# RPT: RESULTANT-POWER TRANSPORT FOR DATA-FREE CONTINUAL EEG MODEL MERGING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Electroencephalography (EEG) supports diverse brain–computer interface and monitoring applications, requiring models that adapt to new subjects while retaining previously acquired knowledge. However, inter-subject variability can undermine knowledge retention during adaptation, and replay-based remedies require storing sensitive EEG recordings, raising privacy concerns and increasing storage demands. Continual model merging (CMM) offers a scalable alternative by incrementally merging subject-specific fine-tuned models without storing or replaying historical EEG. Nevertheless, conflicting parameter updates and differing normalization states across subjects can disrupt accumulated knowledge during merging. We introduce Resultant-Power Transport (RPT), which recursively merges subject models without accessing EEG or calibration samples. RPT uses task-vector agreement to scale layer-specific targets and selects incoming contributions through a closed-form, history-weighted projection that preserves the relative weights of historical contributions. Affine-map BN merging (AM-BN) complements this update by merging the inference transformations jointly defined by batch normalization statistics and affine parameters. Experiments across three EEG tasks demonstrate that RPT outperforms competing CMM methods, enabling data-free continual adaptation to new subjects while sustaining performance on previously encountered subjects.

## 1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive, temporally precise window into brain activity, supporting applications such as brain–computer interfaces (BCIs), emotion recognition, and sleep monitoring (Edelman et al., 2025; Zhang et al., 2025; Wan et al., 2025). As these systems serve new users and patients, a model must accommodate differences in physiology and recording conditions while retaining performance on earlier subjects (Jin et al., 2024; Han et al., 2026). This creates a subject-incremental learning problem: each arriving subject offers useful adaptation, but updating the shared model can erode previously acquired knowledge. Sustained decoding therefore requires both *plasticity* toward new subjects and *stability* on those already encountered.

Continual EEG learning methods address this challenge through replay, regularization, and cross-subject adaptation (Duan et al., 2023; 2024a; Zhou et al., 2025a). Replay is commonly used to mitigate catastrophic forgetting (McCloskey & Cohen, 1989), but storing and repeatedly accessing sensitive recordings can be impractical in EEG applications (Huang et al., 2026; Li et al., 2026a). Recent work offers an alternative through *continual model merging* (CMM), which incrementally merges sequentially arriving fine-tuned checkpoints with a shared backbone into a single model (Tang et al., 2025; Porrello et al., 2025; Qiu et al., 2026), offering advantages in scalability, data privacy, and storage efficiency (Figure 1). However, inter-subject variability in EEG poses a central challenge: merging conflicting parameter updates and differing normalization states while preserving decoding performance on earlier subjects.

Subject-specific fine-tuning modifies both model parameters and normalization statistics. Task vectors defined as parameter changes from a shared initialization differ in direction and magnitude across subjects. Averaging can cancel opposing components, while rescaling the entire merged vector also changes the magnitude of earlier contributions. This motivates determining each incoming contribution from both agreement among subject task vectors and the current merged model. Nor-



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.

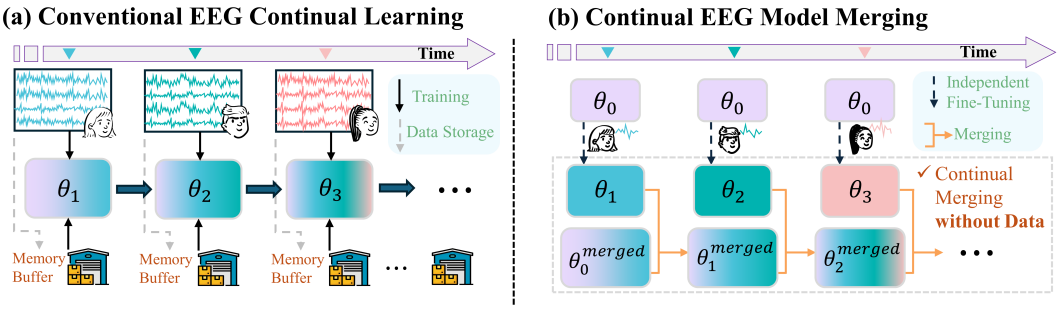



Figure 1: **Conventional continual learning versus continual model merging for EEG.** (a) Conventional continual learning updates a shared backbone using incoming EEG; the replay path reuses historical recordings. (b) Data-free continual model merging (CMM) receives a stream of existing subject fine-tuned models and integrates each arriving model with the previously merged backbone, accumulating knowledge without accessing EEG data.

malization poses a related challenge: differences in feature statistics across subjects affect EEG decoding (Kobler et al., 2022; Li et al., 2026c), and model interpolation can further alter activation statistics (Jordan et al., 2023). In batch normalization (BN), running means and variances interact with learned affine parameters to define the inference transformation. Because this transformation depends nonlinearly on these quantities, averaging them separately does not generally yield the average of the original transformations.

We introduce **Resultant-Power Transport (RPT)** to coordinate these two aspects of continual EEG model merging. The resultant of the encountered task vectors supplies a cross-subject reference, and their agreement controls its scale. A closed-form projection then selects the incoming contribution while preserving the relative weights of historical contributions. Complementing this update, *affine-map BN merging* (AM-BN) uses stored-statistic compatibility to merge BN inference maps and reconstructs the corresponding BN state. This design turns checkpoint information into a continual update rule with three contributions:

- **Resultant-guided task-vector transport.** We formulate continual merging as an update that preserves the relative contributions assigned to earlier subjects. Task-vector agreement sets a layer-specific target, and a closed-form projection determines how much of the incoming adaptation to retain. The rule selects the optimal coefficient within this update family using incrementally maintained checkpoint summaries.
- **Affine-map BN merging.** We introduce AM-BN to merge the inference transformations jointly encoded by running moments and affine parameters. Stored statistics measure compatibility with the pre-trained model and weight subject-specific changes in the map coefficients. Reconstructing the BN state exactly realizes the selected affine map, providing a normalization rule that works directly from checkpoints and complements parameter transport.
- **Improved continual EEG decoding.** Across three EEG tasks, RPT achieves the highest accuracy and macro-F1 among the compared CMM methods and improves aggregate prior-subject retention by 2.3 points. Its sequential adaptation variant also remains competitive with conventional continual learning methods while achieving stronger retention in most comparisons, demonstrating the value of CMM across both evaluation settings.

## 2 Related Work

**Continual EEG decoding.** Replay-based EEG methods preserve prior knowledge through memory selection and transformations of stored recordings (Duan et al., 2023; 2024b;a). Prototype-based approaches use compact class representations and cross-subject alignment (Li et al., 2026b;a), while BrainUICL and SPICED address unlabeled subject streams (Zhou et al., 2025a;b). These approaches establish the importance of retaining knowledge across subject shifts. We pursue the same objective through checkpoint merging and evaluate its use within sequential adaptation through Seq-RPT.

**Continual model merging.** Continual learning is an important application of model merging (Yang et al., 2026), with CMM incrementally merging fine-tuned checkpoints to retain knowledge across tasks, building on task arithmetic and interference-aware merging (Ilharco et al., 2023; Yadav et al., 2023). To reduce interference, OPCM uses orthogonal projection and adaptive scaling (Tang et al., 2025), DOP balances stability and plasticity through dual projections (Yang et al., 2025), and NUFILT couples null-space filtering with low-rank adaptation (Qiu et al., 2026). RPT uses cross-subject agreement to guide a closed-form scalar projection per linear layer, preserving relative historical weights while adjusting incoming contributions. AM-BN complements this update by merging BN inference maps directly from checkpoints.

**Normalization across domains and models.** Domain-specific normalization supports EEG transfer by accounting for feature statistics across subjects (Kobler et al., 2022). HEEGNet places moment and distribution alignment within a framework for domain-invariant EEG representations (Li et al., 2026c). In model merging, REPAIR addresses activation variance collapse after interpolation (Jordan et al., 2023). These studies motivate explicit treatment of normalization states as a component of EEG model merging. AM-BN complements task-vector transport by merging the inference maps jointly defined by BN running moments and affine parameters, using stored-statistic compatibility to weight changes in their coefficients and reconstruct the merged BN state.

## 3 METHODOLOGY

RPT has two complementary components (Figure 2). Resultant-guided task-vector transport determines the incoming parameter contribution, while affine-map BN merging (AM-BN) handles normalization transformations in backbones containing BN.

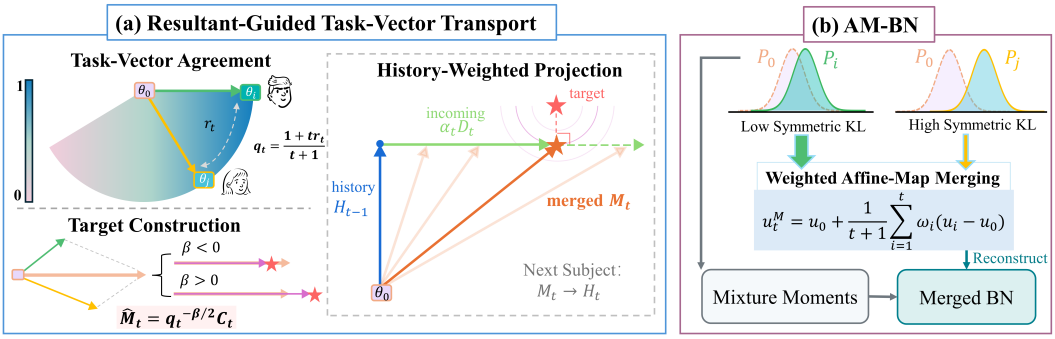



Figure 2: **Overview of RPT.** A simplified schematic summarizes two complementary components. (a) Task-vector agreement sets a target scale, and history-weighted projection selects the incoming contribution to the recursive update. (b) AM-BN weights changes in BN map coefficients by compatibility with the pre-trained model and reconstructs the BN state from the merged map and mixture moments.

### 3.1 PROBLEM SETTING

The inputs to CMM are a pre-trained model $\theta_0$ and a stream of subject checkpoints $\{\theta_t\}_{t=1}^{T}$ sharing its architecture and initialization. Within each EEG task, subject domains $\mathcal{D}_t$ share an input space and label set $\mathcal{Y}$, but differ in their data distributions. At step $t$, the merger receives $\theta_t$ and updates the previous merged model $\theta_{t-1}^{M}$ to a single model $\theta_t^{M}$. It accesses model parameters, stored BN statistics when present, and its incremental summaries, without EEG recordings or calibration samples. The objective is to retain decoding performance on earlier subjects while acquiring the incoming subject's adaptation.

Task vectors express adaptation relative to the common initialization (Ilharco et al., 2023; Yang et al., 2025; Qiu et al., 2026). For layer $g$, define the historical and incoming task vectors relative to the pre-trained backbone:

$$H_{t-1}^{g} = \theta_{t-1}^{M,g} - \theta_0^{g}, \qquad D_t^{g} = \theta_t^{g} - \theta_0^{g}. \tag{1}$$

Here $H^g_{t-1}$ represents the adaptation retained in the previously merged model. We initialize $\theta^M_0 = \theta_0$, so $H^g_0 = 0$.

### 3.2 RESULTANT-GUIDED TASK-VECTOR TRANSPORT

Subject task vectors encode the directions and magnitudes of adaptation. Prior work connects their geometry to task-specific representation subspaces (Cheng et al., 2025; Yang et al., 2025; Qiu et al., 2026); Appendix A.1 gives the corresponding gradient-based interpretation. RPT uses their agreement to set a target and the accumulated model to decide how to approach it. We define a candidate merged task vector relative to the common initialization as

$$M^g_t(\alpha) = \frac{tH^g_{t-1} + \alpha D^g_t}{t+1}, \qquad \alpha \geq 0. \tag{2}$$

This family preserves the relative weights already assigned to historical contributions and adjusts only the incoming coefficient; $\alpha \geq 0$ preserves its direction. One adaptive coefficient is assigned to each `nn.Linear`'s weight and bias, $A = [W \mid b]$. All other non-BN parameters follow Eq. 2 with $\alpha = 1$; Appendix C.4.3 evaluates this parameter scope.

**Resultant-guided target construction.** For each transported layer, the resultant $R_t$ and cumulative squared norm $E_t$ summarize all encountered task vectors. Omitting $g$ and initializing $R_0 = E_0 = 0$, we update

$$R_t = R_{t-1} + D_t, \qquad E_t = E_{t-1} + \|D_t\|^2_F. \tag{3}$$

The reference $C_t = R_t/(t+1)$ is the least-squares consensus: $\theta_0 + C_t$ minimizes the mean squared parameter distance to the pre-trained and encountered subject models (Proposition A.1). To measure agreement among their adaptations, define

$$r_t = \frac{\|R_t\|^2_F}{tE_t}, \qquad q_t = \frac{1 + tr_t}{t+1}. \tag{4}$$

Aligned task vectors of similar magnitude yield a larger $r_t$; cancellation and unequal magnitudes reduce it. This energy-normalized extension of mean resultant geometry (Mardia & Jupp, 2009) satisfies $r_t \in [0, 1]$, with $r_t = 1$ when $E_t = 0$. Its relation to task-vector dispersion is established in Appendix A.2. The adjusted score $q_t \in [1/(t+1), 1]$ is strictly positive.

We use task-vector agreement to scale the consensus while retaining its direction, with transport exponent $\beta$:

$$s_t = q_t^{-\beta/2}, \qquad \widehat{M}_t = s_t C_t = s_t \frac{R_t}{t+1}. \tag{5}$$

Positive $\beta$ permits expansion to offset task-vector cancellation, negative $\beta$ contracts the target toward the pre-trained model, and $\beta = 0$ recovers the consensus. Scaling controls adaptation strength, as in task arithmetic and TIES (Ilharco et al., 2023; Yadav et al., 2023); here its layer-specific value follows the observed agreement. A single dataset-level $\beta$ is fixed across layers and arrivals (Appendix B.5).

**History-weighted projection.** The target treats encountered task vectors symmetrically, while the recursive update retains the relative weights of historical contributions. We therefore use $\widehat{M}_t$ as a reference and select the incoming coefficient within Eq. 2:

$$\alpha_t = \arg\min_{\alpha \geq 0} \left\| M_t(\alpha) - \widehat{M}_t \right\|^2_F. \tag{6}$$

Define the residual needed to reach the target from the weighted historical task vector as

$$K_t = (t+1)\widehat{M}_t - tH_{t-1}. \tag{7}$$

Projecting $K_t$ onto the nonnegative ray spanned by $D_t$ gives the closed-form solution

$$\boxed{\alpha_t = \max\left(0, \frac{\langle D_t, K_t \rangle}{\|D_t\|_F^2}\right)}, \qquad M_t = M_t(\alpha_t). \tag{8}$$

For $D_t = 0$, we set $\alpha_t = 0$. The selected merged task vector yields the recursive parameter update

$$\theta_t^{M,g} = \theta_{t-1}^{M,g} + \frac{\alpha_t^g D_t^g - H_{t-1}^g}{t+1}. \tag{9}$$

Equivalently, $\theta_t^{M,g} = \theta_0^g + M_t^g$. At the next arrival, we recover $H_t^g$ from the merged model using Eq. 1.

Thus, agreement determines the target scale, while alignment with the residual determines the incoming contribution. Equation 8 is optimal within the history-weighted update family and reduces to recursive parameter averaging when $\beta = 0$ throughout the stream (Appendix A.3).

### 3.3 Affine-Map BN Merging

BN is a widely used normalization component whose running moments and affine parameters jointly determine its inference function, so independent averaging of these quantities need not realize the corresponding average inference map. AM-BN merges the slope and offset of BN inference maps, then reconstructs the BN state.

For BN channel $h$ in checkpoint $i$, let $(\mu_{i,h}, v_{i,h})$ be its running mean and variance ($v_{i,h} > 0$), and $(\gamma_{i,h}, \zeta_{i,h})$ its affine parameters. We omit the layer index. At inference, BN is the channel-wise transformation (Ioffe & Szegedy, 2015)

$$y = a_{i,h}x + o_{i,h}, \qquad a_{i,h} = \frac{\gamma_{i,h}}{\sqrt{v_{i,h} + \epsilon}}, \qquad o_{i,h} = \zeta_{i,h} - a_{i,h}\mu_{i,h}, \tag{10}$$

where $\epsilon > 0$ is the numerical constant used by BN.

Using the Gaussian proxy $P_{i,h} = \mathcal{N}(\mu_{i,h}, v_{i,h})$ for the stored feature statistics, we measure each subject model's compatibility with the pre-trained model through the symmetric KL divergence:

$$\begin{aligned} \kappa_{i,h} &= \tfrac{1}{2}\left[\mathrm{KL}(P_{0,h} \| P_{i,h}) + \mathrm{KL}(P_{i,h} \| P_{0,h})\right], \\ \omega_{i,h} &= (1 + \kappa_{i,h})^{-1}. \end{aligned} \tag{11}$$

A larger discrepancy yields a smaller weight for the corresponding subject-specific change in the affine-map coefficients.

