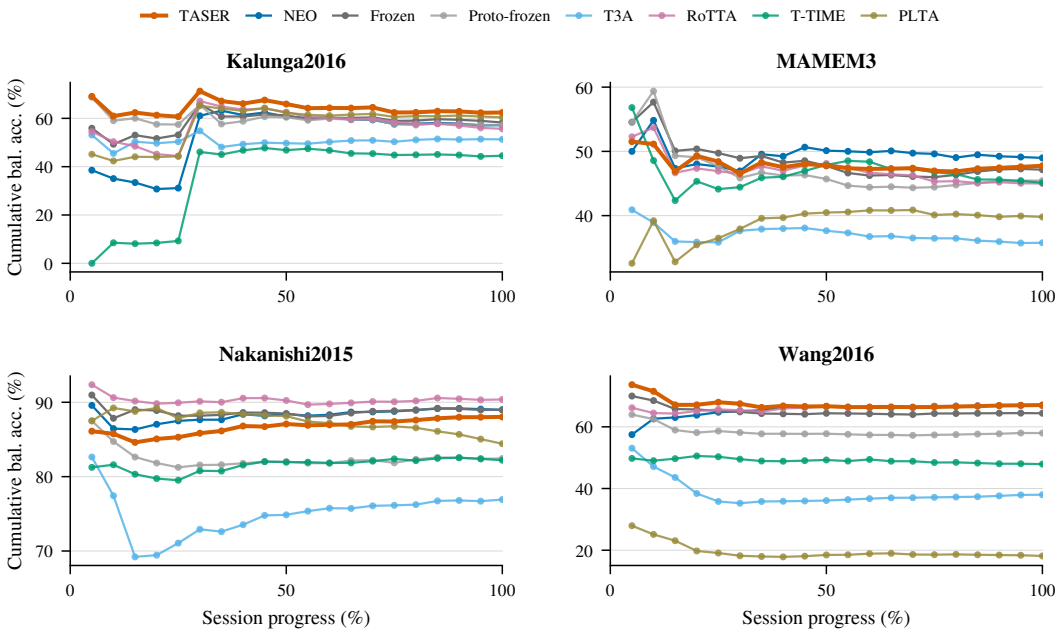


Figure 10: Cumulative balanced accuracy along the natural target stream on the four SSVEP datasets, mean over participants of the mean over four seeds. All methods share the encoder and the trials of every fold.