WAVE: OPTIMAL-TRANSPORT-BASED DOMAIN-INCREMENTAL LEARNING AND GENERALIZATION FOR CROSS-SUBJECT EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain activity varies substantially across individuals, causing Electroencephalography (EEG) decoding models to degrade when applied to new subjects. Each individual shifts the signal underlying probability measure, and models trained on a fixed population generalize poorly to unseen subjects. We cast cross-subject adaptation as domain-incremental learning (DIL), treating each subject as a domain and adapting sequentially through a replay memory. We introduce `WAVE`, a probabilistic approach that represents each subject as a hierarchical probability measure over classes and a replay memory as a measure over subjects, which are natural objects in a Wasserstein over Wasserstein (WoW) space. As such, `WAVE` learns a replay buffer by minimizing an optimal-transport divergence in WoW space and augments data with virtual subjects generated between stored and incoming domains for smoother adaptation and improved generalization. Because the memory stores compressed, synthetic subjects rather than raw recordings, WAVE also avoids retaining privacy-sensitive EEG data. Applied to EEG foundation models, WAVE improves the stability-plasticity-generalization trade-off and matches strong replay baselines without storing real EEG.

1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive measurement of electrical brain activity and is widely used in brain-computer interfaces (BCIs) (Gu et al., 2021), emotion recognition, motor imagery classification and sleep staging (Dadebayev et al., 2022; Altaheri et al., 2021; Roy et al., 2019). Despite recent advances in deep learning for EEG decoding (Wang et al., 2025; Jiang et al., 2024), deploying models across individuals remains challenging due to inter-subject variability in anatomy, neural activity, recording conditions, and behavioral strategies (Saha & Baumert, 2019), resulting in distribution shifts (Quiñonero-Candela et al., 2008). Models trained on fixed populations thus degrade on new subjects without adaptation, motivating methods that continuously adapt to incoming subjects while generalizing to unseen ones (Li et al., 2026b). This has motivated interest in large-scale pretrained models. Recent EEG Foundation Models (FMs) (Wang et al., 2025; Jiang et al., 2024) provide strong representations for subject adaptation, but their large capacity can also lead to overfitting to individual subjects and sensitivity to distribution shifts across subjects.

In real-world scenarios, subjects may arrive sequentially, each providing labeled data for adaptation. Following the Continual Learning (CL) taxonomy of van de Ven & Tolias (2019), we consider each subject as a distinct domain and formulate the problem as DIL (cf. Appendix B.1.1). The model must progressively adapt to these changing subject-specific distributions while retaining knowledge acquired from previous subjects. This setting highlights the stability-plasticity dilemma (Mermillod et al., 2013): *plasticity* refers to the ability to adapt to newly encountered subjects, whereas *stability* denotes the ability to preserve knowledge from previously seen subjects. Insufficient stability can result in *Catastrophic Forgetting (CF)*, where learning from new domains causes a substantial degradation in performance on previously learned domains (Kirkpatrick et al., 2016). Balancing these properties is particularly challenging in EEG due to substantial inter-subject variability, motivating methods that can effectively adapt to new individuals while generalizing to unseen subjects.

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.

Sequential Supervised Finetuning (SFT) can effectively adapt a model to a new subject, but it may progressively bias it toward recently observed subjects, and is thus prone to CF. CL addresses this stability-plasticity trade-off through various approaches. Regularization-based methods add a penalty balancing old and new domains, either over weights, as in Elastic Weight Consolidation (EWC) Kirkpatrick et al. (2016), or outputs, as in Learning without Forgetting (LwF) Li & Hoiem (2016) (cf. Appendix B.1.2). Replay-based methods instead retain data from previously observed domains: episodic memory is an effective, simple strategy for mitigating forgetting under limited budgets (Chaudhry et al., 2019), while GDumb Prabhu et al. (2020) and Dark Experience Replay (DER) Buzzega et al. (2020) further demonstrate their strong performance, and random reservoir sampling maintains a fixed-size memory as new data arrive Vitter (1985). Recent works (Zhou et al., 2025b;a; Li et al., 2026a; Yang et al., 2026) have explored CL for EEG by operating in parameter, sample, or prototype space, but leave the underlying subject-level distributions largely unmodeled.

In our setting, each subject is itself a probability measure over its class-conditional measures, making the memory a measure over subject measures (Figure 1). A distributional perspective is therefore well suited to comparing subjects. Optimal Transport (OT) provides a principled framework for comparing probability measures while accounting for the geometry of the underlying sample space (Peyré & Cuturi, 2019b). Wasserstein distances have consequently been used for distribution alignment and domain adaptation (Courty et al., 2015; Shen et al., 2018), including EEG adaptation across motor-imagery subjects (She et al., 2023). This perspective is also related to dataset distillation and condensation, which preserve the information of large datasets in compact representations (Wang et al., 2020a; Zhao et al., 2021; Zhao & Bilen, 2023; Rosasco et al., 2021). Such hierarchical distributional representations can be studied through Wasserstein-over-Wasserstein (WoW) geometry (Bonet et al., 2025; Nguyen, 2016). Rather than retaining a potentially large collection of individual EEG samples, a distributional memory thus offers a compact representation of previously observed subjects while enabling subject-domain comparison and virtual-subject generation.

To further improve the cross-domain generalization of EEG foundation models, we introduce virtual-subject augmentation based on McCann's interpolation (McCann, 1997) between subjects stored in the replay memory and the current subject. El Hamri et al. (2025) previously used it for gradual domain adaptation along Wasserstein geodesics. In our case, we use the interpolated subjects as additional training data rather than as intermediate domains for progressive adaptation, aiming to improve adaptation to the current subject while enhancing generalization to unseen ones.

In this work, we propose a new method called **W**asserstein-geodesic **A**ugmentation via **V**irtual subjects and **E**xperience-replay (`WAVE`), a domain-incremental learning (DIL) and generalization method for EEG decoding, that treats the incremental SFT of EEG FMs from a distributional perspective. We adopt a replay-based approach in which previously observed subjects are compressed into a fixed-size memory using OT rather than random resampling. Each subject is represented as a probability measure, and the Wasserstein geometry is used to construct a compact memory that preserves subject-level information. During continual adaptation, this memory is replayed alongside the current subject and the generated virtual subjects, enabling knowledge retention, smoother adaptation, and improved generalization while limiting the storage of historical EEG samples.

Paper Organization. The rest of the paper is organized as follows. Section 2 introduces the problem formulation, memory construction, and virtual-subject augmentation strategy. Section 3 describes the experimental protocol, datasets, baselines and evaluation metrics, and presents our results and ablation studies. Section 4 concludes the paper and discusses limitations and future directions.

Contributions. Our contributions are as follows:

1. A WoW memory compression framework representing observed subjects as measures over their EEG measures, enabling compact replay while preserving subject-level geometry.
2. A virtual-subject augmentation strategy based on McCann interpolation between stored and incoming subject measures, enabling smoother adaptation and improved generalization.