The mean and variance of the equal-weight mixture of the pre-trained and subject distributions are

$$\begin{aligned} \bar{\mu}_{t,h} &= \frac{\mu_{0,h} + \sum_{i=1}^{t} \mu_{i,h}}{t+1}, \\ \bar{v}_{t,h} &= \frac{v_{0,h} + \mu_{0,h}^2 + \sum_{i=1}^{t}(v_{i,h} + \mu_{i,h}^2)}{t+1} - \bar{\mu}_{t,h}^2. \end{aligned} \tag{12}$$

Writing $u_{i,h} = (a_{i,h}, o_{i,h})$, we merge changes in these coefficients relative to the pre-trained model as

$$u_{t,h}^M = u_{0,h} + \frac{1}{t+1}\sum_{i=1}^{t} \omega_{i,h}(u_{i,h} - u_{0,h}), \qquad u_{t,h}^M = (a_{t,h}^M, o_{t,h}^M). \tag{13}$$

The fixed denominator $t + 1$ preserves the attenuation of less compatible subject-specific shifts. We assign $(\bar{\mu}_{t,h}, \bar{v}_{t,h})$ as the running moments of the merged BN layer and reconstruct its affine parameters:

$$\gamma_{t,h}^M = a_{t,h}^M \sqrt{\bar{v}_{t,h} + \epsilon}, \qquad \zeta_{t,h}^M = o_{t,h}^M + a_{t,h}^M \bar{\mu}_{t,h}. \tag{14}$$

Using the same running variance and $\epsilon$ during reconstruction and inference, the merged BN layer exactly realizes the transformation $a_{t,h}^M x + o_{t,h}^M$. Appendix A.4 establishes the consistency of the mixture moments, exact affine-map reconstruction, and compatibility-weighted attenuation of individual subject-specific coefficient changes.

All summaries are updated incrementally. Each merging step requires $O(P)$ arithmetic for a model with $P$ parameters, while the auxiliary storage remains independent of the number of encountered subjects. Algorithm 1 and the storage requirements are provided in Appendix A.5.

# 4 EXPERIMENTS

We evaluate whether RPT improves continual decoding, how it balances prior-subject retention and incoming adaptation, and which components account for these gains. Robustness studies and sequential-learning comparisons assess the resulting method beyond the main checkpoint stream.

## 4.1 EXPERIMENTAL SETUP

Table 1: Overview of the three EEG datasets used in our experiments.

| Dataset | Task | Subjects | Classes | Channels | Sampling rate (Hz) | Duration (s) |
|---|---|---|---|---|---|---|
| Physionet-MI | Motor imagery | 103 | 4 | 64 | 160 | 4 |
| FACED | Emotion recognition | 123 | 9 | 32 | 250 | 30 |
| ISRUC | Sleep staging | 98 | 5 | 8 | 100 | 30 |

**Subject-incremental evaluation.** We evaluate motor imagery, emotion recognition, and sleep staging on Physionet-MI (Schalk, 2009), FACED (Chen et al., 2023), and ISRUC (Khalighi et al., 2016) (Table 1). Each dataset is randomly divided into disjoint pre-training and incremental subject pools at an approximate 3:7 ratio. Subjects in the incremental pool are introduced one at a time as incremental domains (Zhou et al., 2025a). A fixed 9:1 training-to-testing ratio is applied to every newly introduced subject, facilitating the fine-tuning of subject-specific models and the subsequent evaluation of the merged model. Main comparisons use ten random subject orders over the same incremental pool; the split-robustness experiment additionally varies the pre-training/incremental subject split. Appendix B provides preprocessing, architecture, implementation, and reporting details.

**Models and baselines.** We adopt the compact CNN–Transformer architecture of Zhou et al. (2025a), with details in Appendix B.3. We compare RPT with weight averaging (WA) (Izmailov et al., 2018), Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), MagMax (Marczak et al., 2025), OPCM (Tang et al., 2025), NUFILT (Qiu et al., 2026), WUDI-Merging (Cheng et al., 2025), and DOP (Yang et al., 2025), and additionally evaluate AdaMerging (Yang et al., 2024) and MINGLE (Qiu et al., 2025) with test-time adaptation (Appendix B.4). We select one dataset-level transport exponent $\beta$ by three-fold cross-validation within the pre-training subject pool (Appendix B.5). It remains fixed across layers, subject arrivals, and evaluation orders.

**Metrics.** We evaluate the final merged model using average accuracy (ACC) and macro-F1 (MF1) across all encountered subjects. Backward transfer (BWT) (Lin et al., 2022), reported as BWT-ACC and BWT-MF1 (abbreviated as $\text{BWT}_A$ and $\text{BWT}_F$), measures the average performance change on earlier subjects from immediately after incorporation to the final merge. Positive BWT indicates improvement and negative BWT indicates forgetting; higher BWT indicates better retention at comparable ACC or MF1. Unless otherwise stated, CMM results are reported as mean $\pm$ standard deviation (SD) across ten random subject orders.

## 4.2 OVERALL CONTINUAL MODEL MERGING PERFORMANCE

**Continual EEG model merging.** RPT achieves the highest ACC and MF1 on all three EEG datasets among the compared CMM methods (Table 2). Its ACC/MF1 gains over the strongest baseline for each metric in the native CMM block are $4.2/4.4$, $6.7/6.6$, $1.4/5.5$ percentage points on Physionet-MI, FACED, and ISRUC, respectively. These consistent gains support effective checkpoint merging across distinct decoding tasks. Pre-trained, Independent FT, and Sequential FT provide reference models.

**Transport gains under matched normalization.** RPT remains best on all six ACC/MF1 endpoints when the eight compatible baselines use the same AM-BN (lower block of Table 2). Holding normalization fixed establishes the additional benefit of its parameter update. AM-BN also improves

Table 2: Continual EEG model merging (ACC/MF1 in %; BWT in pp). All metrics are higher-is-better. CMM rows report mean$_{\pm SD}$ over ten orders; reference rows show means. Bold marks the best merging method in each block; † denotes using test-time adaptation. RPT includes AM-BN.

| Method | Physionet-MI | | | | FACED | | | | ISRUC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ |
| *Reference* | | | | | | | | | | | | |
| Pre-trained | 48.0 | 39.5 | – | – | 41.2 | 31.3 | – | – | 70.9 | 54.3 | – | – |
| Independent FT | 59.3 | 53.6 | – | – | 50.6 | 40.3 | – | – | 83.7 | 70.1 | – | – |
| Sequential FT | 44.3 | 35.8 | -15.2 | -18.0 | 41.4 | 32.5 | -10.4 | -9.9 | 60.2 | 46.6 | -23.5 | -24.5 |
| *CMM* | | | | | | | | | | | | |
| WA | 49.3$_{\pm 2.9}$ | 42.2$_{\pm 3.5}$ | 0.1$_{\pm 0.9}$ | -0.1$_{\pm 1.0}$ | 51.4$_{\pm 2.2}$ | 41.8$_{\pm 2.6}$ | 0.0$_{\pm 0.9}$ | 0.1$_{\pm 1.0}$ | 58.2$_{\pm 16.8}$ | 43.7$_{\pm 17.3}$ | -0.6$_{\pm 0.7}$ | -0.6$_{\pm 0.7}$ |
| Task Arithmetic | 26.2$_{\pm 0.0}$ | 11.1$_{\pm 0.0}$ | -2.8$_{\pm 2.0}$ | -6.6$_{\pm 1.6}$ | 24.7$_{\pm 0.0}$ | 17.5$_{\pm 0.0}$ | -10.2$_{\pm 2.4}$ | -8.4$_{\pm 2.2}$ | 45.3$_{\pm 0.0}$ | 23.0$_{\pm 0.0}$ | -11.4$_{\pm 2.6}$ | -14.4$_{\pm 2.0}$ |
| TIES | 48.6$_{\pm 0.7}$ | 40.5$_{\pm 0.8}$ | -0.1$_{\pm 0.8}$ | -0.1$_{\pm 0.9}$ | 41.3$_{\pm 1.5}$ | 31.4$_{\pm 1.5}$ | -0.2$_{\pm 1.7}$ | 0.3$_{\pm 1.7}$ | 70.9$_{\pm 1.6}$ | 54.0$_{\pm 1.7}$ | -1.8$_{\pm 1.5}$ | -2.0$_{\pm 1.7}$ |
| MagMax | 42.9$_{\pm 0.0}$ | 35.8$_{\pm 0.0}$ | -2.0$_{\pm 0.9}$ | -1.6$_{\pm 0.7}$ | 42.7$_{\pm 0.0}$ | 32.8$_{\pm 0.0}$ | 0.9$_{\pm 0.4}$ | 1.0$_{\pm 0.5}$ | 66.9$_{\pm 0.0}$ | 49.8$_{\pm 0.0}$ | -1.5$_{\pm 1.2}$ | -1.7$_{\pm 1.5}$ |
| OPCM | 25.7$_{\pm 0.3}$ | 12.0$_{\pm 0.7}$ | -0.7$_{\pm 1.4}$ | -4.2$_{\pm 2.0}$ | 12.5$_{\pm 0.0}$ | 6.5$_{\pm 0.0}$ | -1.7$_{\pm 2.2}$ | -2.1$_{\pm 1.5}$ | 31.9$_{\pm 2.2}$ | 12.7$_{\pm 1.9}$ | -3.0$_{\pm 3.9}$ | -4.8$_{\pm 3.9}$ |
| NUFILT | 47.4$_{\pm 4.2}$ | 40.3$_{\pm 5.3}$ | -1.9$_{\pm 2.3}$ | -1.9$_{\pm 2.9}$ | 51.7$_{\pm 2.5}$ | 41.9$_{\pm 2.7}$ | -0.2$_{\pm 1.4}$ | -0.1$_{\pm 1.2}$ | 57.1$_{\pm 16.0}$ | 42.8$_{\pm 16.5}$ | -1.5$_{\pm 1.7}$ | -1.4$_{\pm 1.9}$ |
| WUDI-Merging | 46.2$_{\pm 3.4}$ | 38.3$_{\pm 4.2}$ | -3.5$_{\pm 2.1}$ | -4.2$_{\pm 2.6}$ | 48.2$_{\pm 2.3}$ | 38.4$_{\pm 2.6}$ | -2.3$_{\pm 1.9}$ | -2.1$_{\pm 2.2}$ | 56.6$_{\pm 15.6}$ | 41.7$_{\pm 15.9}$ | -2.4$_{\pm 0.8}$ | -2.6$_{\pm 2.3}$ |
| DOP | 47.9$_{\pm 1.7}$ | 40.9$_{\pm 2.5}$ | -9.0$_{\pm 2.1}$ | -10.3$_{\pm 2.5}$ | 36.3$_{\pm 14.5}$ | 28.1$_{\pm 12.7}$ | -7.4$_{\pm 16.4}$ | -6.7$_{\pm 14.4}$ | 67.6$_{\pm 5.5}$ | 52.7$_{\pm 8.2}$ | -10.6$_{\pm 5.6}$ | -10.6$_{\pm 8.1}$ |
| AdaMerging† | 26.2$_{\pm 0.0}$ | 11.1$_{\pm 0.0}$ | -1.2$_{\pm 0.6}$ | -3.1$_{\pm 0.8}$ | 12.4$_{\pm 0.9}$ | 7.4$_{\pm 0.8}$ | -8.2$_{\pm 2.6}$ | -6.7$_{\pm 1.9}$ | 18.3$_{\pm 5.1}$ | 8.8$_{\pm 2.3}$ | -12.7$_{\pm 4.7}$ | -10.2$_{\pm 2.1}$ |
| MINGLE† | 44.8$_{\pm 3.4}$ | 36.3$_{\pm 4.2}$ | -3.2$_{\pm 3.1}$ | -4.9$_{\pm 4.5}$ | 37.8$_{\pm 16.6}$ | 30.2$_{\pm 14.5}$ | -10.2$_{\pm 16.3}$ | -7.9$_{\pm 14.3}$ | 42.8$_{\pm 12.8}$ | 26.7$_{\pm 12.3}$ | -12.7$_{\pm 13.1}$ | -16.6$_{\pm 12.8}$ |
| **RPT** | **53.5**$_{\pm 0.1}$ | **46.6**$_{\pm 0.1}$ | **0.2**$_{\pm 0.9}$ | **0.4**$_{\pm 0.8}$ | **58.4**$_{\pm 0.2}$ | **48.5**$_{\pm 0.3}$ | **1.8**$_{\pm 0.9}$ | **2.1**$_{\pm 1.0}$ | **72.3**$_{\pm 0.0}$ | **59.5**$_{\pm 0.0}$ | **-0.1**$_{\pm 0.2}$ | **-0.2**$_{\pm 0.5}$ |
| *CMM with AM-BN* | | | | | | | | | | | | |
| WA | 52.4$_{\pm 0.0}$ | 45.9$_{\pm 0.0}$ | -0.6$_{\pm 0.9}$ | -0.6$_{\pm 0.7}$ | 52.9$_{\pm 0.0}$ | 42.6$_{\pm 0.0}$ | 0.7$_{\pm 0.9}$ | 0.7$_{\pm 0.8}$ | 71.9$_{\pm 0.0}$ | 58.7$_{\pm 0.0}$ | -0.3$_{\pm 0.3}$ | -0.6$_{\pm 0.3}$ |
| Task Arithmetic | 26.2$_{\pm 0.0}$ | 11.1$_{\pm 0.0}$ | -4.2$_{\pm 1.9}$ | -8.5$_{\pm 1.8}$ | 25.5$_{\pm 0.0}$ | 16.6$_{\pm 0.0}$ | -13.7$_{\pm 2.3}$ | -12.6$_{\pm 2.4}$ | 47.4$_{\pm 0.0}$ | 24.9$_{\pm 0.0}$ | -11.3$_{\pm 2.3}$ | -14.4$_{\pm 2.4}$ |
| TIES | 51.5$_{\pm 0.9}$ | 45.1$_{\pm 1.1}$ | -1.6$_{\pm 1.3}$ | -1.3$_{\pm 1.4}$ | 51.1$_{\pm 1.0}$ | 41.1$_{\pm 0.9}$ | -0.1$_{\pm 1.3}$ | 0.3$_{\pm 1.1}$ | 70.7$_{\pm 0.9}$ | 57.8$_{\pm 0.8}$ | -1.2$_{\pm 0.9}$ | -1.5$_{\pm 0.8}$ |
| MagMax | 49.4$_{\pm 0.0}$ | 42.7$_{\pm 0.0}$ | -1.1$_{\pm 0.8}$ | -1.0$_{\pm 0.7}$ | 52.9$_{\pm 0.0}$ | 42.4$_{\pm 0.0}$ | 1.0$_{\pm 1.3}$ | 0.7$_{\pm 1.3}$ | 71.2$_{\pm 0.0}$ | 57.6$_{\pm 0.0}$ | -0.5$_{\pm 0.5}$ | -1.2$_{\pm 0.6}$ |
| OPCM | 24.3$_{\pm 0.2}$ | 12.4$_{\pm 0.5}$ | -1.7$_{\pm 1.3}$ | -4.8$_{\pm 1.4}$ | 12.3$_{\pm 0.8}$ | 7.7$_{\pm 0.6}$ | -6.2$_{\pm 3.1}$ | -4.5$_{\pm 2.2}$ | 31.1$_{\pm 4.5}$ | 14.3$_{\pm 2.9}$ | -4.9$_{\pm 4.4}$ | -4.9$_{\pm 3.1}$ |
| NUFILT | 51.7$_{\pm 1.2}$ | 44.9$_{\pm 1.4}$ | -2.1$_{\pm 1.8}$ | -2.3$_{\pm 1.6}$ | 54.0$_{\pm 1.5}$ | 44.1$_{\pm 1.4}$ | 0.4$_{\pm 1.6}$ | 0.6$_{\pm 1.8}$ | 71.4$_{\pm 0.4}$ | 58.9$_{\pm 1.2}$ | -1.0$_{\pm 0.7}$ | -1.1$_{\pm 1.4}$ |
| WUDI-Merging | 49.6$_{\pm 1.8}$ | 42.7$_{\pm 1.8}$ | -3.4$_{\pm 3.1}$ | -3.9$_{\pm 2.8}$ | 51.3$_{\pm 1.4}$ | 41.4$_{\pm 1.6}$ | -0.6$_{\pm 1.8}$ | -0.3$_{\pm 2.0}$ | 70.7$_{\pm 0.6}$ | 57.1$_{\pm 1.1}$ | -2.3$_{\pm 0.7}$ | -3.2$_{\pm 1.2}$ |
| DOP | 49.7$_{\pm 2.0}$ | 43.0$_{\pm 2.4}$ | -5.4$_{\pm 2.6}$ | -6.4$_{\pm 2.9}$ | 53.9$_{\pm 1.2}$ | 43.9$_{\pm 1.1}$ | 0.0$_{\pm 1.5}$ | 0.4$_{\pm 1.6}$ | 70.9$_{\pm 2.2}$ | 57.5$_{\pm 2.5}$ | -1.6$_{\pm 2.3}$ | -2.6$_{\pm 2.4}$ |
| **RPT** | **53.5**$_{\pm 0.1}$ | **46.6**$_{\pm 0.1}$ | **0.2**$_{\pm 0.9}$ | **0.4**$_{\pm 0.8}$ | **58.4**$_{\pm 0.2}$ | **48.5**$_{\pm 0.3}$ | **1.8**$_{\pm 0.9}$ | **2.1**$_{\pm 1.0}$ | **72.3**$_{\pm 0.0}$ | **59.5**$_{\pm 0.0}$ | **-0.1**$_{\pm 0.2}$ | **-0.2**$_{\pm 0.5}$ |

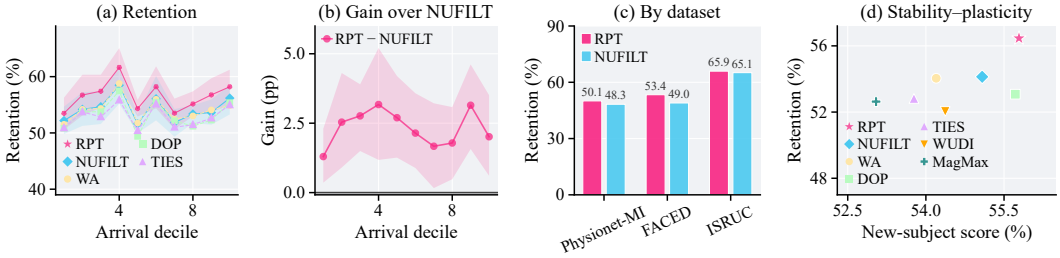



Figure 3: Final prior-subject retention under matched AM-BN. (a) and (b) group earlier subjects by arrival-position decile; (c) reports dataset-level scores, and (d) compares retention with performance immediately after each subject's incorporation. Scores average ACC and MF1 over ten orders, with equal dataset weights in aggregate panels; gains are in percentage points (pp). Bands in (a) and (b) show 95% confidence intervals.

most baseline outcomes, showing that its value extends beyond RPT. The gains therefore reflect both the merging rule and the treatment of normalization.

### 4.3 Retention and New-Subject Adaptation

RPT improves prior-subject retention while maintaining competitive new-subject adaptation (Figure 3). Under matched AM-BN, its aggregate final score on earlier subjects exceeds the strongest retention baseline, NUFILT, by 2.33 points (95% CI $[1.93, 2.71]$), with a positive gain in all ten orders. The advantage spans all three datasets and every aggregate arrival-position group (a–c). Meanwhile, performance immediately after incorporating new subjects remains comparable to DOP (55.79% versus 55.72%; d). The history-weighted update thus preserves more accumulated decoding ability while retaining competitive incoming adaptation.

Table 3: Factorial ablation (mean$_{\pm SD}$ over ten orders; ACC/MF1 in %, BWT in pp). Removing transport uniformly averages parameters, including the pre-trained model; removing AM-BN averages BN parameters and moments independently. Bold marks the best unrounded mean.

| Method | Physionet-MI | | | | FACED | | | | ISRUC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ |
| w/o both | 50.9$_{\pm 0.0}$ | 44.7$_{\pm 0.0}$ | -0.5$_{\pm 1.0}$ | -0.7$_{\pm 0.9}$ | 8.2$_{\pm 0.0}$ | 4.3$_{\pm 0.0}$ | -19.4$_{\pm 8.6}$ | -16.3$_{\pm 7.3}$ | **73.0**$_{\pm 0.0}$ | 56.8$_{\pm 0.0}$ | -0.2$_{\pm 0.8}$ | -0.5$_{\pm 1.4}$ |
| w/o AM-BN | 49.0$_{\pm 0.1}$ | 42.5$_{\pm 0.1}$ | -1.5$_{\pm 1.0}$ | -1.5$_{\pm 1.1}$ | 7.8$_{\pm 0.0}$ | 4.0$_{\pm 0.0}$ | -20.7$_{\pm 7.5}$ | -17.2$_{\pm 6.6}$ | 72.9$_{\pm 0.0}$ | 57.1$_{\pm 0.0}$ | -0.1$_{\pm 0.9}$ | **0.2**$_{\pm 1.4}$ |
| w/o transport | 52.4$_{\pm 0.0}$ | 45.9$_{\pm 0.0}$ | -0.7$_{\pm 0.9}$ | -0.6$_{\pm 0.9}$ | 52.5$_{\pm 0.0}$ | 42.0$_{\pm 0.0}$ | 0.4$_{\pm 0.7}$ | 0.2$_{\pm 0.6}$ | 71.8$_{\pm 0.0}$ | 58.7$_{\pm 0.0}$ | -0.4$_{\pm 0.2}$ | -0.6$_{\pm 0.3}$ |
| **RPT (full)** | **53.5**$_{\pm 0.1}$ | **46.6**$_{\pm 0.1}$ | **0.2**$_{\pm 0.9}$ | **0.4**$_{\pm 0.8}$ | **58.4**$_{\pm 0.2}$ | **48.5**$_{\pm 0.3}$ | **1.8**$_{\pm 0.9}$ | **2.1**$_{\pm 1.0}$ | 72.3$_{\pm 0.0}$ | **59.5**$_{\pm 0.0}$ | **-0.1**$_{\pm 0.2}$ | -0.2$_{\pm 0.5}$ |

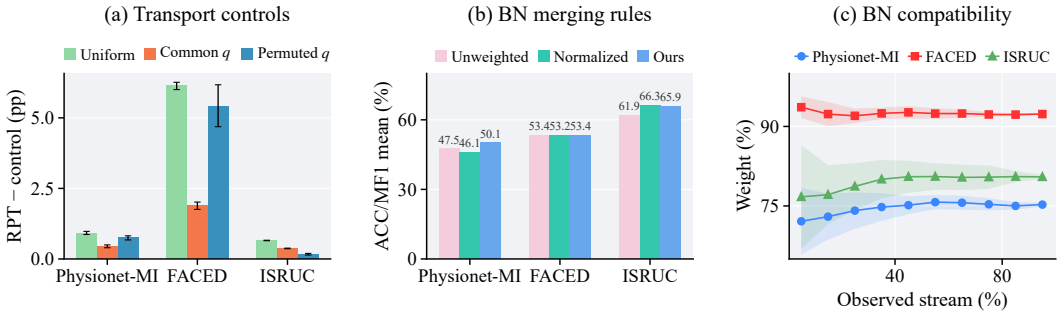



Figure 4: Mechanism controls with matched checkpoints and orders: (a) remove or mismatch layer-specific agreement, (b) vary the BN merging rule (Unweighted denotes affine-map averaging, Normalized denotes normalized overlap, Ours denotes AM-BN), and (c) track BN compatibility. Error bars in (a) show 95% CIs; bands in (c) show SD. Performance averages ACC and MF1; differences in (a) are in pp.

### 4.4 Component and Mechanism Analysis

**Complementary roles of transport and normalization.** Full RPT achieves the highest MF1 and BWT-ACC across datasets (Table 3). With AM-BN fixed, transport improves both decoding metrics and both BWT measures. AM-BN is particularly beneficial on FACED. On ISRUC, full RPT gains 2.7 MF1 points over "w/o both" with a 0.7-point ACC decrease. With GroupNorm (GN) replacing BN, RPT also has the highest mean ACC/MF1 on all three datasets (Appendix C.2).

**Layer-specific agreement guides the update.** Matching each layer to its own agreement score improves merging on all three datasets (Figure 4a). A common score underperforms RPT, and permuting scores across layers causes a larger aggregate loss. Since these controls preserve the checkpoints and BN rule, the benefit depends on where the agreement-derived scale is applied, rather than scaling alone.

**Stored statistics inform affine-map merging.** With transport fixed, compatibility-weighted AM-BN gives the highest aggregate score among the compared rules, exceeding unweighted affine-map averaging by 2.18 points (Figure 4b; Appendix C.4.1). Its weights follow distinct, stable profiles across subject streams (c), supporting the use of stored statistics to modulate changes in affine-map coefficients relative to the pre-trained model.

### 4.5 Sensitivity to the Transport Exponent

The exponent adjusts adaptation strength without changing the merging rule (Figure 5). Physionet-MI favors contraction, FACED benefits from expansion up to a moderate range, and ISRUC peaks near a small positive exponent; excessive expansion weakens decoding and retention. The gains do not require a separate exponent for each dataset: a shared $\beta = 0.5$ exceeds all eight AM-BN baselines on every ACC/MF1 endpoint (Appendix B.5).

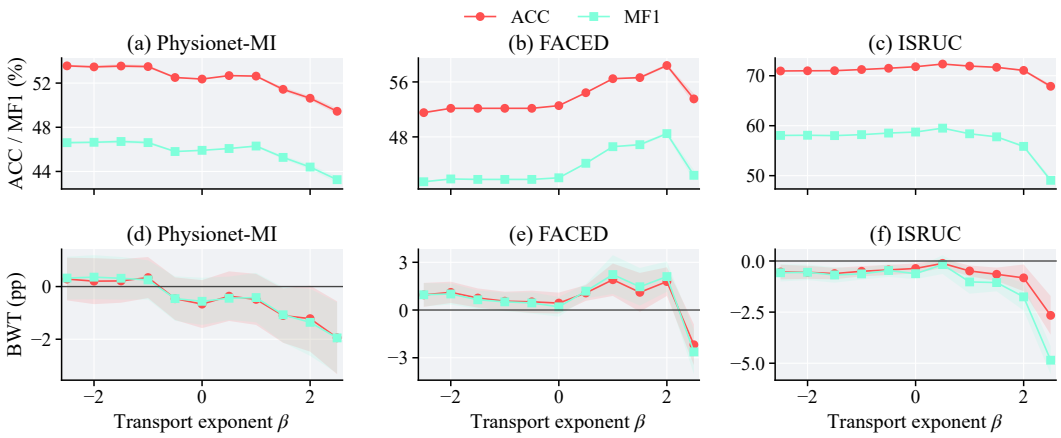



Figure 5: Sensitivity to $\beta$ over ten subject orders (mean $\pm$ SD). Top: ACC and MF1 (%); bottom: their BWT (pp). All eleven grid points are shown for each dataset.

Table 4: Conventional continual learning comparison (mean$_{\pm \mathrm{SD}}$ over ten orders; ACC/MF1 in %, BWT in pp). Bold marks the best result; the last row is Seq-RPT minus the strongest baseline.

| Method | Physionet-MI | | | | FACED | | | | ISRUC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ | ACC | MF1 | BWT$_A$ | BWT$_F$ |
| EWC | 49.1$_{\pm 2.4}$ | 40.9$_{\pm 2.6}$ | -9.8$_{\pm 2.7}$ | -11.6$_{\pm 2.8}$ | **52.3**$_{\pm 15.2}$ | **43.3**$_{\pm 13.8}$ | -4.8$_{\pm 15.6}$ | -4.4$_{\pm 14.0}$ | 42.0$_{\pm 13.4}$ | 21.8$_{\pm 14.0}$ | -15.4$_{\pm 4.5}$ | -15.8$_{\pm 4.1}$ |
| ER | **53.4**$_{\pm 2.1}$ | **46.4**$_{\pm 2.5}$ | -5.3$_{\pm 2.6}$ | -6.3$_{\pm 3.1}$ | 18.2$_{\pm 9.5}$ | 12.3$_{\pm 8.2}$ | -21.2$_{\pm 12.4}$ | -19.2$_{\pm 11.2}$ | 73.8$_{\pm 2.7}$ | 60.6$_{\pm 2.6}$ | -9.0$_{\pm 2.6}$ | -10.0$_{\pm 2.8}$ |
| DER++ | 52.8$_{\pm 2.7}$ | 45.1$_{\pm 3.1}$ | -4.4$_{\pm 2.1}$ | -5.4$_{\pm 2.2}$ | 27.6$_{\pm 21.8}$ | 21.3$_{\pm 19.8}$ | -11.0$_{\pm 19.5}$ | -9.8$_{\pm 17.3}$ | 74.0$_{\pm 3.8}$ | 59.5$_{\pm 4.8}$ | -3.1$_{\pm 3.6}$ | -4.2$_{\pm 4.2}$ |
| MUDVI | 50.4$_{\pm 2.5}$ | 43.4$_{\pm 2.6}$ | -7.0$_{\pm 3.0}$ | -8.2$_{\pm 3.0}$ | 50.9$_{\pm 13.9}$ | 41.2$_{\pm 13.3}$ | **1.2**$_{\pm 16.5}$ | 0.4$_{\pm 15.4}$ | 68.6$_{\pm 11.8}$ | 54.3$_{\pm 12.7}$ | -14.6$_{\pm 12.0}$ | -16.8$_{\pm 12.9}$ |
| ReSNT | 52.7$_{\pm 1.8}$ | 45.8$_{\pm 1.7}$ | -6.4$_{\pm 2.4}$ | -7.1$_{\pm 2.9}$ | 17.4$_{\pm 9.8}$ | 11.5$_{\pm 8.3}$ | -21.9$_{\pm 14.3}$ | -20.0$_{\pm 12.7}$ | **74.2**$_{\pm 2.6}$ | **61.0**$_{\pm 2.3}$ | -8.6$_{\pm 2.7}$ | -9.5$_{\pm 2.5}$ |
| AMBM | 51.9$_{\pm 2.1}$ | 45.4$_{\pm 2.5}$ | -7.2$_{\pm 1.3}$ | -7.9$_{\pm 1.4}$ | 22.3$_{\pm 22.8}$ | 16.5$_{\pm 21.1}$ | -17.8$_{\pm 23.4}$ | -16.1$_{\pm 21.3}$ | 67.8$_{\pm 7.1}$ | 53.2$_{\pm 8.3}$ | -10.0$_{\pm 7.2}$ | -10.3$_{\pm 8.4}$ |
| BrainUICL-SR | 52.1$_{\pm 2.2}$ | 45.0$_{\pm 2.3}$ | -3.3$_{\pm 2.6}$ | -3.7$_{\pm 2.7}$ | 20.0$_{\pm 14.1}$ | 13.7$_{\pm 12.1}$ | -23.8$_{\pm 17.6}$ | -21.8$_{\pm 15.3}$ | 71.4$_{\pm 2.6}$ | 57.2$_{\pm 2.3}$ | -5.7$_{\pm 2.1}$ | -6.3$_{\pm 1.8}$ |
| ProNECL | 43.0$_{\pm 4.0}$ | 34.5$_{\pm 4.9}$ | -15.2$_{\pm 4.5}$ | -17.4$_{\pm 5.4}$ | 49.5$_{\pm 14.1}$ | 40.7$_{\pm 13.0}$ | -2.3$_{\pm 13.7}$ | -1.7$_{\pm 12.6}$ | 62.2$_{\pm 9.0}$ | 45.7$_{\pm 10.4}$ | -19.3$_{\pm 9.7}$ | -19.6$_{\pm 11.5}$ |
| **Seq-RPT** | 52.3$_{\pm 0.3}$ | 45.1$_{\pm 0.3}$ | **-0.7**$_{\pm 0.5}$ | **-0.7**$_{\pm 0.4}$ | 51.8$_{\pm 4.3}$ | 41.6$_{\pm 4.7}$ | 0.4$_{\pm 2.9}$ | **0.5**$_{\pm 3.1}$ | 72.5$_{\pm 0.1}$ | 58.3$_{\pm 0.3}$ | **-0.3**$_{\pm 0.3}$ | **-0.4**$_{\pm 0.5}$ |
| Seq-RPT - Best | -1.1 | -1.3 | +2.6 | +3.0 | -0.5 | -1.7 | -0.8 | +0.1 | -1.7 | -2.7 | +2.8 | +3.8 |

### 4.6 Sequential Adaptation and Resource Use

**Competitive decoding with improved retention.** Seq-RPT remains competitive with conventional CL methods while achieving the highest BWT in five of six dataset–metric comparisons (Table 4). Its ACC and MF1 are within 0.5–2.7 points of the strongest baselines. This variant applies RPT to sequential updates (Appendix B.4), merging checkpoints without accessing subject EEG or replaying historical recordings. The results extend its retention advantage to sequential adaptation.

**Computational and storage costs.** RPT merges each subject checkpoint in 0.42–0.48 seconds with 8.12 MiB peak allocated GPU memory, giving a 5.5–31.1$\times$ speedup over NUFILT, DOP, and WUDI-Merging under matched AM-BN. Its update costs $O(P)$ and maintains state independent of stream length. Including training, updating, and evaluation, Seq-RPT takes 6.7, 3.7, and 9.4 seconds per subject on Physionet-MI, FACED, and ISRUC, with peak memory of 0.30, 0.47, and 1.55 GiB and no historical EEG replay. Appendix C.5 reports both resource comparisons.

## 5 Conclusion

We introduced RPT for continual EEG model merging from subject checkpoints. Task-vector agreement defines a merging target, a history-weighted projection determines the incoming contribution in closed form, and AM-BN merges the normalization transformations. Across three EEG tasks, this design improves complete-stream decoding and prior-subject retention while maintaining competitive new-subject adaptation. These results support checkpoint-derived geometry and BN inference maps for continual EEG adaptation without historical replay. We hope the proposed formulation and findings inform further research on model merging for continual EEG learning.

### AI USE STATEMENT

In this work, we used AI tools to assist with translation. Additionally, we used these tools for manuscript writing and language polishing, including drafting transitional sentences between sections, and for visual refinement of figures and icons. These tools also assisted with writing auxiliary scripts to support the experimental workflow and with data consistency checks. All code, including AI-assisted code, was manually reviewed and tested by the authors. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of AI tools.

### ETHICS STATEMENT

This study uses only publicly available EEG datasets: Physionet-MI, FACED, and ISRUC (Schalk, 2009; Chen et al., 2023; Khalighi et al., 2016). Our work involves secondary analysis of existing recordings, with no new participant recruitment or data collection. We acknowledge the sensitivity of EEG data and the importance of protecting participant privacy.

### REPRODUCIBILITY STATEMENT

Theoretical proofs are provided in Appendix A, and experimental details, including data preprocessing, evaluation protocols, model training, and hyperparameter settings, are documented in Appendix B. An anonymized ZIP archive containing the code is provided as supplementary material for review. The code will be made publicly available on GitHub after the review process concludes.

## REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/b704ea2c39778f07c617f6b7ce480e9e-Paper.pdf.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10(1):740, 2023.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xR9msNaREW.

Tiehang Duan, Zhenyi Wang, Gianfranco Doretto, Fang Li, Cui Tao, and Donald Adjeroh. Replay with stochastic neural transformation for online continual EEG classification. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1874–1879, 2023. doi: 10.1109/BIBM58861.2023.10385807.