2 METHODOLOGY

We consider the problem of continual cross-subject EEG classification. In that sense, we consider the data to be split among a set of N subjects. These subjects are split into source $\mathcal{S}$, incremental $\mathcal{I}$

and generalization $\mathcal{G}$, with N_S, N_I and N_G subjects each, so that $N = N_S + N_I + N_G$. Each subject s has a labeled sample set $\mathcal{D}_s = \{(\mathbf{X}_{s,i}, y_{s,i})\}_{i=1}^{n_s}$, where $\mathbf{X}_{s,i} \in \mathbb{R}^{E\times T}$ is the i-th EEG sample over E electrodes and T time steps, $\mathbf{x}_{s,i} \in \mathbb{R}^d$ its flattened form with $d = E \cdot T$, and $y_{s,i} \in \{1, \dots, C\}$ its class label. Henceforth we denote the joint ambient space via $\Omega = \mathbb{R}^d \times \{1, \cdots, C\}$ whose elements are the tuples $(\mathbf{x}, y)$. We model the measure over subjects hierarchically, at 3 distinct levels: subjects (higher level), classes (intermediate level), and samples (lower level), that is,

$$\mathbb{M} = \frac{1}{N_S}\sum_{s=1}^{N_S}\delta_{\mu_s} \quad \mu_s = \sum_{c=1}^{C}\pi_{s,c}\delta_{\mu_{s,c}} \quad \mu_{s,c} = \frac{1}{n_{s,c}}\sum_{i:y_{s,i}=c}\delta_{(\mathbf{x}_{s,i},y_{s,i})}, \tag{1}$$

where $\delta_{\mathbf{x}}$ denotes the Dirac measure centered at $\mathbf{x}$. Furthermore, $n_{s,c}$ is the number of samples of the $c-$th class of the $s-$th subject, $n_s = \sum_c n_{s,c}$, and $\pi_{s,c} = {}^{n_{s,c}}/_{n_s}$ is the proportion of the $c-$th class in the $s-$th subject. Equation 1 defines a tower of measures, with $\mu_{s,c} \in \mathcal{P}_2(\Omega)$, $\mathcal{P}_2(\Omega)$ being the set of measures over the set Ω, $\mu_s \in \mathcal{P}_2(\mathcal{P}_2(\Omega))$ and $\mathbb{M} \in \mathcal{P}_2(\mathcal{P}_2(\mathcal{P}_2(\Omega)))$. An illustration is shown in Figure 1. We further define the following sets of empirical measures that we work on,

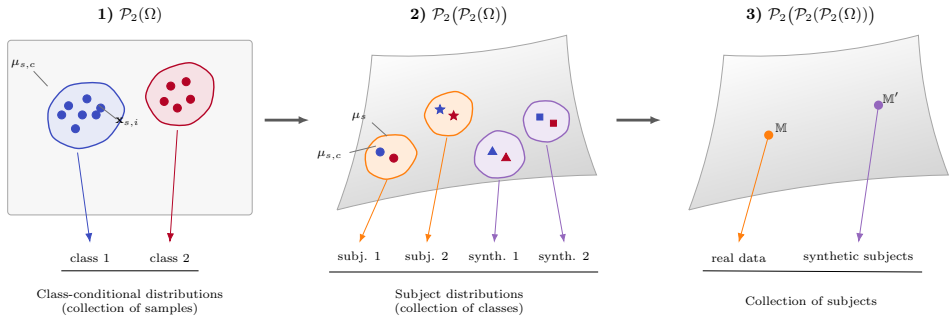


Figure 1: Hierarchical modeling of subject measures. A class-conditional $\mu_{s,c}$ is a set of samples in $\mathcal{P}_2(\Omega)$; a subject collects classes, $\mu_s \in \mathcal{P}_2(\mathcal{P}_2(\Omega))$; subjects form a dataset $\mathbb{M} \in \mathcal{P}_2(\mathcal{P}_2(\mathcal{P}_2(\Omega)))$.

$$\text{Emp}_n(\Omega) = \left\{\mu \in \mathcal{P}_2(\Omega) : \mu = {}^1/_n \sum_{i=1}^{n}\delta_{(\mathbf{x}_i,y_i)}\right\} \subset \mathcal{P}_2(\Omega).$$

This construction works similarly to $\mathcal{P}_2(\Omega)$, meaning, $\mu_s \in \text{Emp}_C(\text{Emp}_{n_{s,c}}(\Omega))$ defines the measure over classes and $\mathbb{M} \in \text{Emp}_{N_S}(\text{Emp}_C(\text{Emp}_{n_{s,c}}(\Omega)))$ denotes the measure over source subjects. For the sake of notation compactness, we use $\text{Emp}_{N_S,C,n_{s,c}}(\Omega)$ to denote the complete tower.

Our main goal is the subject-incremental SFT of EEG FMs. We consider FMs consisting of an encoder $f_\theta : \mathcal{X} \to \mathcal{Z}$, and a classification head $h_\phi : \mathcal{Z} \to \mathcal{Y}$. The model is therefore $g_{\theta,\phi} = h_\phi \circ f_\theta$, mapping $\mathcal{X}$ to $\mathcal{Y}$. To that end, we consider a pool of *source subjects*, $\mathcal{D}_S$ for which we initially minimize the pooled loss via Stochastic Gradient Descent (SGD)

$$\theta_0, \phi_0 = \underset{\theta\in\Theta,\phi\in\Phi}{\arg\min}\, \mathbb{E}_{(\mathbf{X},y)\sim\sum_s \omega_s\mu_s}[\ell(y, g_{\theta,\phi}(\mathbf{X}))], \text{ where } \omega_s = \frac{n_s}{\sum_{s=1}^{N_s} n_s}.$$

After this source pool, we consider a pool of incremental subjects $\mathcal{D}_I$. For notational convenience, we index them by arrival order: the subject μ_{N_S+t}, the t-th subject to arrive after the N_S source subjects, is denoted μ_t for $t \in \{1, \dots, N_I\}$. At each new arrival μ_t, we perform SFT by minimizing,

$$\mathcal{L}_t(\theta, \phi) = \lambda_c \mathbb{E}_{\mu_t}[\ell(y, g_{\theta,\phi}(\mathbf{X}))] + \lambda_r \mathbb{E}_{\mathbb{M}_t}[\ell(y, g_{\theta,\phi}(\mathbf{X}))] + \lambda_a \mathbb{E}_{\sum_k \rho_k \bar{\mu}_k}[\ell(y, g_{\theta,\phi}(\mathbf{X}))] \tag{2}$$

where $\mathbb{M}_t$ is a *replay memory* at stage t (see Section 2.1), and $\bar{\mu}_k$ are virtual augmented subjects (see Section 2.2). Formally,

$$\theta_t, \phi_t = \underset{\theta\in\Theta,\phi\in\Phi}{\arg\min}\, \mathcal{L}_t(\theta, \phi).$$

Throughout the SFT process, we keep a family of N_G unseen subjects $\{\mu_s\}_{s=N_S+N_I+1}^{N}$, which are used to measure the generalization capabilities of the model. ℓ denotes label-smoothed cross-entropy, using soft cross-entropy for Mixup targets and standard cross-entropy otherwise. Both the memory compression and the virtual augmentation operate on the flattened representation $\mathbf{x}_{s,i} \in \mathbb{R}^d$; the resulting vectors are then reshaped back into the EEG space $\mathbb{R}^{E\times T}$ for model training.

Algorithm 1 Continual SFT of an EEG FM with replay and augmentation

Require: source subjects $\{\mu_s\}_{s=1}^{N_S}$ with weights w_s; incremental stream $\{\mu_{N_S+t}\}_{t=1}^{N_I}$; hold-out pool $\{\mu_{N_S+N_I+g}\}_{g=1}^{N_G}$; loss weights $\lambda_c, \lambda_r, \lambda_a \in [0,1]$

$\theta_0, \phi_0 \leftarrow \arg\min_{\theta\in\Theta,\phi\in\Phi} \mathbb{E}_{(\mathbf{X},y)\sim\sum_s \omega_s\mu_s}[\ell(y, g_{\theta,\phi}(\mathbf{X}))]$ ▷ pretrain on source subjects via SGD

$\mathbb{M}_1 \leftarrow \text{INITIALIZEMEMORY}(\{\mu_s\}_{s=1}^{N_S})$

for each incoming μ_t, $t = 1, 2, \ldots, N_I$ **do**

  $\{\bar{\mu}_k, \rho_k\}_{k=1}^{K} \leftarrow \text{AUGMENTSUBJECTS}(\mu_t, \mathbb{M}_t)$

  $\mathcal{L}_t(\theta,\phi) \leftarrow \lambda_c\, \mathbb{E}_{\mu_t}[\ell(y, g_{\theta,\phi}(\mathbf{X}))] + \lambda_r\, \mathbb{E}_{\mathbb{M}_t}[\ell(y, g_{\theta,\phi}(\mathbf{X}))] + \lambda_a\, \mathbb{E}_{\sum_k \rho_k\bar{\mu}_k}[\ell(y, g_{\theta,\phi}(\mathbf{X}))]$

  $\theta_t, \phi_t \leftarrow \arg\min_{\theta\in\Theta,\phi\in\Phi} \mathcal{L}_t(\theta,\phi)$ ▷ SFT via SGD, initialized from the previous iterate

  $\mathbb{M}_{t+1} \leftarrow \text{UPDATEMEMORY}(\mathbb{M}_t, \mu_t)$

  Evaluate $g_{\theta,\phi}$ on $\{\mu_{N_S+N_I+g}\}_{g=1}^{N_G}$ ▷ track generalization

end for

2.1 MEMORY CONSTRUCTION

The replay memory prevents CF, keeping the model from overfitting to each new incremental subject. We denote it by $\mathbb{M}_t$, where $t = 1, \ldots, N_I$ indexes the adaptation step, and assume a budget of at most $B \geq N_S$ virtual subjects, each with m samples. Initialize $\mathbb{M}_1 := \mathbb{M}_{\text{src}} \in \text{Emp}_{N_S,C,n_{s,c}}$, drawing m samples per source subject at random ($\sum_c n_{s,c} = m$). At any time t, the memory holds $N_t \leq B$ subjects $\nu_s \in \text{Emp}_{C,n_{s,c}}(\Omega)$, with $\sum_c n_{s,c} = m$ and $s = 1, \cdots, N_t$. A fixed budget is necessary for two main reasons. First, storing all observed subjects is infeasible in a continual setting: the memory grows unbounded as $t \to +\infty$. Second, replaying an ever-growing memory raises the per-step adaptation cost and progressively biases training toward the past, undermining plasticity on the current subject. Each incremental subject is a measure over classes $\mu_t \in \mathcal{P}_2(\mathcal{P}_2(\Omega))$, lifted into the same space as $\mathbb{M}_t$ via $\mu_t \mapsto \delta_{\mu_t}$. From this lifting, we construct an augmented memory,

$$\tilde{\mathbb{M}}_t = \sum_{s=1}^{N_t} \frac{1}{N_t+1}\delta_{\nu_s} + \frac{1}{N_t+1}\delta_{\mu_t} \in \text{Emp}_{N_t+1,C,n_{s,c}}(\Omega),$$

which then contains $N_t + 1$ virtual subjects. We therefore propose a memory compression mechanism, inspired by the WoW geometry Nguyen (2016); Bonet et al. (2025), that maintains a memory of N_{t+1} subjects, where $N_{t+1} = min(N_t+1, B)$, while keeping the number of samples per subject fixed at m. The overall memory therefore holds at most B subjects,

$$\mathbb{M}_{t+1} = \underset{\mathbb{P}\in\text{Emp}_{N_{t+1},C,n_{s,c}}(\Omega)}{\arg\min} \mathbb{MMD}_\kappa(\mathbb{P}, \tilde{\mathbb{M}}_t)^2, \sum_c n_{s,c} = m \tag{3}$$

which is an optimization problem in the WoW space over the functional $\mathbb{P} \mapsto \mathbb{MMD}_\kappa(\mathbb{P}, \tilde{\mathbb{M}}_t)^2$. Let $\mathbb{P} = N^{-1}\sum_{s=1}^{N}\delta_{\mu_s}$ and $\mathbb{Q} = M^{-1}\sum_{s'=1}^{M}\delta_{\nu_{s'}}$, then,

$$\mathbb{MMD}_\kappa(\mathbb{P},\mathbb{Q})^2 = \frac{1}{N^2}\sum_{s=1}^{N}\sum_{s'=1}^{N}\kappa(\mu_s,\mu_{s'}) + \frac{1}{M^2}\sum_{s=1}^{M}\sum_{s'=1}^{M}\kappa(\nu_s,\nu_{s'}) - \frac{2}{NM}\sum_{s=1}^{N}\sum_{s'=1}^{M}\kappa(\mu_s,\nu_{s'}),$$

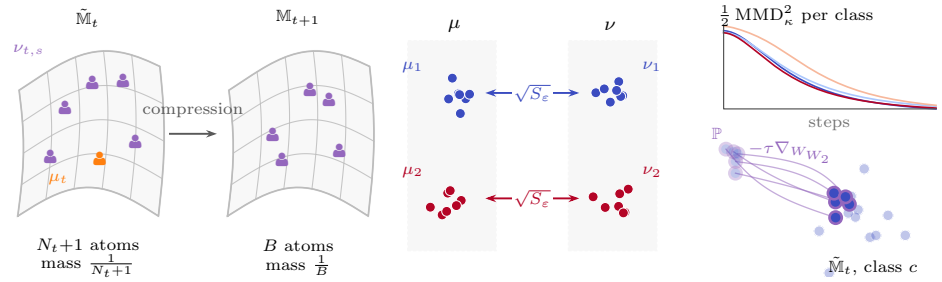
where $\kappa : \mathcal{P}_2(\mathcal{P}_2(\Omega))\times\mathcal{P}_2(\mathcal{P}_2(\Omega)) \to \mathbb{R}$ is a kernel. In this work we consider the Riesz kernel Bonet et al. (2025) between class-conditionals. Considering ν and μ as measures over classes,

$$\kappa(\mu,\nu) = -\frac{1}{C}\sum_{c=1}^{C}\sqrt{S_\epsilon(\mu_c,\nu_c)}, \text{where } \mu_c \text{ and } \nu_c \text{ are the class-conditional measures}$$

and S_ϵ is the debiased Sinkhorn divergence (cf. Eq. 22, Appendix B.2.2). Note that these functionals are ultimately a function of the samples $(\mathbf{x}_{s,i}, y_{s,i})$ that constitute the support of $\mathbb{P}$. Optimizing over $\mathbb{P}$ therefore amounts to optimizing directly over these samples, and the minimizer yields the compressed memory $\mathbb{M}_{t+1}$ that is carried to the next step. As Alvarez-Melis & Fusi (2021), we keep labels fixed and flow $\mathbf{x}_{s,i}$ using the gradient descent strategy of Bonet et al. (2025),

$$\mathbf{x}_{s,i} = \mathbf{x}_{s,i} - \tau\,\nabla_{W_{W_2}}\mathbb{MMD}_\kappa(\mathbb{P}, \tilde{\mathbb{M}}_t)^2(\mu_s)(\mathbf{x}_{s,i}), \text{ for } \mathbb{P} = N_{t+1}^{-1}\sum_{s=1}^{N_{t+1}}\delta_{\mu_s} \tag{4}$$

where $\nabla_{W_{W_2}}$ is the WoW gradient. An illustration is shown in Figure 2.



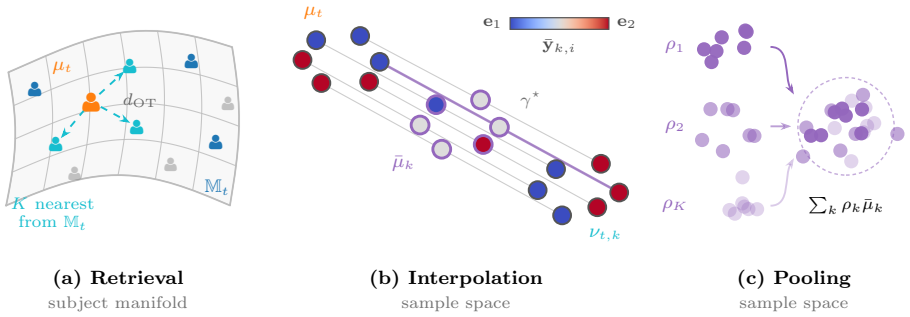
(a) Memory update subject level — **(b) Class-additive kernel** subject pair — **(c) WoW flow** sample space

Figure 2: Overview of WoW memory compression. *(a) Memory update (subject level):* the augmented memory $\tilde{\mathbb{M}}_t$, the N_t stored subjects plus the incoming subject μ_t as N_t+1 atoms of mass $\frac{1}{N_t+1}$, is compressed into $\mathbb{M}_{t+1}$ with at most B atoms of mass $\frac{1}{B}$. *(b) Class-additive kernel (subject pair):* subjects μ, ν are compared class by class, each pair (μ_c, ν_c) through $\sqrt{S_\varepsilon}$; summing the per-class terms forbids cross-class transport. *(c) WoW flow (sample space):* compression minimizes the per-class MMD$^2_\kappa$ by a WoW gradient flow moving the support of $\mathbb{P}$ along $-\tau\nabla_{W_{W_2}}$, per class c.