Tiehang Duan, Zhenyi Wang, Fang Li, Gianfranco Doretto, Donald A. Adjeroh, Yiyi Yin, and Cui Tao. Online continual decoding of streaming EEG signal with a balanced and informative memory buffer. *Neural Networks*, 176:106338, 2024a. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106338. URL https://www.sciencedirect.com/science/article/pii/S0893608024002624.

Tiehang Duan, Zhenyi Wang, Li Shen, Gianfranco Doretto, Donald A. Adjeroh, Fang Li, and Cui Tao. Retain and adapt: Online sequential EEG classification with subject shift. *IEEE Transactions on Artificial Intelligence*, 5(9):4479–4492, 2024b. doi: 10.1109/TAI.2024.3385390.

Bradley J. Edelman, Shuailei Zhang, Gerwin Schalk, Peter Brunner, Gernot Müller-Putz, Cuntai Guan, and Bin He. Non-invasive brain-computer interfaces: State of the art and trends. *IEEE Reviews in Biomedical Engineering*, 18:26–49, 2025. doi: 10.1109/RBME.2024.3449790.

Yuyang Han, Ziyu Li, Zhiying Long, and Xia Wu. Consolidating memories, not data: A privacy-aware and stable framework for continual learning on EEG. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6786–6790, 2026. doi: 10.1109/ICASSP55912.2026.11462186.

Xiongfeng Huang, Yongyi Lu, and Zhijing Yang. CrossMindFlex: Privacy-preserving cross-subject EEG visual reconstruction with incremental adaptation. In *2026 International Conference on Signal Image Processing and Communication (ICSIPC)*, pp. 362–366, 2026. doi: 10.1109/ICSIPC69751.2026.11583932.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. arXiv preprint arXiv:2212.04089, 2023. URL https://arxiv.org/abs/2212.04089v3.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pp. 876–885. AUAI Press, 2018. URL https://auai.org/uai2018/proceedings/papers/313.pdf.

Jiarui Jin, Zongnan Chen, Honghua Cai, and Jiahui Pan. Affective EEG-based person identification with continual learning. *IEEE Transactions on Instrumentation and Measurement*, 73:1–16, 2024. doi: 10.1109/TIM.2024.3406836.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=gU5sJ6ZggcX.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. ISRUC-Sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0169260715002734.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Reinmar J Kobler, Jun ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pp7onaiM4VB.

Dan Li, Hye-Bin Shin, and Yeon-Woo Choi. Prototype-guided non-exemplar continual learning for cross-subject EEG decoding. In *2026 14th International Conference on Brain-Computer Interface (BCI)*, pp. 1–4, 2026a. doi: 10.1109/BCI69045.2026.11435042.

Dan Li, Hye-Bin Shin, and Seong-Whan Lee. Toward memory-efficient continual adaptation for MI-EEG decoding in BCIs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56 (1):766–778, 2026b. doi: 10.1109/TSMC.2025.3630757.

Shanglin Li, Chu Shiwen, Okan Koç, Yi Ding, Qibin Zhao, Motoaki Kawanabe, and Ziheng Chen. HEEGNet: Hyperbolic embeddings for EEG. In *The Fourteenth International Conference on Learning Representations*, 2026c. URL https://openreview.net/forum?id=CNDNRjpVIL.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 16165–16177. Curran Associates, Inc., 2022. doi: 10.52202/068431-1176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6728fcf94660c59c938319a6833a6073-Paper-Conference.pdf.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. MagMax: Leveraging model merging for seamless continual learning. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 379–395, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73013-9.

Kanti V Mardia and Peter E Jupp. *Directional statistics*. John Wiley & Sons, 2009.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gordon H. Bower (ed.), *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Angelo Porrello, Lorenzo Bonicelli, Pietro Buzzega, Monica Millunzi, Simone Calderara, and Rita Cucchiara. A second-order perspective on model compositionality and incremental learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 75098–75128, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bb0eb2b4c6e88ad546e20339f7a4783b-Paper-Conference.pdf.

Zihuan Qiu, Yi Xu, Chiyuan He, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. MINGLE: Mixture of null-space gated low-rank experts for test-time continual model merging. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8DCyv8x58O.

Zihuan Qiu, Lei Wang, Yang Cao, Runtong ZHANG, Bing Su, Yi Xu, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Null-space filtering for data-free continual model merging: Preserving stability, promoting plasticity. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 128492–128518, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d0da30e312b75a3fffd9e9191f8bc1b0-Paper-Conference.pdf.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Paper.pdf.

Gerwin Schalk. EEG motor movement/imagery dataset. *PhysioNet*, 2009. doi: 10.13026/C28G6P. URL https://physionet.org/content/eegmmidb/1.0.0/. Version 1.0.0.

Anke Tang, Enneng Yang, Li Shen, Yong Luo, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Merging on the fly without retraining: A sequential approach to scalable continual model merging. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=rdGMyTPhui.

Cheng Wan, Micky C. Nnamdi, Wenqi Shi, Benjamin Smith, Chad Purnell, and May D. Wang. Advancing sleep disorder diagnostics: A transformer-based EEG model for sleep stage classification and OSA prediction. *IEEE Journal of Biomedical and Health Informatics*, 29(2):878–886, 2025. doi: 10.1109/JBHI.2024.3512616.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xtaX3WyCj1.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 22743–22763, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/62868cc2fc1eb5cdf321d05b4b88510c-Paper-Conference.pdf.

Enneng Yang, Anke Tang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, and Jie Zhang. Continual model merging without data: Dual projections for balancing stability and plasticity. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 39275–39305. Curran Associates, Inc., 2025. doi: 10.52202/085713-1311. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/37d9f19150fce07bced2a81fc87d47a6-Paper-Conference.pdf.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Comput. Surv.*, 58(8), February 2026. ISSN 0360-0300. doi: 10.1145/3787849. URL https://doi.org/10.1145/3787849.

Qingzhu Zhang, Jiani Zhong, Zongsheng Li, Xinke Shen, and Quanying Liu. Multi-dataset joint pre-training of emotional eeg enables generalizable affective computing. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 165434–165461. Curran Associates, Inc., 2025. doi: 10.52202/085713-5515. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f1b8d443042f376aa3654b6c68de6297-Paper-Conference.pdf.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. BrainUICL: An unsupervised individual continual learning framework for EEG applications. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=6jjAYmppGQ.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. SPICED: A synaptic homeostasis-inspired framework for unsupervised continual EEG decoding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=qcdoHkkHcb.

# A THEORETICAL PROPERTIES

This appendix supplies the assumptions and proofs underlying Section 3, followed by the complete algorithm. We follow the notation of Section 3: $i$ indexes models, $t$ indexes subject arrivals, $g$ indexes layers, and $h$ indexes BN channels. Layer and channel indices are omitted where unambiguous, and matrix inner products are Frobenius inner products. The results describe the specified updates in exact arithmetic, rather than guaranteeing decoding accuracy or absence of forgetting.

## A.1 TASK VECTORS AND SUBJECT REPRESENTATIONS

Model Soups demonstrates that averaging weights of models fine-tuned from a shared initialization can improve accuracy without increasing inference time (Wortsman et al., 2022). Task Arithmetic expresses adaptation relative to a common pre-trained model and merges the resulting task vectors through scaled addition (Ilharco et al., 2023). TIES-Merging addresses interference from redundant parameter changes and sign conflicts through magnitude-based trimming and sign-consistent merging (Yadav et al., 2023). These works highlight the importance of parameter magnitudes and compatibility when merging knowledge across models. In continual EEG model merging, this perspective motivates controlling interference between successive subject adaptations while retaining knowledge from earlier subjects.

Understanding how task vectors encode subject-specific information provides a basis for designing such continual updates. Yang et al. (2025) connect task vectors to input subspaces through the accumulation of training gradients. Cheng et al. (2025) show that task vectors in linear layers can be approximated by linear combinations of their input representations. Qiu et al. (2026) further establish approximate alignment between task-vector and representation subspaces. Together, these analyses relate parameter changes to the feature information acquired during adaptation, motivating the use of task-vector geometry to characterize subject models.

To express this connection, consider a linear layer with weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Let $X_i \in \mathbb{R}^{N_i \times d_{\text{in}}}$ collect the input representations of subject $i$ at this layer, and let $D_i^W = W_i - W_0$ denote its weight task vector. The representation relationship developed in these studies can be summarized as

$$D_i^W \approx B_i X_i, \tag{15}$$

where $B_i \in \mathbb{R}^{d_{\text{out}} \times N_i}$ collects coefficients associated with the accumulated learning updates. This expression describes each row of the task vector as an approximate linear combination of the subject's input features, linking parameter adaptation to directions in the input representation subspace. For an affine operator $[W \mid b]$, the same interpretation uses input representations augmented with a constant coordinate.

In subject-incremental EEG decoding, this connection motivates treating each subject task vector as a parameter-space representation of the feature information acquired during local adaptation. Task-vector geometry therefore provides an indirect basis for characterizing similarities and differences among subject adaptations without accessing their EEG recordings. RPT follows this perspective by summarizing cross-subject task-vector agreement through the resultant and using it to guide each incoming contribution to the merged model.

## A.2 CONSENSUS TARGET AND RESULTANT PROPERTIES

**Proposition A.1** (Least-squares consensus)**.** *For the task vectors of a fixed transported layer and $t \geq 1$, let $D_0 = 0$. The unique minimizer of*

$$\mathcal{J}_t(X) = \frac{1}{2(t+1)} \sum_{i=0}^{t} \|X - D_i\|_F^2 \tag{16}$$

*is $C_t = R_t/(t+1)$, and for every $X$,*

$$\mathcal{J}_t(X) - \mathcal{J}_t(C_t) = \frac{1}{2}\|X - C_t\|_F^2. \tag{17}$$

*Proof.* Using $R_t = \sum_{i=1}^{t} D_i$ and $E_t = \sum_{i=1}^{t} \|D_i\|_F^2$, completing the square gives

$$\mathcal{J}_t(X) = \frac{1}{2}\|X - C_t\|_F^2 + \frac{E_t}{2(t+1)} - \frac{1}{2}\|C_t\|_F^2.$$

The last two terms are independent of $X$, and the squared term vanishes only at $X = C_t$, proving uniqueness and the stated gap. □

Thus $\theta_0 + C_t = (\theta_0 + \sum_{i=1}^{t} \theta_i)/(t+1)$ is the equal-weight parameter consensus for this layer.

For $E_t > 0$, let $\bar{D}_t = R_t/t$. Expanding the squared deviations yields

$$\sum_{i=1}^{t} \|D_i - \bar{D}_t\|_F^2 = E_t - \frac{\|R_t\|_F^2}{t} = E_t(1 - r_t). \tag{18}$$

Together with $r_t \geq 0$, this identity gives $0 \leq r_t \leq 1$. It identifies $1 - r_t$ as the energy-normalized dispersion of subject adaptations, with $r_t = 1$ exactly when $D_1 = \cdots = D_t$.

Mardia & Jupp (2009) define the mean resultant length as the norm of the mean of unit direction vectors. For $D_i = dU_i$, with $d > 0$ and $\|U_i\|_F = 1$, vectorizing the matrices shows that our score recovers its square:

$$r_t = \left\| \frac{1}{t} \sum_{i=1}^{t} U_i \right\|_F^2. \tag{19}$$

Since $q_t = (1 + tr_t)/(t+1) \in [1/(t+1), 1]$, the scale $s_t = q_t^{-\beta/2}$ is positive for every fixed $\beta \in \mathbb{R}$. It satisfies $s_t \geq 1$ for $\beta \geq 0$, $s_t \leq 1$ for $\beta < 0$, and $s_t = 1$ for $\beta = 0$, providing an agreement-based adjustment of the consensus magnitude. Under $D_i \mapsto cD_i$ with $c \in \mathbb{R} \setminus \{0\}$, $R_t \mapsto cR_t$ and $E_t \mapsto c^2 E_t$, leaving $r_t$, $q_t$, and $s_t$ unchanged. Because $R_t$ and $E_t$ are symmetric sums, the target $\widehat{M}_t = s_t C_t$ is invariant to permutations of a fixed set of subject models. If $E_t = 0$, all task vectors vanish; the convention $r_t = 1$ gives $q_t = s_t = 1$ and $\widehat{M}_t = 0$.

### A.3 Exact Projection and Recursive Averaging

For a fixed transported layer, Eq. 2 induces the parameter update

$$\begin{aligned} \theta_0 + M_t &= \theta_0 + H_{t-1} + \frac{\alpha_t D_t - H_{t-1}}{t+1} \\ &= \theta_{t-1}^M + \frac{\alpha_t D_t - H_{t-1}}{t+1}. \end{aligned}$$

Thus, $M_t$ is the displacement from the common initialization, whereas $M_t - H_{t-1}$ is the current parameter increment. We next derive the projection coefficient and the equal-weight averaging limit.

**Proposition A.2** (Closed-form projection)**.** *For fixed $H_{t-1}$ and $\widehat{M}_t$, if $D_t \neq 0$, the coefficient $\alpha_t$ in Eq. 8 is the unique minimizer of*

$$\min_{\alpha \geq 0} \left\| \frac{tH_{t-1} + \alpha D_t}{t+1} - \widehat{M}_t \right\|_F^2.$$

*Proof.* With $K_t = (t+1)\widehat{M}_t - tH_{t-1}$, expanding the objective gives

$$\begin{aligned} (t+1)^2 \|M_t(\alpha) - \widehat{M}_t\|_F^2 &= \|\alpha D_t - K_t\|_F^2 \\ &= \alpha^2 \|D_t\|_F^2 - 2\alpha \langle D_t, K_t \rangle + \|K_t\|_F^2. \end{aligned}$$

This quadratic has second derivative $2\|D_t\|_F^2 > 0$ and unconstrained minimizer $\langle D_t, K_t \rangle / \|D_t\|_F^2$. Restricting the minimizer to $[0, \infty)$ gives

$$\alpha_t = \max\left(0, \frac{\langle D_t, K_t \rangle}{\|D_t\|_F^2}\right).$$

□

If $D_t = 0$, every feasible coefficient gives the same update, and we set $\alpha_t = 0$.

**Corollary (recursive averaging).** Starting from $H_0 = M_0 = 0$ and using the identity $H_t = M_t$, the selected merged task vector satisfies

$$M_t = \frac{1}{t+1} \sum_{i=1}^{t} \alpha_i D_i. \tag{20}$$

If $\beta = 0$ throughout the stream, it reduces to

$$M_t = C_t = \frac{R_t}{t+1}, \qquad \theta_t^M = \frac{\theta_0 + \sum_{i=1}^{t} \theta_i}{t+1}. \tag{21}$$

*Proof.* Multiplying the recurrence by $t+1$ and using $H_{t-1} = M_{t-1}$ in the induction yields

$$(t+1)M_t = tH_{t-1} + \alpha_t D_t = \sum_{i=1}^{t} \alpha_i D_i.$$

For $\beta = 0$, we have $s_t = 1$. The initialization satisfies $H_0 = M_0 = R_0 = 0$. Assuming $H_{t-1} = M_{t-1} = R_{t-1}/t$ gives

$$K_t = R_t - tH_{t-1} = R_t - R_{t-1} = D_t.$$

Hence $\alpha_t D_t = D_t$, including the zero-vector case, and the update gives $M_t = R_t/(t+1)$. Substituting $D_i = \theta_i - \theta_0$ proves the parameter-averaging identity. □

### A.4 Moment and Affine-Map Consistency of AM-BN

We consider a fixed BN layer and channel, omitting their indices. For $i = 0, \ldots, t$, let $P_i = \mathcal{N}(\mu_i, v_i)$ denote the stored feature-distribution proxy of the pre-trained model ($i = 0$) or a subject model ($i \geq 1$), with $\mu_i \in \mathbb{R}$ and $0 < v_i < \infty$.

**Moment consistency.** Let $J$ be uniform on $\{0, \ldots, t\}$ and $X \mid J = i \sim P_i$. By total expectation and total variance,

$$\begin{aligned} \mathbb{E}[X] &= \frac{1}{t+1} \sum_{i=0}^{t} \mu_i = \bar{\mu}_t, \\ \mathrm{Var}(X) &= \frac{1}{t+1} \sum_{i=0}^{t} \left[ v_i + (\mu_i - \bar{\mu}_t)^2 \right] \\ &= \frac{1}{t+1} \sum_{i=0}^{t} (v_i + \mu_i^2) - \bar{\mu}_t^2 = \bar{v}_t. \end{aligned} \tag{22}$$

Thus Eq. 12 gives the mixture moments and ensures $\bar{v}_t \geq 0$. The variance accounts for both within-component variation and differences between component means.

**Exact affine-map reconstruction.** Kobler et al. (2022) express BN with fixed inference statistics as a channel-wise affine transformation.

**Proposition A.3** (Exact affine-map reconstruction)**.** *When evaluated with running moments $(\bar{\mu}_t, \bar{v}_t)$ and the same constant $\epsilon > 0$ used in reconstruction, the BN parameters in Eq. 14 realize $a_t^M x + o_t^M$ for every $x \in \mathbb{R}$.*

*Proof.* Since $\bar{v}_t + \epsilon > 0$, substituting the reconstructed parameters gives

$$\begin{aligned} \gamma_t^M \frac{x - \bar{\mu}_t}{\sqrt{\bar{v}_t + \epsilon}} + \zeta_t^M &= a_t^M (x - \bar{\mu}_t) + o_t^M + a_t^M \bar{\mu}_t \\ &= a_t^M x + o_t^M. \end{aligned} \tag{23}$$

□

**Compatibility-weighted affine-map merging.** Expanding the symmetric Gaussian KL divergence in Eq. 11 yields

$$\kappa_i = \frac{(v_i - v_0)^2 + (\mu_i - \mu_0)^2(v_0 + v_i)}{4v_0 v_i} \geq 0, \tag{24}$$

so $0 < \omega_i = (1 + \kappa_i)^{-1} \leq 1$. Writing $u_i = (a_i, o_i)$, Eq. 13 becomes

$$u_t^M = \left(1 - \frac{\sum_{i=1}^t \omega_i}{t+1}\right) u_0 + \sum_{i=1}^t \frac{\omega_i}{t+1} u_i. \tag{25}$$

The coefficients are nonnegative and sum to one, with the pre-trained coefficient at least $1/(t + 1)$. This gives a convex combination of channel-wise affine maps, while the factors $\omega_i$ attenuate subject-specific changes in affine-map coefficients according to their compatibility with the pre-trained model.

### A.5 Algorithmic Details

Algorithm 1 summarizes the sequential merging procedure. Each transported linear layer treats its weight matrix and optional bias jointly as one affine operator, sharing a single incoming coefficient.

**Algorithm 1** RPT for subject-incremental EEG model merging

**Require:** Pre-trained model $\theta_0$, stream of subject models $\{\theta_t\}_{t=1}^T$, fixed transport exponent $\beta$
1: $\theta_0^M \leftarrow \theta_0$; initialize $(R_0^g, E_0^g) \leftarrow (0, 0)$ for each transported layer $g$
2: Initialize BN first- and second-moment sums from $\theta_0$, and sums of BN coefficient changes to zero
3: **for** $t = 1, \ldots, T$ **do**
4: Receive $\theta_t$; initialize $\theta_t^M \leftarrow \theta_{t-1}^M$
5: **for all** transported linear layers $g$ **do**
6: $D_t^g \leftarrow \theta_t^g - \theta_0^g$, $H_{t-1}^g \leftarrow \theta_{t-1}^{M,g} - \theta_0^g$
7: Update $R_t^g$ and $E_t^g$ using Eq. 3
8: Compute $r_t^g$ and $q_t^g$ using Eq. 4, setting $r_t^g = 1$ if $E_t^g = 0$
9: Construct $\widehat{M}_t^g$ and $K_t^g$ using Eqs. 5 and 7
10: Compute $\alpha_t^g$ using Eq. 8, setting $\alpha_t^g = 0$ if $D_t^g = 0$
11: $M_t^g \leftarrow M_t^g(\alpha_t^g)$ using Eq. 2
12: $\theta_t^{M,g} \leftarrow \theta_0^g + M_t^g$
13: **end for**
14: For each remaining non-BN layer $g$, set $\theta_t^{M,g} \leftarrow \theta_{t-1}^{M,g} + (\theta_t^g - \theta_{t-1}^{M,g})/(t + 1)$
15: Compute the incoming BN affine maps and their compatibility with $\theta_0$ using Eqs. 10 and 11
16: Accumulate BN moment sums and compatibility-weighted coefficient changes; compute mixture moments and merged maps using Eqs. 12 and 13
17: Assign the mixture moments as BN running moments and reconstruct the affine parameters using Eq. 14
18: Deploy $\theta_t^M$ and release $\theta_t$
19: **end for**
20: **return** $\theta_T^M$

**State and computation.** Let $P$ denote the number of model parameters, $P_{\text{aff}}$ the number of parameters in transported linear layers, and $C_{\text{BN}}$ the total number of BN channels. Each transported layer maintains one resultant and one scalar energy. Each BN channel maintains sums of means, second moments $v_i + \mu_i^2$, and compatibility-weighted changes in affine-map coefficients $\omega_i(u_i - u_0)$. These persistent summaries require $O(P_{\text{aff}} + C_{\text{BN}})$ storage. Each subject arrival requires $O(P)$ arithmetic, and total working memory, including the pre-trained, merged, and incoming models, is $O(P)$ and independent of stream length.

# B EXPERIMENTAL DETAILS

## B.1 DATASETS AND PREPROCESSING

**Physionet-MI.** Physionet-MI comprises the motor imagery recordings from the PhysioNet EEG Motor Movement/Imagery Dataset (Schalk, 2009). The dataset contains EEG recordings collected during visually cued motor execution and imagery tasks, using 64 scalp channels sampled at 160 Hz. We use the motor imagery recordings for four-class classification: imagined movements of the left fist, right fist, both fists, and both feet. The resulting evaluation pool contains 103 subjects.

**FACED.** Chen et al. (2023) introduce the Finer-grained Affective Computing EEG Dataset (FACED), which contains 32-channel EEG recordings from 123 subjects watching emotion-eliciting video clips. Its nine emotion categories comprise amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and a neutral state. This coverage supports fine-grained recognition across positive, negative, and neutral emotional experiences. We retain all 123 subjects and use nine-class emotion recognition as the decoding task.

**ISRUC.** Khalighi et al. (2016) introduce ISRUC-Sleep, a polysomnographic dataset with overnight physiological recordings and expert sleep-stage annotations. We use subgroup 1, which contains recordings from individuals with sleep-related disorders, and retain 98 subjects. Inputs comprise six EEG channels and two electrooculography (EOG) channels. Each 30-second epoch is assigned one of five sleep-stage labels: Wake, N1, N2, N3, or REM. We formulate the task as sequence labeling to incorporate temporal context across consecutive sleep epochs.

**Preprocessing.** Our data preparation follows the preprocessing procedures described by Zhou et al. (2025a); further procedural details can be found in their work. Briefly, we exclude Physionet-MI subjects 38, 88, 89, 92, 100, and 104 because of inconsistent sampling rates or task durations, and ISRUC subjects 8 and 40 because of missing channels. For Physionet-MI, we extract one 4-second epoch (640 points per channel at 160 Hz) from each imagery-cue onset in runs 4, 6, 8, 10, 12, and 14. FACED uses the released preprocessed data, with one 30-second input (7,500 points per channel at 250 Hz) from the end of each video trial. ISRUC recordings are band-pass filtered at 0.3–35 Hz, resampled to 100 Hz, and segmented into 30-second epochs containing 3,000 time points per channel. Consecutive ISRUC epochs are grouped into non-overlapping sequences of length 20 with a stride of 20 epochs, retaining one label per epoch.

## B.2 SUBJECT-INCREMENTAL PROTOCOL AND METRICS

**Evaluation protocol.** We study continual EEG adaptation across successive subjects. Under the subject-incremental learning formulation, we treat each subject as an incremental domain. For each dataset, subjects are randomly partitioned into disjoint pre-training and incremental pools at an approximate 3:7 ratio. The respective pool sizes are 32/71 for Physionet-MI, 38/85 for FACED, and 31/67 for ISRUC. The pre-training pool establishes a common initialization, while the incremental pool provides a stream of subject-specific adaptations.

Each incremental subject's complete preprocessed inputs are randomly partitioned into approximately 90% training and 10% test subsets. Each trial or ISRUC sequence is assigned wholly to one subset, and the split remains fixed across methods and arrival orders. Subject models are independently fine-tuned from the common pre-trained model using the training subset. The merger receives these models sequentially and updates a shared model using their parameters and stored BN statistics. After each update, the merged model is evaluated on the fixed test sets of all encountered subjects, tracking both incoming-subject adaptation and historical retention. Main CMM comparisons use the same subject models and ten shared random arrival orders, enabling paired comparisons under changes in subject order. The split-robustness study additionally evaluates five pre-training/incremental subject partitions, with ten arrival orders per partition.

**Metrics.** Let $T > 1$ denote the number of incremental subjects, indexed by arrival order. For $m \in \{\mathrm{ACC}, \mathrm{MF1}\}$, let $m_{t,i}$ denote the performance of the merged model on subject $i$'s fixed test set after update $t$, where $1 \leq i \leq t \leq T$. Here, ACC denotes classification accuracy and MF1

denotes macro-F1. We compute

$$A_m = \frac{1}{T}\sum_{i=1}^{T} m_{T,i}, \qquad P_m = \frac{1}{T}\sum_{i=1}^{T} m_{i,i}, \tag{26}$$

$$S_m = \frac{1}{T-1}\sum_{i=1}^{T-1} m_{T,i}, \qquad \mathrm{BWT}_m = \frac{1}{T-1}\sum_{i=1}^{T-1}\big(m_{T,i} - m_{i,i}\big). \tag{27}$$

The final average $A_m$ gives the ACC or MF1 reported in the main comparisons. The plasticity score $P_m$ averages performance immediately after each subject's arrival, while the stability score $S_m$ measures final performance on earlier subjects. Backward transfer measures their average performance change since arrival: positive values indicate improvement and negative values indicate forgetting. These complementary metrics characterize overall decoding quality, immediate adaptation, and retention, with equal weight assigned to each subject.

For the reference models, Pre-trained evaluates the common initial model on each subject, while Independent FT evaluates each subject-specific model on its corresponding test set. Their averages are invariant to arrival order and are therefore reported as single values.

**Statistical reporting.** Unless otherwise stated, results are reported as the mean and sample standard deviation across ten random subject orders. Methods share the same orders within each dataset, preserving paired comparisons. For direct-retention analyses, subject scores are first aggregated within each order, and 95% confidence intervals are obtained by bootstrap resampling the ten order-level summaries. Inter-method differences use paired order-level differences. For split robustness, results are first averaged across the ten orders within each subject partition, then summarized by the mean and sample standard deviation across the five partitions. This evaluates sensitivity to both subject order and the composition of the pre-training and incremental pools.

In aggregate visualizations, ACC and MF1 receive equal weight, and cross-dataset summaries additionally assign equal weight to the three datasets. Dataset-specific ACC, MF1, and BWT results accompany these summaries. Mechanism controls vary the component under study while holding the subject models, data splits, arrival orders, and remaining settings fixed.

### B.3 Backbone and Subject Model Training

We use a common CNN–Transformer backbone design across all three EEG tasks, following the feature-extractor, temporal-encoder, and classifier organization of Zhou et al. (2025a). The CNN feature extractor produces 512-dimensional tokens, three Transformer blocks encode their temporal interactions, and an MLP predicts the classes (Figure 6). All datasets share the convolutional channel progression, Transformer configuration, and MLP hidden-layer widths. Within this design, front-end branch and kernel settings accommodate the recording channels and temporal structure, while the readout matches input-level or epoch-level prediction. Table 5 gives the common configuration and these task settings. Within each dataset, all compared methods use the same instantiated backbone. Appendix C.2 evaluates its GN counterpart under the same protocol.

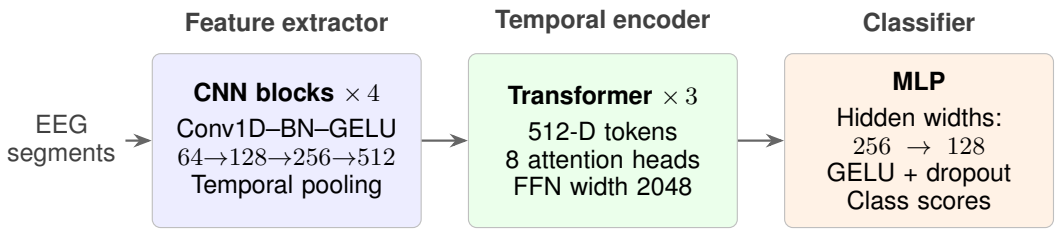



Figure 6: Common CNN–Transformer backbone design for the three EEG tasks. A CNN feature extractor maps temporal segments to tokens, a Transformer encoder models their interactions, and an MLP produces class scores. Branch settings, temporal grouping, and task readouts are detailed in Table 5.

**Temporal feature extraction.** Each convolutional branch contains four Conv1D–BN–GELU stages with output channels $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$. Max pooling follows the first and fourth stages, and adaptive average pooling produces one feature vector per temporal segment. For Physionet-MI, each 640-point input is partitioned into ten consecutive 64-point segments and processed by a single branch. For FACED, each 7,500-point input is partitioned into ten 750-point segments; parallel branches with shorter and longer temporal kernels extract complementary temporal features. For ISRUC, each input sequence contains twenty 3,000-point sleep epochs, with EEG and EOG processed by separate branches. In the two-branch front ends, the pooled features are concatenated and linearly projected from 1,024 to 512 dimensions. These front ends produce ten tokens for Physionet-MI and FACED and twenty tokens for ISRUC.

**Temporal encoding and prediction.** The encoder uses three Transformer blocks, each with model width 512, eight attention heads, and feed-forward width 2,048. The classifier follows the same two-hidden-layer design across datasets, with widths 256 and 128, GELU activations, dropout of 0.1, and a final linear layer without bias. For Physionet-MI and FACED, the ten encoded tokens are flattened into a 5,120-dimensional vector for input-level classification. For ISRUC, the classifier is applied to each 512-dimensional encoded token to produce one sleep-stage prediction per epoch. The output dimensions are four, nine, and five, respectively.

Table 5: Common backbone configuration and dataset-specific input/output settings. Convolution and max-pooling entries use $(k, s, p)$ for kernel size, stride, and padding. Paired entries for FACED correspond to its short- and long-kernel branches; ISRUC settings apply to both EEG and EOG branches.

| **Common configuration** | | | |
|---|---|---|---|
| Convolutional stages | Four Conv1D–BN–GELU stages | | |
| Convolutional channels | $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$ | | |
| Temporal pooling | Max pooling after stages 1 and 4; adaptive average pooling to one position | | |
| Transformer | 3 blocks; width 512; 8 attention heads; FFN width 2048 | | |
| MLP hidden layers | $256 \rightarrow 128$ with GELU; final linear layer without bias | | |
| Dropout | 0.1; attention-probability dropout 0.25 | | |
| **Input/output configuration** | **Physionet-MI** | **FACED** | **ISRUC** |
| Input channels | 64 EEG | 32 EEG | 6 EEG + 2 EOG |
| Temporal grouping | $10 \times 64$ points | $10 \times 750$ points | $20 \times 3000$ points |
| Convolutional branches | Single | Short / long kernels | EEG / EOG |
| Conv 1 $(k, s, p)$ | $(50, 6, 24)$ | $(50, 6, 24)$ / $(400, 50, 200)$ | $(50, 6, 0)$ |
| Conv 2–4 $(k, s, p)$ | $(6, 1, 3)$ | $(6, 1, 3)$ / $(8, 1, 3)$ | $(8, 1, 0)$ |
| Pool after Conv 1 | $(8, 8, 4)$ | $(8, 8, 4)$ / $(4, 4, 2)$ | $(8, 8, 0)$ |
| Pool after Conv 4 | $(4, 4, 0)$ | $(4, 4, 0)$ / $(2, 2, 1)$ | $(4, 4, 0)$ |
| Branch projection | – | $1024 \rightarrow 512$ | $1024 \rightarrow 512$ |
| Encoder input | $10 \times 512$ | $10 \times 512$ | $20 \times 512$ |
| Classifier input | 5120, flattened tokens | 5120, flattened tokens | 512, per token |
| Output classes | 4 | 9 | 5 |
| Prediction unit | Input segment | Input segment | Sleep epoch |

**Subject model training.** For each dataset, we train the initial model $\theta_0$ on the pre-training subject pool for 100 epochs using Adam with a learning rate of $10^{-4}$, optimizer coefficients $(\beta_1, \beta_2) = (0.9, 0.99)$, and weight decay $3 \times 10^{-4}$. Each subject model is then independently fine-tuned from $\theta_0$ on its designated training subset for 50 epochs. Fine-tuning uses Adam with the same optimizer coefficients and weight decay, a learning rate of $10^{-5}$, and a batch size of 32. The architecture and label space remain fixed throughout the subject stream. All CMM methods receive the same pre-trained model and subject checkpoints, providing a common basis for comparing their merging rules. Subject partitions are detailed in Appendix B.2.

### B.4 BASELINES

**CMM baselines.** We evaluate weight averaging (WA) (Izmailov et al., 2018), Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), MagMax (Marczak et al., 2025), OPCM (Tang et al., 2025), NUFILT (Qiu et al., 2026), WUDI-Merging (Cheng et al., 2025), and DOP (Yang et al., 2025). These baselines cover parameter averaging, task-vector addition, sign and magnitude selection, and subspace-based merging. We also include AdaMerging (Yang et al., 2024) and MINGLE (Qiu et al., 2025), which use unlabeled samples for adaptation and are marked with † in Table 2. All merging methods receive the same pre-trained model and subject checkpoints in the same evaluation orders.

Table 6 summarizes the key hyperparameter settings of the CMM baselines.

Table 6: Key hyperparameters of CMM baselines. Learning rates and steps refer to merging optimization or test-time adaptation.

| Method | Settings |
|---|---|
| Task Arithmetic | Task-vector scale $\lambda = 0.3$. |
| TIES | Scale $\lambda = 0.5$; top-20% magnitude retention; sign-consistent summation. |
| MagMax | Task-vector scale $\lambda = 0.5$. |
| OPCM | SVD projection threshold $\alpha = 0.5$. |
| NUFILT | Null-space rank $r_p = 128$, correction rank $r_l = 64$, task-projection rank $r_v = 8$; learning rate $10^{-3}$; 50 steps. |
| WUDI-Merging | Learning rate $10^{-3}$; 50 steps. |
| DOP | SVD threshold $0.9$; two-sided projection; MGDA with initial weight $0.5$ and EMA decay $0.999$; learning rate $10^{-4}$; 50 steps. |
| AdaMerging | Initial coefficient 1, constrained to $[0, 1]$; learning rate $10^{-4}$; 50 steps. |
| MINGLE | LoRA rank $64$; projection rank 2; gate parameters $\gamma = 1$, $\beta = 0.99$; distillation temperature 2; learning rate $10^{-4}$; 50 steps. |

The upper CMM block in Table 2 evaluates each baseline with its own normalization handling. The lower block embeds AM-BN into the eight data-free baselines while preserving their non-BN merging rules and incoming checkpoints. This controlled comparison evaluates parameter merging under a common normalization procedure. Pre-trained, Independent FT, and Sequential FT provide reference results for the initial model, independent subject adaptation, and sequential fine-tuning, respectively.