2.2 Virtual Subject Augmentation

Besides the continual adaptation to new subjects, our method seeks to *generalize* to unseen subjects $\{\mu_{N_S+N_I+s}\}_{s=1}^{N_G}$. To that end, inspired by the established literature of domain generalization through interpolations Zhang et al. (2017); Xu et al. (2019); Wang et al. (2020b), we interpolate between incoming subjects μ_t and the *virtual subjects* in the memory $\mathbb{M}_t$. While previous methods perform sample-level interpolation, we do measure-level interpolation based on the WoW geometry we introduced in Section 2.1.

Our approach is based on 3 steps. (i) Subject retrieval. Given μ_t, we select K subjects within the memory for performing the augmentation; (ii) Interpolation. Given a pair $(\mu_t, \nu_{t,k})$, we interpolate their point clouds using McCann's interpolant McCann (1997); El Hamri et al. (2025) and (iii) Pooling. We compose a pooled measure with the set of augmented subjects. We describe each of these in the following. A summary is shown in Figure 3.



(a) Retrieval subject manifold — **(b) Interpolation** sample space — **(c) Pooling** sample space

$$\mathcal{L}_t = \lambda_c \, \mathbb{E}_{\mu_t}[\ell] + \lambda_r \, \mathbb{E}_{\mathbb{M}_t}[\ell] + \lambda_a \, \mathbb{E}_{\sum_k \rho_k \bar{\mu}_k}[\ell]$$

Figure 3: Virtual subject interpolation. (a) We initially retrieve, in the memory $\mathbb{M}_t$, the K-nearest neighbors of the incoming incremental subject μ_t. (b) Then, we interpolate between $(\mu_t, \nu_{t,k})$ via OT. (c) Finally, we pool together the interpolated samples, weighted by their importance ρ_k.

Subject Retrieval. The memory is composed of at most B virtual subjects $\mathbb{M}_t = N_t^{-1} \sum_{s=1}^{N_t} \delta_{\nu_s}$. We then select the $K-$nearest neighbors of μ_t in $\mathbb{M}_t$, $\text{KNN}(\mu_t, \mathbb{M}_t)$, using,

$$d_{\text{OT}}(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \int d\big((\mathbf{x}, y), (\mathbf{x}', y')\big)\, \mathrm{d}\gamma, \tag{5}$$

where $d((\mathbf{x}, y), (\mathbf{x}', y')) = \sqrt{\|\mathbf{x} - \mathbf{x}'\|_2^2 + \|\text{Mean}[\mu_y] - \text{Mean}[\nu_{y'}]\|_2^2}$. The operator $\mu \mapsto \text{Mean}[\mu] = \int x \mathrm{d}\mu$ is the mean embedding of the measure μ, which is similar to the metric of Alvarez-Melis & Fusi (2020).

Virtual Subject Generation. For each selected neighbor $\{\nu_{t,k}\}_{k=1}^K$ we draw a single mixing weight $b_k \sim \text{Beta}(\alpha, \alpha)$, shared across all pairs from that neighbor. A virtual subject is produced by displacing samples of μ_t toward $\nu_{t,k}$ along the optimal-transport geodesic McCann (1997),

$$\bar{\mathbf{x}}_{k,i} = \mathbf{x}_{t,i} + (1 - b_k)\big(T_\gamma(\mathbf{x}_{t,i}) - \mathbf{x}_{t,i}\big), \quad \bar{\mathbf{y}}_{k,i} = \mathbf{y}_{t,i} + (1 - b_k)\big(T_\gamma(\mathbf{y}_{t,i}) - \mathbf{y}_{t,i}\big), \tag{6}$$

where T_γ denotes the barycentric mapping of Courty et al. (2015),

$$\gamma^\star = \arg\min_{\gamma \in \Gamma(\mu,\nu)} \langle \gamma, \mathbf{C} \rangle_F \quad T_\gamma(\mathbf{x}_{t,i}) = \sum_{j=1}^m \frac{\gamma_{ij}}{\sum_{j'=1}^m \gamma_{ij'}} \tilde{\mathbf{x}}_{t,k,j} \quad T_\gamma(\mathbf{y}_{t,i}) = \sum_{j=1}^m \frac{\gamma_{ij}}{\sum_{j'=1}^m \gamma_{ij'}} \tilde{\mathbf{y}}_{t,k,j} \tag{7}$$

where $\mathbf{x}_{t,i}$ denotes the $i-$th sample from μ_t, and $\tilde{\mathbf{x}}_{t,k,j}$ denotes the $j-$th sample from the $k-$nearest neighbor in $\mathbb{M}_t$. Note that we use one-hot encoded labels for blending the classes, meaning $\mathbf{y}_{t,i} \in \Delta_C$, the $C-$simplex. The interpolated samples $\{(\bar{\mathbf{x}}_{k,i}, \bar{\mathbf{y}}_{k,i})\}_{i=1}^m$ form the interpolation measure $\bar{\mu}_k$.

Pooling. We pool the K interpolated subjects $\{\bar{\mu}_k\}_{k=1}^K$ into a single sampling distribution by inverse distance weighting, where each neighbor $\nu_{t,k}$ is weighted by its relevance to μ_t:

$$\rho_k = \frac{(d_{\text{OT}}(\mu_t, \nu_{t,k}) + \varepsilon_\rho)^{-1}}{\sum_{k'} (d_{\text{OT}}(\mu_t, \nu_{t,k'}) + \varepsilon_\rho)^{-1}},$$

3 Experiments

We partition the dataset at the subject level into three disjoint sets: **source domains**, of N_S source subjects for an initial SFT step, **incremental domains** of N_I subjects that arrive sequentially during continual adaptation, and **generalization domains** of N_G unseen subjects. Source and incremental subjects are further partitioned into training/held-out test subsets for measuring *generalization to unseen samples*. We denote by $g_0 = g_{\theta_0,\phi_0}$ the model after SFT on the sources, and $g_t = g_{\theta_t,\phi_t}$ for the model after adaptation on each new incremental subject. At each step t, we evaluate g_t on the test partition of each incremental subject $t' \leq t$, and on the generalization subjects. We report Balanced Accuracy (BAcc) and Macro-F1 (MF1) along three axes: *plasticity*, *stability*, and *generalization*, whose full definitions are in Appendix E.2.

Aggregate score. To summarize the three axes in a single number, we combine them with a harmonic mean, as in the F1 score. It penalizes a collapse on any one axis and cannot be inflated by strong performance on the others, unlike an arithmetic mean. For each metric $m \in \{\text{BACC}, \text{MF1}\}$, the three axes are the current-subject adaptation $P(m) = \overline{m}_{\text{c}}$, the retention on previous incremental subjects $S(m) = \text{Ret}_I(m)$, and the anytime unseen-subject score $G(m) = \text{AA}_{N_I}(m)$. The per-metric score is their harmonic mean, and the final score averages over the two metrics:

$$H(m) = \frac{3}{P(m)^{-1} + S(m)^{-1} + G(m)^{-1}}, \qquad \text{Score} = \frac{1}{2}\big(H(\text{BACC}) + H(\text{MF1})\big). \tag{8}$$

Datasets We evaluate our method on five EEG datasets: BCI-IV-2a (Brunner et al., 2008), PhysioNet-MI (Goldberger et al., 2000), SHU-MI (Ma et al., 2022), FACED (Chen et al., 2023), and ISRUC (Khalighi et al., 2016). All datasets are resampled to 200 Hz for consistency with CBraMod (Wang et al., 2025). Table 1 summarizes the dataset characteristics and subject-level partitioning used in our experiments.

Table 1: EEG datasets and subject-level partitioning.