The uniform averaging control in Table 3 includes the pre-trained model as an equally weighted member, whereas WA averages only subject models. AdaMerging and MINGLE use unlabeled inputs to adapt merging coefficients or expert routing. In subject-incremental EEG, limited adaptation samples, temporal dependence, and uneven class coverage can make the adaptation signal less representative of the distributions encountered across subjects, causing unreliable predictions.

**Conventional continual learning.** We compare with elastic weight consolidation (EWC) (Kirkpatrick et al., 2017), experience replay (ER) (Rolnick et al., 2019), dark experience replay with label replay (DER++) (Buzzega et al., 2020), and the EEG-oriented methods MUDVI (Duan et al., 2024a), ReSNT (Duan et al., 2023), and AMBM (Duan et al., 2024b). We additionally include BrainUICL-SR, a supervised replay adaptation of Zhou et al. (2025a) that combines classification losses on current and replayed labeled samples, and ProNECL (Li et al., 2026a), which maintains class prototypes for cross-subject feature regularization.

All methods use the same backbone, subject partitions, evaluation orders, and local optimization settings described in Appendix B.3, while retaining their regularization, replay, or prototype mechanisms. During supervised adaptation, BN running means and variances are updated from training-mode forward passes. Methods without replay use current-subject samples, whereas replay-based methods additionally expose BN to buffered historical samples according to their minibatch construction.

Table 7 lists the default method-specific settings for conventional CL. Shared training settings follow Appendix B.3.

Table 7: Default hyperparameters of conventional CL baselines. Replay methods use a replay-to-current batch ratio of 1. Memory budgets are 200 inputs for Physionet-MI and FACED, and 500 sleep epochs (25 sequences) for ISRUC.

| Method | Settings |
|---|---|
| EWC | Regularization weight $\lambda = 10^3$; Fisher decay 1. |
| ER | Shared replay budget and batch ratio. |
| DER++ | Logit-replay weight $\alpha = 0.1$; label-replay weight $\beta = 0.5$. |
| MUDVI | Memory-selection coefficient $\lambda = 1$; 6 representatives per class. |
| ReSNT | Transformation intensity $G = 0.1$; JS-divergence weight 2; 1 transformation step. |
| AMBM | 6 retained heads; rate-adaptation step size $10^{-3}$; mutual-information weight 1; contrastive temperature 0.1. |
| BrainUICL-SR | Current-sample loss weight $\alpha = 0.01$; replay loss weight $1 - \alpha$. |
| ProNECL | Prototype-consistency weight $\lambda_p = 0.1$; alignment weight $\lambda_a = 0.3$; prototype EMA decay 0.9. |

**Seq-RPT.** Seq-RPT applies the recursive CMM rule to checkpoints obtained through sequential adaptation. At step $t$, we fine-tune the previously merged model $\theta^M_{t-1}$ on the arriving subject to obtain $\widetilde{\theta}_t$. For each non-BN layer $g$, we express the local adaptation relative to the common initialization by constructing the incoming checkpoint

$$\theta^g_t = \theta^g_0 + \big(\widetilde{\theta}^g_t - \theta^{M,g}_{t-1}\big). \tag{28}$$

Its task vector is therefore $D^g_t = \theta^g_t - \theta^g_0 = \widetilde{\theta}^g_t - \theta^{M,g}_{t-1}$, while the historical task vector remains $H^g_{t-1} = \theta^{M,g}_{t-1} - \theta^g_0$. RPT uses these task vectors to construct the resultant-guided target and determine the incoming coefficient through history-weighted projection. The merged model is then updated recursively:

$$\theta^{M,g}_t = \theta^{M,g}_{t-1} + \frac{\alpha^g_t D^g_t - H^g_{t-1}}{t+1}. \tag{29}$$

For transported linear layers, $\alpha^g_t$ follows Eq. 8; remaining non-BN parameters use $\alpha^g_t = 1$. BN updates are translated to the common initialization in $(\mu, \log v, a, o)$ coordinates before applying AM-BN. The merging step operates on checkpoints and incremental summaries without accessing EEG recordings or replaying historical data.

**Computing environment.** All experiments were conducted on a server with two Intel Xeon Platinum 8368Q CPUs, 512 GB of system memory, and NVIDIA GeForce RTX 4090 GPUs under Ubuntu 22.04. Each training and evaluation run used one 24-GB RTX 4090. For conventional CL, end-to-end runtime includes local training, continual updating or merging, and evaluation. Peak GPU memory denotes the maximum CUDA memory allocated within the corresponding measurement scope.

### B.5 Transport Exponent

**Transport-exponent selection.** We select $\beta$ by three-fold subject-level cross-validation within the pre-training pool. In each round, two folds train an initialization, and the remaining fold forms a development stream. Each development checkpoint is independently fine-tuned from this initialization using 90% of the subject's samples, with 10% reserved for validation; training settings follow Appendix B.3. We search $\beta \in \{-2.5, -2.0, \ldots, 2.5\}$ at intervals of 0.5. The selection score equally weights final validation ACC and MF1 and averages over all folds, yielding $\beta = -2$, 2, and 0.5 for Physionet-MI, FACED, and ISRUC, respectively. Main incremental subjects are excluded from selection. The main experiments construct the initialization from the full pre-training pool and keep the selected exponent fixed across transported layers, subject arrivals, and evaluation orders.

**Sensitivity and shared exponent.** Figure 5 evaluates all eleven grid points over ten random subject orders. For each shared value $\beta \in \{0, 0.5, 1, 1.5, 2\}$, RPT achieves higher mean ACC and MF1 on all three datasets than every baseline in the original CMM comparison block of Table 2. At the shared value $\beta = 0.5$, it also exceeds the baselines using AM-BN on all six ACC/MF1 endpoints, as

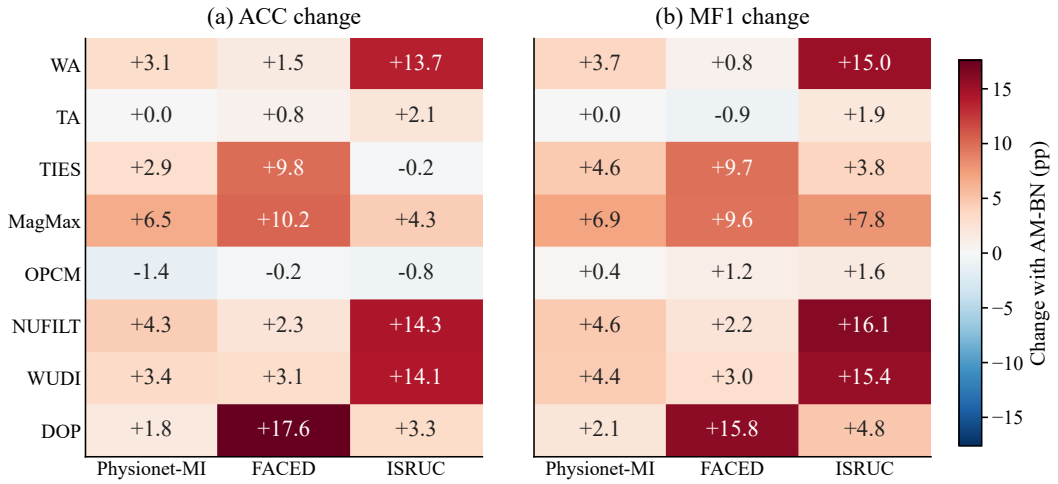



Figure 7: Change in ACC and MF1 after replacing each method's native normalization handling with AM-BN. Columns are datasets and rows are the eight compatible CMM methods; positive values favor AM-BN. Each cell is the difference in pp between the corresponding ten-order means in the two blocks of Table 2.

summarized in Table 8. Thus, a single exponent supports competitive merging across the three EEG tasks, while dataset-level selection further improves performance on Physionet-MI and FACED.

Table 8: Shared-exponent comparison (%). RPT uses $\beta = 0.5$ for all three datasets. Values are means over ten random subject orders. Baseline entries are the highest means among the eight baselines using AM-BN in Table 2, determined separately for each dataset and metric.

| | RPT ($\beta = 0.5$) | | Best baseline + AM-BN | |
|---|---|---|---|---|
| Dataset | ACC | MF1 | ACC | MF1 |
| Physionet-MI | 52.68 | 46.08 | 52.36 | 45.92 |
| FACED | 54.43 | 44.16 | 54.04 | 44.14 |
| ISRUC | 72.34 | 59.51 | 71.85 | 58.87 |

# C Additional Experimental Results

This section expands the main experimental evidence through normalization controls, including a GN backbone without BN, followed by stability–plasticity, component mechanisms, resource trade-offs, and robustness.

## C.1 AM-BN Transfer Across Merging Methods

Figure 7 expands the controlled comparison in Table 2. AM-BN improves ACC in 19 of 24 method–dataset cases and MF1 in 22 of 24. The largest gains occur for DOP on FACED ($+17.6/+15.8$ ACC/MF1 points) and for NUFILT on ISRUC ($+14.3/+16.1$). At the same time, OPCM loses ACC on all three datasets, despite small MF1 gains. Merging BN inference maps improves performance for most parameter-merging rules, with gains that depend on the remaining parameter update. RPT remains above all eight transferred baselines on every ACC/MF1 endpoint in the lower block of Table 2.

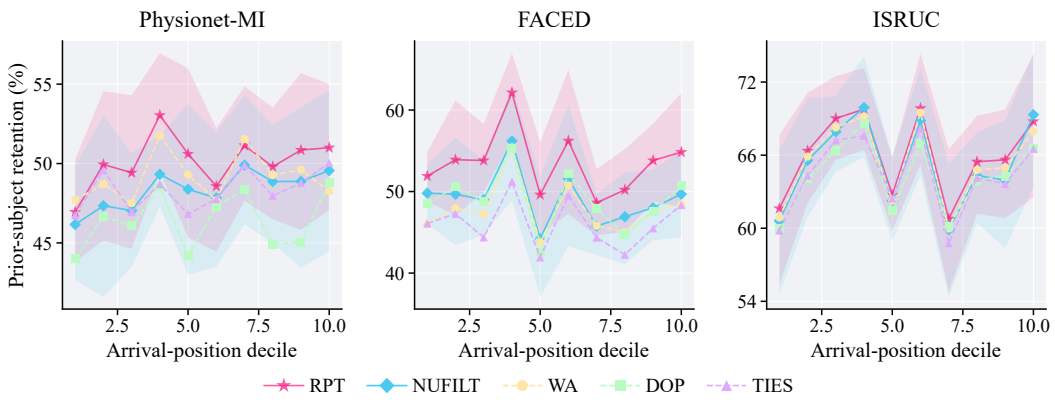



Figure 8: Final performance on previously encountered subjects, grouped by their arrival-position decile. Panels show Physionet-MI, FACED, and ISRUC from left to right. Curves average ACC and MF1 within each of the ten random subject orders; bands are 95% confidence intervals across orders. The horizontal coordinate denotes a normalized position in the complete subject stream, not a fixed set of ten subjects.

### C.2 Task-Vector Transport without Batch Normalization

RPT retains the highest mean ACC and MF1 among the compared CMM methods on all three datasets when every backbone BN layer is replaced by GroupNorm (GN; Table 9). This control evaluates its parameter-transport advantage on a network without BN.

We repeat pre-training and subject-specific fine-tuning with the GN backbone under the main experimental settings. Subject partitions, training/test splits, optimization settings, ten arrival orders, and method hyperparameters remain unchanged. Within each dataset, all methods receive the same GN pre-trained model and subject checkpoints. RPT retains adaptive transport on fully connected layers, while convolutional parameters and GN and LayerNorm affine parameters follow uniform averaging. With no BN layers, AM-BN has no state to merge. AdaMerging and MINGLE retain their native unlabeled-data adaptation protocol.

Relative to WA, RPT improves ACC/MF1 by $1.41/1.82$, $1.69/1.24$, and $0.76/0.64$ points on Physionet-MI, FACED, and ISRUC, respectively. Its mean BWT is positive on all six dataset–metric endpoints and highest on five. The exception is ISRUC BWT-MF1, where AdaMerging reaches $1.21$ versus RPT's $0.58$ points, alongside lower ACC/MF1 ($20.37/11.34\%$ versus $68.29/51.51\%$). These results support the benefit of task-vector transport without AM-BN.

### C.3 Dataset-Specific Stability–Plasticity

The macro curve in Figure 3(a) summarizes the three tasks equally, whereas Figure 8 exposes their different scales and trajectories. RPT achieves the highest mean retention in eight of ten arrival-position deciles on Physionet-MI and all ten on FACED. On ISRUC, NUFILT has slightly higher point estimates at deciles 4 and 10, with overlapping confidence intervals. After aggregation over prior subjects, RPT achieves the highest dataset-level retention across all three tasks.

Figure 9 gives the corresponding two-dimensional view. RPT's prior-subject retention exceeds NUFILT by $1.79$, $4.40$, and $0.79$ points on Physionet-MI, FACED, and ISRUC. Its aggregate new-subject score is $55.79\%$, compared with $55.72\%$ for DOP. These results show improved prior-subject retention with competitive incoming-subject performance.

Table 9: Continual model merging with GroupNorm replacing every backbone BN layer (ACC/MF1 in %; BWT in pp; mean$_{\pm SD}$ over ten subject orders). All metrics are higher-is-better. The first three rows in each panel are reference models. Bold marks the best CMM mean; † denotes native unlabeled-data adaptation. RPT uses task-vector transport without AM-BN.

| Method | ACC | MF1 | BWT$_A$ | BWT$_F$ |
|---|---|---|---|---|
| **Physionet-MI** | | | | |
| Pre-trained | 47.26$_{\pm 0.00}$ | 41.02$_{\pm 0.00}$ | – | – |
| Independent FT | 54.76$_{\pm 0.00}$ | 50.07$_{\pm 0.00}$ | – | – |
| Sequential FT | 49.80$_{\pm 2.70}$ | 42.24$_{\pm 2.85}$ | -7.46$_{\pm 3.52}$ | -9.53$_{\pm 3.69}$ |
| WA | 48.60$_{\pm 0.00}$ | 41.49$_{\pm 0.00}$ | -0.11$_{\pm 0.34}$ | -0.36$_{\pm 0.34}$ |
| Task Arithmetic | 30.76$_{\pm 0.00}$ | 19.47$_{\pm 0.00}$ | -4.53$_{\pm 1.45}$ | -4.83$_{\pm 1.50}$ |
| TIES | 49.10$_{\pm 1.08}$ | 42.66$_{\pm 1.13}$ | -1.66$_{\pm 1.29}$ | -1.91$_{\pm 1.30}$ |
| MagMax | 46.61$_{\pm 0.00}$ | 39.25$_{\pm 0.00}$ | -0.32$_{\pm 0.78}$ | -0.71$_{\pm 0.96}$ |
| OPCM | 45.30$_{\pm 0.31}$ | 38.85$_{\pm 0.35}$ | -1.05$_{\pm 0.72}$ | -0.81$_{\pm 0.97}$ |
| NUFILT | 49.27$_{\pm 1.71}$ | 42.60$_{\pm 2.04}$ | -2.51$_{\pm 1.71}$ | -2.78$_{\pm 2.13}$ |
| DOP | 48.41$_{\pm 1.75}$ | 41.71$_{\pm 2.28}$ | -4.24$_{\pm 1.99}$ | -4.97$_{\pm 2.48}$ |
| AdaMerging† | 23.57$_{\pm 0.00}$ | 9.76$_{\pm 0.00}$ | -5.00$_{\pm 2.72}$ | -6.65$_{\pm 2.38}$ |
| MINGLE† | 49.37$_{\pm 2.37}$ | 43.16$_{\pm 2.49}$ | 0.35$_{\pm 1.18}$ | 0.39$_{\pm 1.40}$ |
| WUDI-Merging | 42.96$_{\pm 1.50}$ | 35.49$_{\pm 2.06}$ | -7.82$_{\pm 1.18}$ | -9.43$_{\pm 1.45}$ |
| **RPT** | **50.01**$_{\pm 0.08}$ | **43.31**$_{\pm 0.06}$ | **0.56**$_{\pm 0.34}$ | **0.58**$_{\pm 0.31}$ |
| **FACED** | | | | |
| Pre-trained | 44.88$_{\pm 0.00}$ | 35.07$_{\pm 0.00}$ | – | – |
| Independent FT | 48.24$_{\pm 0.00}$ | 38.35$_{\pm 0.00}$ | – | – |
| Sequential FT | 47.06$_{\pm 1.79}$ | 37.82$_{\pm 2.12}$ | -3.89$_{\pm 2.20}$ | -4.21$_{\pm 2.66}$ |
| WA | 49.41$_{\pm 0.00}$ | 39.34$_{\pm 0.00}$ | 0.04$_{\pm 0.76}$ | 0.09$_{\pm 0.70}$ |
| Task Arithmetic | 22.35$_{\pm 0.00}$ | 14.71$_{\pm 0.00}$ | -11.43$_{\pm 1.73}$ | -10.03$_{\pm 1.71}$ |
| TIES | 46.39$_{\pm 0.96}$ | 36.59$_{\pm 0.85}$ | -0.87$_{\pm 0.81}$ | -0.43$_{\pm 0.92}$ |
| MagMax | 49.80$_{\pm 0.00}$ | 39.24$_{\pm 0.00}$ | 0.16$_{\pm 0.84}$ | 0.03$_{\pm 1.07}$ |
| OPCM | 50.00$_{\pm 0.21}$ | 40.04$_{\pm 0.19}$ | -0.36$_{\pm 0.51}$ | 0.08$_{\pm 0.61}$ |
| NUFILT | 49.10$_{\pm 1.48}$ | 39.08$_{\pm 1.67}$ | 0.91$_{\pm 1.90}$ | 0.66$_{\pm 2.00}$ |
| DOP | 49.18$_{\pm 1.20}$ | 39.15$_{\pm 1.33}$ | 0.44$_{\pm 1.29}$ | -0.08$_{\pm 1.56}$ |
| AdaMerging† | 8.24$_{\pm 0.00}$ | 4.27$_{\pm 0.00}$ | -9.64$_{\pm 3.74}$ | -7.82$_{\pm 3.10}$ |
| MINGLE† | 45.80$_{\pm 1.38}$ | 36.02$_{\pm 1.45}$ | 0.12$_{\pm 0.82}$ | 0.16$_{\pm 0.63}$ |
| WUDI-Merging | 44.94$_{\pm 2.04}$ | 35.02$_{\pm 2.02}$ | -1.03$_{\pm 2.71}$ | -1.03$_{\pm 2.60}$ |
| **RPT** | **51.10**$_{\pm 0.19}$ | **40.58**$_{\pm 0.22}$ | **1.15**$_{\pm 0.69}$ | **1.18**$_{\pm 0.86}$ |
| **ISRUC** | | | | |
| Pre-trained | 66.85$_{\pm 0.00}$ | 47.99$_{\pm 0.00}$ | – | – |
| Independent FT | 73.41$_{\pm 0.00}$ | 56.48$_{\pm 0.00}$ | – | – |
| Sequential FT | 55.16$_{\pm 4.64}$ | 40.65$_{\pm 4.66}$ | -19.73$_{\pm 4.61}$ | -17.19$_{\pm 4.42}$ |
| WA | 67.53$_{\pm 0.00}$ | 50.87$_{\pm 0.00}$ | -0.60$_{\pm 0.31}$ | 0.02$_{\pm 0.16}$ |
| Task Arithmetic | 8.94$_{\pm 0.00}$ | 3.15$_{\pm 0.00}$ | -12.88$_{\pm 1.74}$ | -9.43$_{\pm 1.30}$ |
| TIES | 66.89$_{\pm 1.42}$ | 48.86$_{\pm 1.50}$ | -4.12$_{\pm 1.51}$ | -4.21$_{\pm 1.69}$ |
| MagMax | 62.53$_{\pm 0.00}$ | 46.25$_{\pm 0.00}$ | -1.86$_{\pm 1.78}$ | -1.60$_{\pm 1.48}$ |
| OPCM | 45.63$_{\pm 0.23}$ | 30.02$_{\pm 0.25}$ | -5.81$_{\pm 3.70}$ | -5.04$_{\pm 2.89}$ |
| NUFILT | 65.00$_{\pm 2.76}$ | 48.01$_{\pm 1.97}$ | -2.18$_{\pm 2.71}$ | -1.59$_{\pm 2.50}$ |
| DOP | 64.05$_{\pm 3.86}$ | 46.96$_{\pm 4.08}$ | -7.45$_{\pm 4.06}$ | -6.64$_{\pm 4.17}$ |
| AdaMerging† | 20.37$_{\pm 0.08}$ | 11.34$_{\pm 0.03}$ | 0.03$_{\pm 4.14}$ | **1.21**$_{\pm 2.27}$ |
| MINGLE† | 67.70$_{\pm 0.84}$ | 49.23$_{\pm 1.39}$ | -0.05$_{\pm 0.39}$ | 0.04$_{\pm 0.63}$ |
| WUDI-Merging | 61.30$_{\pm 3.67}$ | 44.23$_{\pm 2.98}$ | -7.57$_{\pm 3.35}$ | -6.25$_{\pm 2.96}$ |
| **RPT** | **68.29**$_{\pm 0.01}$ | **51.51**$_{\pm 0.02}$ | **0.06**$_{\pm 0.23}$ | 0.58$_{\pm 0.17}$ |

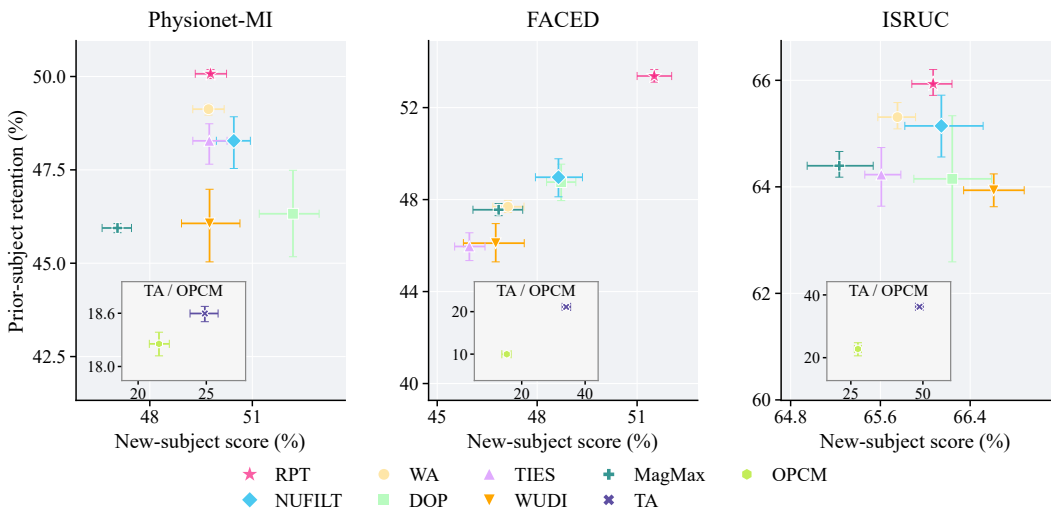



Figure 9: Dataset-specific stability–plasticity planes. Stability is performance on previously encountered subjects after continual merging; plasticity is the mean performance immediately after each subject's incorporation. Points and error bars summarize ten random subject orders with 95% confidence intervals. The main axes show seven methods; insets show TA and OPCM on separate scales, with the same score definitions. The preferred region is toward the upper right.

### C.4 Component Behavior and Transport Scope

The endpoint ablation in Table 3 is complemented by Figure 10. Independent averaging of BN parameters and moments produces larger performance losses over the stream, especially with task-vector transport. AM-BN improves the resulting stability-plasticity balance, and adding RPT transport improves it further over uniform parameter averaging. These trajectories localize when the factorial differences arise; the controlled endpoint comparisons in Figure 4 remain the basis for quantitative component attribution.

#### C.4.1 Paired Mechanism Controls

Layer-specific agreement and compatibility-weighted AM-BN each improve aggregate final performance across all ten subject orders (Table 10). Correct layer assignment improves on both common and permuted $q$, while compatibility weighting improves on both unweighted affine-map averaging and normalized overlap in macro score. These paired comparisons support matching agreement to its source layer and using stored-statistic compatibility to weight changes in BN affine-map coefficients.

All comparisons pair the same subject checkpoints and arrival order within each dataset. The $q$ controls retain the selected $\beta$ and AM-BN. Common $q$ assigns the arithmetic mean agreement of the linear layers at each step. Permuted $q$ preserves these scores while changing their layer assignment; its score averages three deterministic layer permutations. The BN controls retain RPT transport and vary only the normalization rule. Parameter averaging independently averages BN parameters and buffers, whereas affine-map averaging sets all compatibility weights to one. Normalized overlap rescales the subject weights to unit sum before assigning total subject mass $t/(t+1)$.

We compute each paired difference within an order before averaging over orders. The macro confidence intervals use 20,000 paired bootstrap resamples of the ten order-level differences, with random seed 20260901. All five macro contrasts remain positive in every evaluated order, supporting robustness of the component gains to subject arrival order.

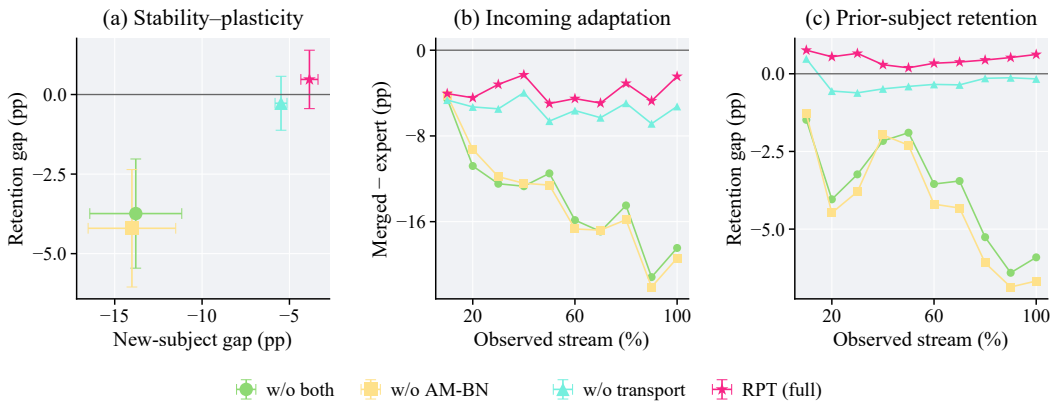



Figure 10: Stream-level behavior of the two-component factorial in Table 3. In panel (a), both axes average gaps over the stream: the horizontal gap compares new-subject performance with the independently fine-tuned subject model, and the vertical gap compares prior-subject performance at each step with performance at incorporation. (a) Macro stability–plasticity plane; (b) new-subject performance relative to the independently fine-tuned subject model over the stream; and (c) prior-subject retention relative to performance at incorporation. Performance gaps are in pp; error bars in (a) show SD over ten orders. Each curve aggregates the three datasets, ACC/MF1, and ten random subject orders.

Table 10: Paired gains over mechanism controls across ten subject orders, in pp. Final performance averages ACC and MF1 equally; the macro score also weights the three datasets equally. Positive differences favor RPT with AM-BN. Dataset columns report mean paired differences; the interval and positive-order count refer to the macro difference.

| Control | Physionet-MI | FACED | ISRUC | Macro | 95% CI | Positive orders |
|---|---|---|---|---|---|---|
| *Agreement controls with AM-BN fixed* | | | | | | |
| Common $q$ | +0.44 | +1.89 | +0.37 | **+0.90** | [0.85, 0.95] | 10/10 |
| Permuted $q$ | +0.75 | +5.40 | +0.17 | **+2.11** | [1.86, 2.37] | 10/10 |
| *BN controls with RPT transport fixed* | | | | | | |
| Parameter averaging | +4.29 | +47.51 | +0.90 | **+17.57** | [17.51, 17.62] | 10/10 |
| Affine-map averaging | +2.52 | +0.03 | +3.98 | **+2.18** | [2.08, 2.26] | 10/10 |
| Normalized overlap | +3.96 | +0.25 | -0.40 | **+1.27** | [1.19, 1.35] | 10/10 |

### C.4.2 Adaptive incoming coefficients

Figure 11 shows that RPT does not impose a fixed interpolation weight. From the first to the last stream bin, mean agreement decreases from $0.80$ to $0.66$ on Physionet-MI, from $0.74$ to $0.55$ on FACED, and from $0.68$ to $0.38$ on ISRUC. The resulting $\alpha$ reduces the incoming task-vector contribution on Physionet-MI but increases it on FACED and ISRUC. This dataset-dependent response is consistent with the role of $\beta$ in Eq. 5; causal support for layer assignment comes from the common-$q$ and permuted-$q$ controls in Figure 4(a), rather than from coefficient trajectories alone.

### C.4.3 Parameter scope

Table 11 tests whether the transport rule should simply be extended to every non-BN tensor. A single global treatment of all non-BN parameters reduces ACC/MF1 by $0.37/0.21$, $2.43/2.42$, and $0.03/0.07$ points on Physionet-MI, FACED, and ISRUC. Applying module-wise coefficients to all non-BN tensors is worse, most clearly on FACED ($-25.64/-24.69$ points). The selected scope also gives the highest BWT in five of the six dataset–metric comparisons. The exception is ISRUC BWT-ACC, where global all-non-BN gives $-0.05$ versus $-0.12$ for the selected scope, a $0.07$-point difference. These comparisons support applying adaptive transport to linear layers in the evaluated

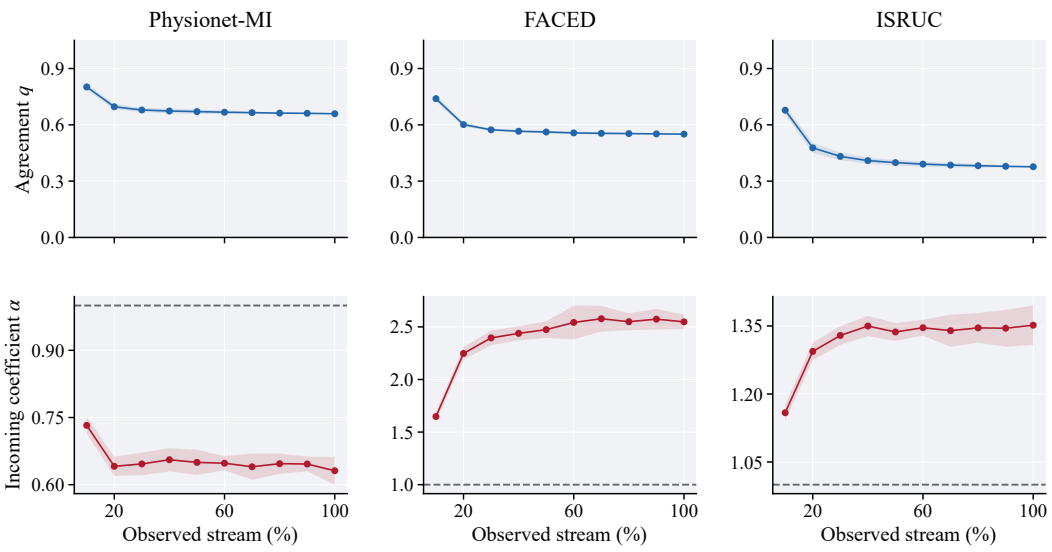



Figure 11: Checkpoint-derived RPT coefficients over the observed subject stream: agreement $q$ (top, blue) and incoming coefficient $\alpha$ (bottom, red). Within each order, values are averaged across transported layers at each step and then across steps in each stream decile. Curves show means over ten orders; bands show sample SD across orders. The dashed line denotes $\alpha = 1$, the incoming coefficient used by uniform averaging.

Table 11: Parameter-scope control (ACC/MF1 in %; BWT in pp; mean$_{\pm SD}$ over ten orders). The selected scope applies adaptive transport to fully connected layers (`nn.Linear`); alternatives extend transport to all non-BN tensors. Bold marks the best mean per dataset and metric.

| Dataset | Transport scope | ACC | MF1 | BWT-ACC | BWT-MF1 |
|---|---|---|---|---|---|
| Physionet-MI | All non-BN (global) | 53.10$_{\pm 0.05}$ | 46.43$_{\pm 0.03}$ | -0.50$_{\pm 0.76}$ | -0.51$_{\pm 0.82}$ |
| | All non-BN (module-wise) | 51.99$_{\pm 0.14}$ | 45.53$_{\pm 0.23}$ | -0.73$_{\pm 0.96}$ | -0.58$_{\pm 0.97}$ |
| | **Linear layers (selected)** | **53.47**$_{\pm 0.11}$ | **46.64**$_{\pm 0.09}$ | **0.21**$_{\pm 0.87}$ | **0.36**$_{\pm 0.84}$ |
| FACED | All non-BN (global) | 55.96$_{\pm 0.19}$ | 46.06$_{\pm 0.09}$ | 1.47$_{\pm 0.79}$ | 1.76$_{\pm 0.69}$ |
| | All non-BN (module-wise) | 32.75$_{\pm 1.53}$ | 23.79$_{\pm 1.35}$ | -10.99$_{\pm 3.85}$ | -9.94$_{\pm 3.48}$ |
| | **Linear layers (selected)** | **58.39**$_{\pm 0.22}$ | **48.48**$_{\pm 0.25}$ | **1.79**$_{\pm 0.90}$ | **2.10**$_{\pm 0.95}$ |
| ISRUC | All non-BN (global) | 72.31$_{\pm 0.02}$ | 59.44$_{\pm 0.02}$ | **-0.05**$_{\pm 0.28}$ | -0.20$_{\pm 0.51}$ |
| | All non-BN (module-wise) | 72.07$_{\pm 0.06}$ | 58.82$_{\pm 0.08}$ | -0.24$_{\pm 0.33}$ | -0.57$_{\pm 0.33}$ |
| | **Linear layers (selected)** | **72.34**$_{\pm 0.02}$ | **59.51**$_{\pm 0.02}$ | -0.12$_{\pm 0.22}$ | **-0.18**$_{\pm 0.46}$ |

backbone, with one coefficient shared by each weight matrix and its bias. The remaining non-BN parameters use uniform averaging. The GN control in Appendix C.2 retains this scope and shows that its decoding advantage persists without BN.