Dataset	Task	Subjects	E	T	C	Sampling	N_S	N_I	N_G	samples
BCI-IV 2a	Motor Imagery	9	22	800	4	250 Hz	2	5	2	5088
PhysioNet-MI	Motor Imagery	109	64	800	4	160 Hz	19	70	20	9837
SHU-MI	Motor Imagery	25	32	800	2	250 Hz	5	15	5	11988
FACED	Emotion Recognition	123	32	2000	9	250 Hz	20	80	23	10332
ISRUC	Sleep Staging	100	6	6000	5	200 Hz	20	60	20	89240

Compared Baselines: We compare our approach with several baselines. **Source-only:** the model g_0 is trained only on the source subjects and kept fixed, measuring the performance without adaptation to incremental subjects. **Oracle (Joint Training):** the model is jointly trained using data from both source and incremental subjects, assuming that all subjects are available simultaneously, and serves as an upper-bound reference without sequential-learning constraints. **Naive fine-tuning:** the lower bound, the model is sequentially fine-tuned on each incoming subject without preserving previous knowledge. **Random replay:** a fixed-budget of size $B{\cdot}m$ replay buffer is maintained by randomly sampling and storing previously observed training samples; when a new domain arrives, the budget is split evenly across domains and each is randomly subsampled to its quota. This domain-balanced rehearsal adapts the fixed-buffer, balanced-sampling strategy GDumb (Prabhu et al., 2020), originally class-balanced for class-incremental learning, to the domain-incremental setting, and serves as a strong baseline against WoW's gradient-flow distillation. **EWC:** penalizes changes to parameters important for past knowledge via a Fisher-information term (Kirkpatrick et al., 2016). **LwF:** distills the previous model's predictions while adapting to the current subject (Li & Hoiem, 2016). See Appendix B.1.2 for details.

We use CBraMod (Wang et al., 2025), with a two-layer Multilayer Perceptron (MLP) classifier, trained on the source subjects to obtain g_0 and then fine-tuned on each incoming subject. Training details and hyperparameter grids for all methods are in Appendix F (Tables 10-11), and our full hyperparameter sweep across all the datasets in Appendix G.1. We report here the incremental-subject stability. The source-subject stability, for main and ablation experiments, is in Appendix I.

Tables 2, 3, 4, 5 and 6 report the results on all the datasets, with each method evaluated under its best *Score* hyperparameter configuration. `WAVE` outperforms all baselines on FACED. On PhysioNet-MI all methods perform similarly: `WAVE` and random replay reach the same Score (50.94) and differ only in variance. On BCI-IV-2a and SHU-MI it ranks a close second, and on ISRUC it remains competitive with the top method. The strongest competing baseline throughout is random replay, which is known to be difficult to surpass. Notably, `WAVE` performs best on FACED and PhysioNet-MI, the two datasets with the most incremental subjects, suggesting that our distributional memory scales better as the sequence grows. Crucially, unlike random replay, which retains real EEG samples, our method stores only the synthetic data produced by memory compression, matching a strong baseline while avoiding the retention of privacy-sensitive raw signals.

Table 2: Continual-learning results on **BCI-IV-2a**. All methods share the same source model g_0 and report their best-*Score* configuration. Oracle is a non-adapting baseline, marked '-'. **Bold** marks the best among adapting methods (Oracle excluded); mean ± std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Source-only (g_0)	-	41.53 ± 2.43	-	36.10 ± 2.36	-	-	-	-	45.55 ± 1.58	45.29 ± 1.72	-
Oracle (joint training)	-	74.75 ± 1.09	-	74.72 ± 1.13	-	-	-	-	50.30 ± 1.23	49.11 ± 1.34	-
Naive fine-tuning	44.58 ± 1.41	**71.66 ± 0.62 (↑30.12)**	39.30 ± 2.14	**71.53 ± 0.34 (↑35.43)**	53.26 ± 2.04	18.40 ± 2.40	51.51 ± 2.34	20.02 ± 2.54	41.90 ± 0.73 (↓3.65)	36.69 ± 0.79 (↓8.61)	51.21 ± 0.71
Random Replay ($\lambda_r = 0.5, \lambda_c = 1$)	**47.59 ± 1.42**	70.05 ± 1.14 (↑28.51)	**42.98 ± 1.72**	69.84 ± 1.16 (↑33.74)	**66.66 ± 1.76**	**3.39 ± 1.33**	**66.83 ± 1.85**	**3.01 ± 1.21**	46.19 ± 1.10 (↑0.64)	45.33 ± 1.08 (↑0.03)	**58.66 ± 1.25**
EWC ($\lambda = 10$)	42.27 ± 0.75	70.05 ± 0.82 (↑28.51)	36.43 ± 1.63	69.56 ± 0.90 (↑33.46)	53.75 ± 1.77	16.30 ± 1.73	52.03 ± 2.24	17.54 ± 2.39	42.39 ± 1.34 (↓3.16)	37.11 ± 1.74 (↓8.19)	51.32 ± 1.45
LwF ($w = 1$)	43.75 ± 1.67	68.13 ± 0.15 (↑26.60)	37.56 ± 1.99	66.92 ± 0.26 (↑30.82)	56.73 ± 1.30	11.41 ± 1.41	56.14 ± 1.92	10.78 ± 2.08	45.66 ± 1.26 (↑0.10)	42.50 ± 1.65 (↓2.79)	54.31 ± 1.02
WAVE	46.69 ± 1.58	70.75 ± 1.06 (↑29.21)	41.75 ± 1.71	70.49 ± 1.22 (↑34.39)	64.39 ± 1.17	6.36 ± 1.05	64.45 ± 1.31	6.04 ± 1.01	**46.63 ± 0.98 (↑1.08)**	**45.54 ± 0.92 (↑0.24)**	58.38 ± 0.78

We conduct ablations along four axes: (i) the distributional geometry used to compare and compress subject measures, (ii) the replay-memory budget (B, m), (iii) the loss weights $\lambda_c, \lambda_r, \lambda_a$ that balance the current-subject, replay, and augmentation terms, and (iv) the order in which subjects arrive during the incremental phase. We present (i) and (ii) here; (iii) and (iv) are reported in Appendix G.

Ablation I: Distributional Geometry We compare the proposed WoW geometry with a measure-level formulation based on the Sinkhorn distance, which condenses past subjects into a single dis-

Table 3: Continual-learning results on **PhysioNet-MI**.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_t}	g_{N_t}	Score
Source-only (g_0)	-	40.93 ± 0.93	-	37.11 ± 0.08	-	-	-	-	39.09 ± 0.72	39.13 ± 0.68	-
Oracle	-	63.99 ± 1.04	-	64.17 ± 1.16	-	-	-	-	55.29 ± 1.17	55.26 ± 1.21	-
Naive fine-tuning	46.21 ± 0.46	55.73 ± 1.60 (↑14.80)	43.57 ± 0.87	54.65 ± 1.66 (↑17.54)	53.93 ± 1.92	1.80 ± 1.15	53.98 ± 2.07	0.66 ± 0.80	44.62 ± 1.11 (↑5.53)	44.22 ± 1.09 (↑5.09)	50.70 ± 1.32
Random Replay ($\lambda_c = 1, \lambda_r = 0.5$)	46.96 ± 0.90	55.12 ± 1.57 (↑14.19)	**44.86 ± 0.71**	53.95 ± 1.60 (↑16.84)	**54.52 ± 0.66**	**0.60 ± 1.46**	**54.68 ± 0.79**	**-0.73 ± 1.12**	**45.04 ± 1.14 (↑5.95)**	**44.95 ± 1.13 (↑5.82)**	**50.94 ± 0.85**
EWC ($\lambda = 10$)	45.91 ± 0.87	55.51 ± 1.48 (↑14.57)	43.24 ± 0.76	54.25 ± 1.77 (↑17.14)	53.62 ± 1.29	1.89 ± 1.24	53.73 ± 1.37	0.52 ± 1.13	44.43 ± 1.03 (↑5.34)	44.05 ± 1.02 (↑4.93)	50.45 ± 1.03
LwF ($w = 0.5$)	46.61 ± 0.36	55.47 ± 1.24 (↑14.54)	44.09 ± 0.67	54.44 ± 1.74 (↑17.32)	54.41 ± 1.27	1.06 ± 0.89	54.32 ± 1.40	0.11 ± 1.06	44.91 ± 1.15 (↑5.82)	44.58 ± 1.14 (↑5.46)	50.89 ± 1.03
WAVE	**47.14 ± 0.82**	**57.01 ± 1.75 (↑16.07)**	44.23 ± 0.10	**55.98 ± 1.10 (↑18.87)**	53.49 ± 1.37	3.52 ± 1.56	53.36 ± 1.42	2.63 ± 1.15	44.73 ± 1.01 (↑5.64)	44.30 ± 1.00 (↑5.18)	**50.94 ± 0.98**

Table 4: Continual-learning results on **SHU-MI**.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_t}	g_{N_t}	Score
Source-only (g_0)	-	57.57 ± 1.22	-	56.86 ± 1.10	-	-	-	-	57.73 ± 0.94	57.53 ± 0.79	-
Oracle	-	63.73 ± 1.30	-	63.91 ± 1.46	-	-	-	-	55.14 ± 1.02	55.10 ± 1.03	-
Naive fine-tuning	56.94 ± 0.30	**65.10 ± 1.33 (↑7.53)**	56.39 ± 0.61	64.88 ± 1.39 (↑8.02)	58.73 ± 3.08	6.37 ± 3.30	58.28 ± 3.73	6.60 ± 3.88	55.95 ± 1.29 (↓1.78)	55.78 ± 1.35 (↓1.75)	59.51 ± 1.39
Random Replay ($\lambda_c = \lambda_r = 0.5$)	**58.31 ± 0.99**	63.98 ± 1.30 (↑6.41)	**58.02 ± 1.41**	63.86 ± 1.35 (↑7.00)	**63.87 ± 0.85**	**0.12 ± 0.92**	**63.85 ± 0.87**	**0.01 ± 0.91**	57.21 ± 1.37 (↓0.52)	57.05 ± 1.39 (↓0.48)	**61.46 ± 1.09**
EWC ($\lambda = 4000$)	58.15 ± 1.18	63.85 ± 1.44 (↑6.28)	57.74 ± 1.60	63.60 ± 1.59 (↑6.75)	61.23 ± 1.54	2.62 ± 1.01	61.04 ± 1.48	2.56 ± 0.76	55.77 ± 1.20 (↓1.96)	55.65 ± 1.21 (↓1.88)	60.00 ± 1.33
LwF ($w = 5$)	57.46 ± 1.11	62.68 ± 0.74 (↑5.11)	56.95 ± 1.41	62.51 ± 1.12 (↑5.66)	60.88 ± 2.42	1.80 ± 2.05	60.85 ± 2.49	1.67 ± 1.89	**57.51 ± 1.19 (↓0.22)**	**57.41 ± 1.23 (↓0.12)**	60.22 ± 1.22
WAVE	57.40 ± 0.96	**65.10 ± 1.37 (↑7.53)**	56.81 ± 1.27	**64.89 ± 1.39 (↑8.04)**	60.21 ± 1.03	4.89 ± 0.51	59.81 ± 0.66	5.08 ± 0.99	57.20 ± 1.05 (↓0.53)	56.83 ± 1.09 (↓0.71)	60.49 ± 0.85

Table 5: Continual-learning results on **FACED**.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_t}	g_{N_t}	Score
Source-only (g_0)	-	27.12 ± 1.54	-	27.84 ± 1.78	-	-	-	-	26.69 ± 0.89	26.16 ± 0.95	-
Oracle	-	72.05 ± 1.29	-	71.83 ± 1.45	-	-	-	-	56.25 ± 1.01	55.96 ± 1.03	-
Naive fine-tuning	41.79 ± 1.06	44.66 ± 0.99 (↑17.54)	45.24 ± 1.74	48.81 ± 1.47 (↑20.97)	55.00 ± 2.71	-10.34 ± 1.73	54.43 ± 2.93	-5.62 ± 1.57	36.14 ± 0.84 (↑9.45)	35.71 ± 0.87 (↑9.55)	44.41 ± 1.40
Random Replay ($\lambda_c = \lambda_r = 0.5$)	42.54 ± 1.26	45.14 ± 0.86 (↑18.02)	46.18 ± 1.98	49.58 ± 1.19 (↑21.74)	58.66 ± 2.47	**-13.52 ± 1.92**	58.21 ± 2.63	**-8.62 ± 2.22**	37.92 ± 0.90 (↑11.23)	37.41 ± 0.91 (↑11.24)	46.27 ± 1.19
EWC ($\lambda = 10$)	41.52 ± 1.26	44.06 ± 1.18 (↑16.95)	45.01 ± 1.79	48.51 ± 1.91 (↑20.67)	54.05 ± 1.76	-9.99 ± 0.59	53.37 ± 1.92	-4.86 ± 0.52	36.06 ± 0.82 (↑9.37)	35.61 ± 0.83 (↑9.45)	44.01 ± 1.25
LwF ($w = 1$)	41.78 ± 1.21	44.67 ± 0.58 (↑17.55)	45.07 ± 1.69	48.66 ± 1.16 (↑20.82)	56.27 ± 1.96	-11.60 ± 1.44	55.69 ± 2.29	-7.03 ± 1.81	36.36 ± 0.82 (↑9.67)	36.04 ± 0.88 (↑9.88)	44.80 ± 1.10
WAVE	**43.31 ± 1.83**	**45.78 ± 1.70 (↑18.66)**	**47.25 ± 2.22**	**50.57 ± 1.60 (↑22.73)**	**59.30 ± 1.70**	**-13.52 ± 0.51**	**58.62 ± 1.92**	-8.05 ± 0.83	**37.98 ± 0.76 (↑11.29)**	**37.50 ± 0.82 (↑11.34)**	**46.69 ± 1.26**

Table 6: Continual-learning results on **ISRUC**.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_t}	g_{N_t}	Score
Source-only (g_0)	-	60.39 ± 0.53	-	60.79 ± 0.88	-	-	-	-	74.91 ± 0.21	77.42 ± 0.12	-
Oracle	-	79.72 ± 0.29	-	82.60 ± 0.38	-	-	-	-	76.63 ± 0.14	79.49 ± 0.23	-
Naive fine-tuning	60.83 ± 0.23	75.31 ± 0.34 (↑14.92)	64.53 ± 0.37	81.61 ± 0.34 (↑20.81)	57.74 ± 0.21	17.57 ± 0.36	60.37 ± 0.28	21.23 ± 0.25	65.80 ± 0.38 (↓9.11)	67.43 ± 0.38 (↓9.98)	67.12 ± 0.23
Random Replay ($\lambda_c = \lambda_r = 0.5$)	**68.36 ± 0.64**	**75.92 ± 0.38 (↑15.52)**	**73.28 ± 0.47**	81.62 ± 0.59 (↑20.83)	**75.22 ± 0.22**	**0.69 ± 0.29**	**78.45 ± 0.32**	**3.17 ± 0.48**	**74.22 ± 0.14 (↓0.68)**	**77.08 ± 0.12 (↓0.34)**	**77.06 ± 0.23**
EWC ($\lambda = 4000$)	64.07 ± 0.45	75.50 ± 0.31 (↑15.10)	68.31 ± 0.18	81.31 ± 0.66 (↑20.51)	68.57 ± 1.41	6.92 ± 1.39	70.64 ± 1.48	10.67 ± 1.44	69.27 ± 0.58 (↓5.63)	70.89 ± 0.49 (↓6.53)	72.46 ± 0.53
LwF ($w = 1$)	62.27 ± 0.34	75.51 ± 0.21 (↑15.12)	66.02 ± 0.34	81.22 ± 0.34 (↑20.43)	58.50 ± 0.95	17.02 ± 1.03	61.11 ± 0.72	20.11 ± 1.02	67.99 ± 0.30 (↓6.91)	69.75 ± 0.26 (↓7.66)	68.17 ± 0.36
WAVE	66.38 ± 0.92	75.89 ± 0.36 (↑15.50)	71.07 ± 0.56	**81.76 ± 0.41 (↑20.97)**	68.82 ± 1.07	7.07 ± 0.89	71.67 ± 0.97	10.09 ± 1.22	72.48 ± 0.53 (↓2.43)	75.42 ± 0.43 (↓1.99)	74.17 ± 0.52

tilled memory, a distribution-matching dataset-distillation baseline (Zhao & Bilen, 2023; Rosasco et al., 2021). This ablation evaluates whether the hierarchical geometry over subjects and signals is important for continual adaptation. Both variants use the same debiased Sinkhorn ground distance and differ only in how the memory is structured; details are in Appendix D.