### C.5 Performance–Resource Trade-offs

**CMM update costs.** Table 12 reports complete-stream measurements on an RTX 4090 over ten subject orders. The nine checkpoint-only methods use the same AM-BN as the controlled comparison in Table 2. Timing includes initialization, all update computations, and model reconstruction, with CUDA synchronized at both boundaries; checkpoint loading, upstream training, metric evaluation, and output I/O are excluded. Total time is divided by the number of subject arrivals before computing the mean and sample SD across orders. Peak memory is the maximum allocated GPU memory across updates and orders.

RPT completes updates substantially faster than the decomposition- or optimization-based NUFILT, DOP, and WUDI-Merging, while using less peak GPU memory. WA provides the lowest-cost ref-

Table 12: CMM update resources over ten subject orders. Time is mean$_{\pm \mathrm{SD}}$ seconds per subject; peak is maximum allocated GPU memory in MiB. Checkpoint-only methods use matched AM-BN.

| | Physionet-MI | | FACED | | ISRUC | |
|---|---|---|---|---|---|---|
| Method | Time | Peak | Time | Peak | Time | Peak |
| WA | $0.029_{\pm 0.001}$ | 8.12 | $0.050_{\pm 0.002}$ | 8.12 | $0.032_{\pm 0.001}$ | 8.12 |
| Task Arithmetic | $0.034_{\pm 0.002}$ | 8.12 | $0.052_{\pm 0.001}$ | 8.12 | $0.038_{\pm 0.001}$ | 8.12 |
| TIES | $2.490_{\pm 0.058}$ | 8.12 | $2.756_{\pm 0.076}$ | 8.12 | $2.611_{\pm 0.060}$ | 8.12 |
| MagMax | $0.606_{\pm 0.008}$ | 8.12 | $0.676_{\pm 0.013}$ | 8.12 | $0.622_{\pm 0.009}$ | 8.12 |
| OPCM | $1.889_{\pm 0.063}$ | 156.98 | $1.914_{\pm 0.070}$ | 157.20 | $1.764_{\pm 0.037}$ | 66.21 |
| NUFILT | $2.505_{\pm 0.051}$ | 468.92 | $2.654_{\pm 0.048}$ | 468.92 | $2.662_{\pm 0.066}$ | 115.34 |
| WUDI-Merging | $2.638_{\pm 0.013}$ | 328.47 | $2.781_{\pm 0.031}$ | 361.01 | $2.793_{\pm 0.027}$ | 336.93 |
| DOP | $12.944_{\pm 0.089}$ | 285.92 | $13.484_{\pm 0.296}$ | 286.14 | $13.139_{\pm 0.133}$ | 112.00 |
| **RPT** | $0.443_{\pm 0.017}$ | 8.12 | $0.482_{\pm 0.016}$ | 8.12 | $0.423_{\pm 0.017}$ | 8.12 |

erence. RPT retains the original backbone topology at deployment, and its incremental summaries have size independent of the number of encountered subjects; the GPU peak does not measure this CPU-resident state.

**Sequential-learning costs.** All sequential-learning runs use one RTX 4090. End-to-end time includes training, model updating or merging, and evaluation, and is divided by the number of arriving subjects. Table 13 reports the mean and SD over ten orders and the maximum allocated GPU memory. These measurements use the same method configurations as the decoding comparison in Table 4.

Table 13 reports decoding performance, end-to-end runtime, peak GPU memory, and replay capacity for these same method configurations. Seq-RPT retains no historical EEG, while the included replay baselines store 200 segments on Physionet-MI and FACED and 500 on ISRUC. This comparison describes the computational and replay-storage costs associated with sequential decoding performance.

Table 13: Decoding performance and resources for sequential learning. Score is the mean of ACC and MF1 (%). Time is mean$_{\pm \mathrm{SD}}$ seconds per subject over ten orders, including training, updating, and evaluation. Peak is maximum allocated GPU memory (GiB); replay is the capacity in historical EEG segments. All method configurations match Table 4.

| Method | Score (%) | Time (s) | Peak (GiB) | Replay |
|---|---|---|---|---|
| **Physionet-MI** | | | | |
| EWC | 44.97 | $8.13_{\pm 0.74}$ | 0.48 | 0 |
| ER | 49.92 | $5.33_{\pm 0.40}$ | 0.35 | 200 |
| DER++ | 48.94 | $10.75_{\pm 0.28}$ | 0.43 | 200 |
| MUDVI | 46.88 | $8.36_{\pm 0.51}$ | 0.35 | 200 |
| ReSNT | 49.24 | $6.87_{\pm 0.30}$ | 1.72 | 200 |
| AMBM | 48.65 | $12.92_{\pm 0.83}$ | 0.65 | 200 |
| BrainUICL-SR | 48.54 | $6.75_{\pm 0.09}$ | 0.35 | 200 |
| ProNECL | 38.73 | $6.59_{\pm 0.10}$ | 0.29 | 0 |
| **Seq-RPT** | 48.70 | $6.68_{\pm 0.10}$ | 0.30 | 0 |
| **FACED** | | | | |
| EWC | 47.80 | $4.21_{\pm 0.10}$ | 0.60 | 0 |
| ER | 15.22 | $3.96_{\pm 0.15}$ | 0.60 | 200 |
| DER++ | 24.46 | $4.68_{\pm 0.22}$ | 0.51 | 200 |
| MUDVI | 46.07 | $6.96_{\pm 0.48}$ | 0.60 | 200 |
| ReSNT | 14.45 | $6.37_{\pm 0.10}$ | 7.03 | 200 |
| AMBM | 19.38 | $5.83_{\pm 0.05}$ | 0.92 | 200 |
| BrainUICL-SR | 16.87 | $4.04_{\pm 0.09}$ | 0.60 | 200 |
| ProNECL | 45.10 | $3.14_{\pm 0.09}$ | 0.47 | 0 |
| **Seq-RPT** | 46.68 | $3.66_{\pm 0.16}$ | 0.47 | 0 |
| **ISRUC** | | | | |
| EWC | 31.90 | $11.96_{\pm 0.73}$ | 1.70 | 0 |
| ER | 67.23 | $13.29_{\pm 0.25}$ | 2.68 | 500 |
| DER++ | 66.77 | $19.67_{\pm 0.27}$ | 3.63 | 500 |
| MUDVI | 61.48 | $14.99_{\pm 0.45}$ | 2.68 | 500 |
| ReSNT | 67.62 | $17.73_{\pm 0.39}$ | 14.62 | 500 |
| AMBM | 60.50 | $17.97_{\pm 0.43}$ | 4.15 | 500 |
| BrainUICL-SR | 64.31 | $12.91_{\pm 0.22}$ | 2.68 | 500 |
| ProNECL | 53.96 | $10.68_{\pm 0.36}$ | 1.55 | 0 |
| **Seq-RPT** | 65.41 | $9.38_{\pm 0.19}$ | 1.55 | 0 |

### C.6 Robustness Across Subject Splits and Stream Lengths

Table 14: Robustness across five subject splits (ACC/MF1 in %; BWT in pp). Results first average ten random subject orders within each split; entries show mean ± SD across splits. Bold marks the best result per dataset and metric.

| Method | ACC (%) | MF1 (%) | $BWT_A$ (pp) | $BWT_F$ (pp) |
|---|---|---|---|---|
| **Physionet-MI** | | | | |
| WA | $52.82_{\pm 2.49}$ | $45.82_{\pm 3.26}$ | $-0.14_{\pm 0.12}$ | $-0.15_{\pm 0.32}$ |
| Task Arithmetic | $24.04_{\pm 1.64}$ | $10.52_{\pm 0.73}$ | $-5.86_{\pm 0.98}$ | $-8.42_{\pm 0.53}$ |
| TIES | $51.14_{\pm 1.48}$ | $44.27_{\pm 1.77}$ | $-1.23_{\pm 0.72}$ | $-1.26_{\pm 0.99}$ |
| MagMax | $48.96_{\pm 3.63}$ | $41.87_{\pm 3.06}$ | $-1.02_{\pm 0.47}$ | $-1.07_{\pm 0.28}$ |
| OPCM | $23.59_{\pm 2.05}$ | $11.77_{\pm 2.07}$ | $-1.98_{\pm 0.47}$ | $-3.33_{\pm 0.96}$ |
| NUFILT | $51.43_{\pm 3.14}$ | $44.03_{\pm 3.77}$ | $-1.63_{\pm 1.12}$ | $-1.95_{\pm 1.09}$ |
| WUDI-Merging | $48.64_{\pm 1.54}$ | $40.50_{\pm 2.22}$ | $-4.47_{\pm 1.80}$ | $-5.61_{\pm 2.75}$ |
| DOP | $49.94_{\pm 1.83}$ | $42.88_{\pm 2.06}$ | $-5.17_{\pm 1.88}$ | $-6.12_{\pm 2.25}$ |
| AdaMerging | $24.91_{\pm 1.40}$ | $12.13_{\pm 1.29}$ | $-1.45_{\pm 0.46}$ | $-3.94_{\pm 1.26}$ |
| MINGLE | $41.94_{\pm 2.45}$ | $32.77_{\pm 2.80}$ | $-7.17_{\pm 3.13}$ | $-9.26_{\pm 3.71}$ |
| **RPT** | $\mathbf{53.96}_{\pm 1.60}$ | $\mathbf{47.05}_{\pm 2.61}$ | $\mathbf{0.20}_{\pm 0.15}$ | $\mathbf{0.38}_{\pm 0.28}$ |
| **FACED** | | | | |
| WA | $50.20_{\pm 1.54}$ | $40.20_{\pm 1.72}$ | $0.43_{\pm 0.56}$ | $0.44_{\pm 0.60}$ |
| Task Arithmetic | $17.51_{\pm 8.29}$ | $11.31_{\pm 6.51}$ | $-12.40_{\pm 4.65}$ | $-10.82_{\pm 3.93}$ |
| TIES | $46.56_{\pm 4.47}$ | $36.80_{\pm 4.32}$ | $0.57_{\pm 0.54}$ | $0.70_{\pm 0.42}$ |
| MagMax | $47.22_{\pm 3.95}$ | $37.46_{\pm 4.08}$ | $0.33_{\pm 0.57}$ | $0.41_{\pm 0.59}$ |
| OPCM | $12.08_{\pm 1.48}$ | $7.21_{\pm 1.13}$ | $-7.02_{\pm 3.69}$ | $-5.78_{\pm 3.09}$ |
| NUFILT | $51.35_{\pm 1.50}$ | $41.32_{\pm 1.57}$ | $1.54_{\pm 1.13}$ | $1.62_{\pm 1.09}$ |
| WUDI-Merging | $47.04_{\pm 1.35}$ | $37.40_{\pm 1.18}$ | $-0.80_{\pm 1.29}$ | $-0.43_{\pm 0.98}$ |
| DOP | $46.19_{\pm 10.91}$ | $36.72_{\pm 9.90}$ | $-0.45_{\pm 3.40}$ | $-0.37_{\pm 3.64}$ |
| AdaMerging | $10.87_{\pm 1.07}$ | $5.95_{\pm 0.90}$ | $-6.45_{\pm 2.59}$ | $-5.38_{\pm 2.08}$ |
| MINGLE | $40.60_{\pm 2.77}$ | $31.60_{\pm 2.19}$ | $-6.48_{\pm 3.29}$ | $-5.46_{\pm 2.33}$ |
| **RPT** | $\mathbf{54.85}_{\pm 2.08}$ | $\mathbf{44.81}_{\pm 2.28}$ | $\mathbf{1.58}_{\pm 0.44}$ | $\mathbf{1.68}_{\pm 0.54}$ |
| **ISRUC** | | | | |
| WA | $67.37_{\pm 5.03}$ | $54.65_{\pm 5.76}$ | $-0.22_{\pm 0.18}$ | $-0.20_{\pm 0.22}$ |
| Task Arithmetic | $30.01_{\pm 10.87}$ | $14.04_{\pm 6.63}$ | $-14.74_{\pm 2.60}$ | $-15.01_{\pm 1.15}$ |
| TIES | $69.56_{\pm 1.73}$ | $56.23_{\pm 2.45}$ | $-1.00_{\pm 0.51}$ | $-1.49_{\pm 0.46}$ |
| MagMax | $68.64_{\pm 2.59}$ | $55.38_{\pm 3.61}$ | $-0.72_{\pm 0.61}$ | $-0.68_{\pm 0.62}$ |
| OPCM | $26.07_{\pm 4.30}$ | $10.10_{\pm 2.46}$ | $-4.82_{\pm 2.03}$ | $-5.49_{\pm 1.50}$ |
| NUFILT | $66.21_{\pm 5.11}$ | $53.63_{\pm 5.44}$ | $-1.34_{\pm 0.61}$ | $-1.16_{\pm 0.61}$ |
| WUDI-Merging | $67.17_{\pm 3.93}$ | $53.93_{\pm 4.52}$ | $-2.34_{\pm 0.63}$ | $-2.61_{\pm 0.75}$ |
| DOP | $68.74_{\pm 2.59}$ | $55.51_{\pm 2.46}$ | $-2.37_{\pm 2.20}$ | $-3.28_{\pm 2.56}$ |
| AdaMerging | $22.99_{\pm 5.62}$ | $9.51_{\pm 2.08}$ | $-7.10_{\pm 4.57}$ | $-7.21_{\pm 2.19}$ |
| MINGLE | $41.60_{\pm 3.85}$ | $27.59_{\pm 3.06}$ | $-12.94_{\pm 2.58}$ | $-15.42_{\pm 2.97}$ |
| **RPT** | $\mathbf{71.52}_{\pm 1.66}$ | $\mathbf{59.07}_{\pm 1.16}$ | $\mathbf{0.04}_{\pm 0.61}$ | $\mathbf{0.02}_{\pm 0.24}$ |

The main experiments hold the pre-training and incremental subject pools fixed to isolate variability from subject arrival. Table 14 additionally changes that division five times for every CMM method. RPT has the highest across-split mean for all twelve dataset–metric endpoints, including both BWT metrics on all three datasets. Its ACC/MF1 margins over the strongest comparator are $1.14/1.23$ points on Physionet-MI, $3.50/3.49$ on FACED, and $1.96/2.84$ on ISRUC. Its BWT-ACC/BWT-MF1 means are $0.20/0.38$, $1.58/1.68$, and $0.04/0.02$ points, respectively. This experiment addresses sensitivity to which subjects form the pre-training and incremental pools; it is distinct from the order SD reported for the fixed main split.

Figure 12 evaluates RPT after 25%, 50%, 75%, and 100% of each stream. ACC and MF1 remain stable or improve as more subjects are incorporated, while BWT stays close to zero on Physionet-MI and ISRUC and positive on FACED. Variation across arrival orders also narrows toward the full

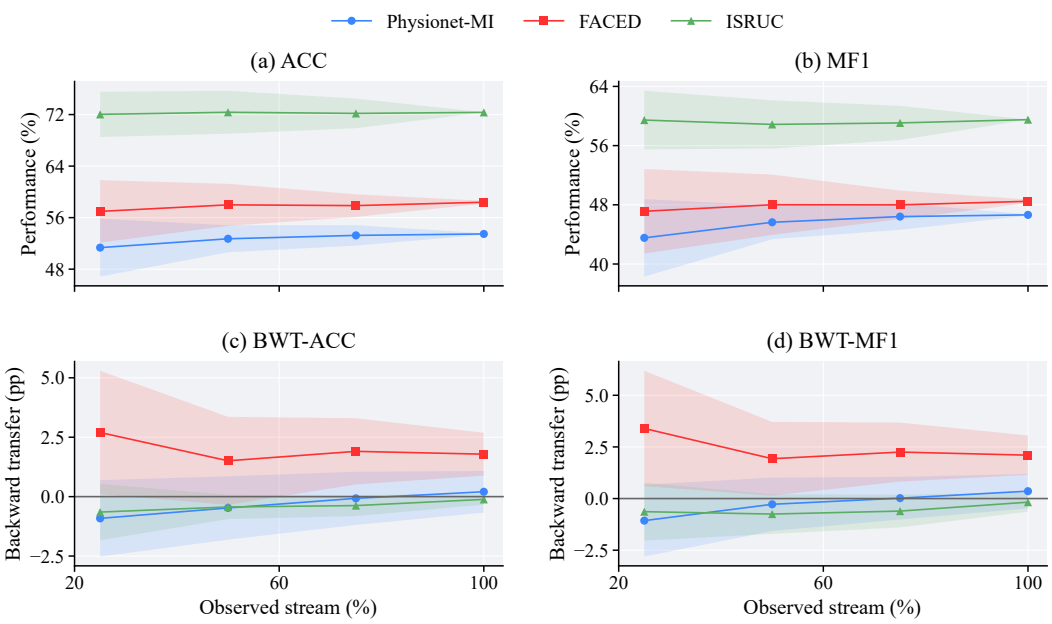



Figure 12: RPT across increasing fractions of each subject stream. Curves report ACC, MF1, BWT-ACC, and BWT-MF1; bands show SD across ten random subject orders. ACC/MF1 are in %, and BWT is in pp. All fractions use prefixes of the same complete streams and the dataset-level settings fixed in the main experiment.

stream. Together with Table 14, these results assess sensitivity to stream length, subject order, and the composition of the subject pools.

## D LIMITATIONS

The present study considers subject-incremental EEG decoding with a fixed label space. Extending checkpoint-based merging to streams that introduce new classes or recording configurations is a useful direction for broader deployment. RPT also uses a dataset-level transport exponent that remains fixed throughout each stream. Although the sensitivity analysis identifies effective parameter regions, selecting this exponent directly from evolving checkpoint statistics could reduce development effort and better accommodate changes in the arriving subject population.