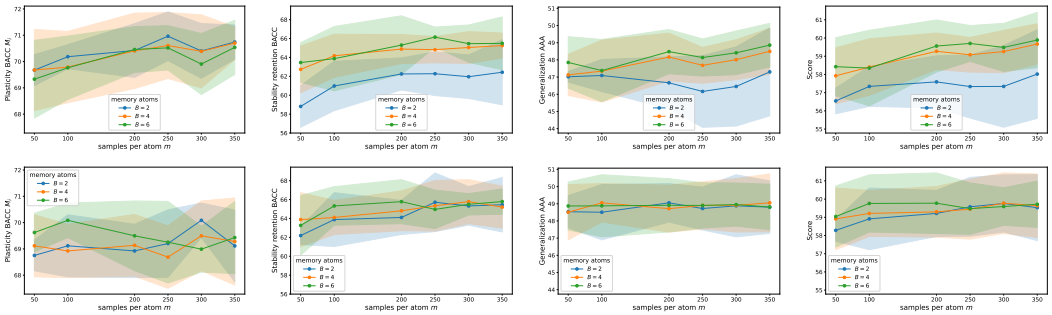
For a fair comparison, we sweep hyperparameters per dataset (Table 11) and report the best configuration. Table 7 shows that the Sinkhorn variant is outperformed by our framework on four of the five datasets, the exception being PhysioNet-MI. Plasticity is comparable between the two, as expected: adaptation to the current subject is driven by the current-subject loss term, in which the memory is not involved, so the choice of memory geometry does not affect it. The gap instead appears on stability and generalization, which rely more on the memory. This is consistent with the design principle discussed above: the flat Sinkhorn memory uses soft labels and a single joint transport plan with no per-class isolation, exactly the treatment we deliberately avoid for the compressed memory. Because distant classes are free to mix within the pooled atom, its stored measures drift from the real data; the worse stability of the Sinkhorn variant confirms this, indicating that the pooled memory represents previous subjects poorly and is more prone to CF. Generalization also degrades: augmentation is still applied, but the flat memory behaves as a single subject, so each virtual subject is interpolated from only one neighbor rather than the B available in our framework, yielding far less diverse augmentation. The exception is PhysioNet-MI, whose subjects are likely more similar, so there is little subject structure to keep and a single distilled memory works just as well.

Ablation II: Replay-Memory Size For BCI-IV-2a, we vary the replay-memory budget the number of stored subjects B and samples per subject m. This ablation quantifies the trade-off between memory size and CL performance, and also compares our framework against the strong random

Table 7: Continual-learning results for **Sinkhorn** on all the datasets, mean ± std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Dataset	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_t}	g_{N_t}	Score
BCI-IV-2a	45.48 ± 1.30	70.66 ± 0.40 (↑29.12)	40.21 ± 1.53	70.33 ± 0.31 (↑34.23)	56.08 ± 2.43	14.58 ± 2.39	54.74 ± 2.61	15.60 ± 2.54	43.79 ± 1.09 (↓1.76)	39.85 ± 1.26 (↓5.45)	53.39 ± 0.96
PhysioNet-MI	47.36 ± 1.33	56.44 ± 2.11 (↑15.50)	45.61 ± 0.96	55.60 ± 1.42 (↑18.49)	54.73 ± 0.90	1.71 ± 2.35	54.77 ± 0.97	0.83 ± 1.49	45.66 ± 1.26 (↑6.58)	45.37 ± 1.29 (↑6.24)	51.64 ± 1.11
SHU-MI	57.60 ± 0.58	64.74 ± 1.42 (↑7.16)	57.13 ± 0.29	64.56 ± 1.47 (↑7.71)	59.03 ± 2.93	5.71 ± 3.00	58.48 ± 3.69	6.09 ± 3.70	56.71 ± 1.19 (↓1.03)	56.52 ± 1.27 (↓1.02)	59.78 ± 1.49
FACED	42.05 ± 1.21	44.78 ± 1.70 (↑17.66)	45.68 ± 1.77	48.82 ± 2.44 (↑20.98)	55.96 ± 2.31	-11.18 ± 1.12	55.30 ± 2.42	-6.48 ± 1.15	36.77 ± 0.84 (↑10.08)	36.40 ± 0.84 (↑10.24)	44.96 ± 1.52
ISRUC	63.93 ± 0.43	75.55 ± 0.25 (↑15.15)	68.84 ± 0.30	81.39 ± 0.46 (↑20.59)	67.30 ± 0.15	8.24 ± 0.14	71.40 ± 0.24	9.99 ± 0.29	69.09 ± 0.28 (↓5.81)	71.75 ± 0.19 (↓5.67)	72.52 ± 0.17

replay baseline across buffer sizes. Figure 4 reports balanced accuracy for plasticity, stability, and generalization, together with the aggregate score, as a function of memory size for both methods. Plasticity is nearly flat for both methods, since it is driven by the current-subject loss term and does



(a) Plasticity (b) Stability (c) Generalization (d) Score

Figure 4: Memory-size effect B, m on BCI-IV-2a. Top: our method, bottom: random replay.

not depend on the memory. For our method, generalization improves with the number of subjects B: a memory with more subjects offers better neighbors for k-NN retrieval, whereas at $B = 2$ both stored subjects are always selected, so interpolation cannot exploit proximity. Generalization also improves with m, as more samples per stored subject yield richer virtual subjects. Stability follows the same trend: larger B and m both improve retention of previous subjects, so the aggregate score increases with memory size overall. For random replay, only the product $B \times m$ matters: its domain-balanced buffer has no per-subject constraint, so performance depends on the total sample count rather than on B and m separately. Comparing the two methods, our approach outperforms random replay on plasticity, the two are comparable on stability, and random replay is slightly ahead on generalization and the aggregate score.

4 Conclusion

In this paper, we introduced `WAVE`, a method for domain generalization in the sequential setting of EEG decoding across subjects. We formulate cross-subject EEG decoding as domain-incremental learning, treating each subject as a domain. Our contributions are twofold. First, we propose a new way of constructing the replay buffer for continual learning that stores synthetic samples summarizing all past subjects, rather than randomly resampling real data (Section 2.1). Second, we generate virtual subjects by interpolating between the stored subjects and the incoming subject, yielding smoother adaptation and better generalization to unseen subjects (Section 2.2). Empirically, across the five datasets, `WAVE` performs best on the two with the most incremental subjects (FACED and PhysioNet-MI) and remains a close second to random replay on the others, with rankings largely stable across subject arrival orders. Crucially, unlike random replay, our method attains these results while storing only synthetic samples, which is appealing for privacy-sensitive data such as EEG. Our approach has some limitations. `WAVE` requires labeled samples from the incremental subjects, and while storing optimized synthetic atoms is motivated by privacy, we do not formally quantify the privacy it provides in this paper. These open directions motivate future work: extending `WAVE` to the unsupervised setting where incremental subjects are unlabeled, and formally studying the privacy of the synthetic EEG produced by our compression mechanism.

AI USE STATEMENT

We disclose the use of LLMs in accordance with the ICLR 2027 policies. We used LLMs only to polish the writing (grammar, phrasing, and clarity), editing, and formatting of the manuscript. They were not used for the core methodology, scientific contributions, or originality of the research. All AI-assisted text retained in the paper was reviewed and edited by the authors, who take full responsibility for the final content.

ETHICS STATEMENT

Our work adheres to the ICLR Code of Ethics. This work does not involve the collection of new human-subject data. Our experiments rely on publicly available datasets: BCI-IV-2a (Brunner et al., 2008), PhysioNet-MI (Goldberger et al., 2000), SHU-MI (Ma et al., 2022), FACED (Chen et al., 2023) and ISRUC (Khalighi et al., 2016), which contain electroencephalography recordings, and we follow the data usage and access conditions associated with each of them. We refer readers to each dataset reference for more information about the ethics behind the studies conducted. We discuss in Appendix H potential impacts of our method.

REPRODUCIBILITY STATEMENT

We provide enough theoretical details for the faithful reproduction of our method. The problem formulation, memory-compression procedure, and virtual-subject augmentation are described in Section 2, with the full continual training loop in Algorithm 1. The pretrained model is specified in the paper; the data splits are described in Appendix E, and the training hyperparameters and search grids in Appendix F (Tables 10 and 11).

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Amin, Ghadir Altuwaijri, Wadood Abdul, Mohamed Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35, 08 2021. doi: 10.1007/s00521-021-06352-5.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

David Alvarez-Melis and Nicolò Fusi. Dataset dynamics via gradient flows in probability space. In *International conference on machine learning*, pp. 219–230. PMLR, 2021.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Eden Belouadah, Adrian Popescu, and Ioannis Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with wasserstein over wasserstein gradient flows, 2025. URL https://arxiv.org/abs/2506.07534.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):1, 2008.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline, 2020. URL https://arxiv.org/abs/2004.07211.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet Kumar Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. Continual learning with tiny episodic memories. *CoRR*, abs/1902.10486, 2019. URL http://arxiv.org/abs/1902.10486.

Jingjing Chen, Xiaobin Wong, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10, 10 2023. doi: 10.1038/s41597-023-02650-w.

Zhiyuan Chen and Bing Liu. *Lifelong machine learning*. Springer Nature, 2017.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *CoRR*, abs/1507.00504, 2015. URL http://arxiv.org/abs/1507.00504.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Didar Dadebayev, Wei Wei Goh, and Ee Xion Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University - Computer and Information Sciences*, 34(7):4385–4401, 2022. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2021.03.009. URL https://www.sciencedirect.com/science/article/pii/S1319157821000732.

Mourad El Hamri, Issam Falih, and Yves Rozenholc. Mccann's interpolation for gradual domain adaptation on the wasserstein geodesic. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2025. doi: 10.1109/IJCNN64981.2025.11228688.

Eduardo Fernandes Montesuma, Yassir Bendou, and Mike Gartrell. Wasserstein gradient flows for scalable and regularized barycenter computation. In Emilija Perković and Daniel Malinsky (eds.), *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4595–4606. PMLR, 17–21 Aug 2026. URL https://proceedings.mlr.press/v337/fernandes-montesuma26a.html.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International conference on artificial intelligence and statistics*, pp. 1608–1617. PMLR, 2018.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Xiaotong Gu, Zehong Cao, Alireza Jolfaei, Peng Xu, Dongrui Wu, Tzyy-Ping Jung, and Chin-Teng Lin. Eeg-based brain-computer interfaces (bcis): A survey of recent studies on signal sensing technologies and computational intelligence approaches and their applications. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(5):1645–1666, 2021.

Wei-Bang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/47393e8594c82ce8fd83adc672cf9872-Paper-Conference.pdf.

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.

James Kirkpatrick, Razvan Pascanu, Neil C. 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. *CoRR*, abs/1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

Dan Li, Hye-Bin Shin, and Yeon-Woo Choi. Prototype-guided non-exemplar continual learning for cross-subject eeg decoding, 2026a. URL https://arxiv.org/abs/2511.20696.

Taida Li, Yujun Yan, Fei Dou, Wenzhan Song, and Xiang Zhang. Cross-subject generalization for eeg decoding: a survey of deep learning methods. *Progress in Biomedical Engineering*, 8(2): 022013, May 2026b. ISSN 2516-1091. doi: 10.1088/2516-1091/ae65f0. URL http://dx.doi.org/10.1088/2516-1091/ae65f0.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *CoRR*, abs/1606.09282, 2016. URL http://arxiv.org/abs/1606.09282.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1): 153–179, 1997. doi: 10.1006/aima.1997.1634.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in psychology*, 4:504, 08 2013. doi: 10.3389/fpsyg.2013.00504.

XuanLong Nguyen. Borrowing strengh in hierarchical bayes: Posterior concentration of the dirichlet base measure. *Bernoulli*, 22(3), 2016. ISSN 1350-7265. doi: 10.3150/15-bej703. URL http://dx.doi.org/10.3150/15-BEJ703.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

Gabriel Peyré. Optimal transport for machine learners. *CoRR*, abs/2505.06589, 2025. URL https://arxiv.org/abs/2505.06589.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Now Publishers, 2019a.

Gabriel Peyré and Marco Cuturi. *Computational optimal transport*, volume 11. Now Publishers Inc, January 2019b. doi: 10.1561/2200000073. Publisher Copyright: © 2019 Now Publishers Inc. All Rights Reserved.

Ameya Prabhu, Philip H. S. Torr, and Puneet K. Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, pp. 524–540, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58535-8. doi: 10.1007/978-3-030-58536-5_31. URL https://doi.org/10.1007/978-3-030-58536-5_31.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Distilled replay: Overcoming forgetting through synthetic samples. In *International Workshop on Continual Semi-Supervised Learning*, pp. 104–117. Springer, 2021.

Yannick Roy, Hubert Banville, Isabela Maria Carneiro de Albuquerque, Alexandre Gramfort, Tiago Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16, 08 2019. doi: 10.1088/1741-2552/ab260c.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Simanto Saha and Mathias Baumert. Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Frontiers in Computational Neuroscience*, 13:87, 2019. doi: 10.3389/fncom.2019.00087. URL https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00087.

Qingshan She, Tie Chen, Feng Fang, Jianhai Zhang, Youmei Gao, and Yingchun Zhang. Improved domain adaptation network based on wasserstein distance for motor imagery eeg classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1137–1148, 02 2023. doi: 10.1109/TNSRE.2023.3241846.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation, 2018. URL https://arxiv.org/abs/1707.01217.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.

Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning. *CoRR*, abs/1904.07734, 2019. URL http://arxiv.org/abs/1904.07734.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2008.

Jeffrey Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11:37–57, 03 1985. doi: 10.1145/3147.3165.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding, 2025. URL https://arxiv.org/abs/2412.07236.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020a. URL https://arxiv.org/abs/1811.10959.

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3622–3626. IEEE, 2020b.

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup, 2019. URL https://arxiv.org/abs/1912.01805.

Fan Yang, Xu Tong, Anping Zeng, Chunlin He, Yi Liu, Chunhua Ren, and Ran Tian. Drift-aware experience replay for continual eeg-based disease recognition. *Neurocomputing*, 699:134388, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2026.134388. URL https://www.sciencedirect.com/science/article/pii/S0925231226017868.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6503–6512, 2023. doi: 10.1109/WACV56688.2023.00645.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. 2021. URL https://arxiv.org/abs/2006.05929.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024. doi: 10.1109/TPAMI.2024.3429383.

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, 2025b. URL https://arxiv.org/abs/2509.17439.

A Introduction

In this appendix, we provide supplementary material on the background, methodology, and the evaluation of our method, WAVE. It is organized as follows. Section B reviews the related literature and the optimal-transport tools used in our method. Section C illustrates the distributional shift across subjects on the BCI-IV-2a dataset, through topographies of the subject-averaged Pearson correlation coefficients. Section D details the memory compression using the Sinkhorn divergence on the joint space of features and soft labels. Section E describes the data split and the evaluation metrics used in our experiments. Section F reports the training hyperparameters and the search grids used for tuning each baseline and our method. Section G presents additional ablation studies; the loss-weight sweep for our framework and the subject arrival-order sweep. Section H discusses both the positive and negative societal impacts of our work. Finally, Section I extends the stability analysis to the source subjects across all the experiments above.

B Background

B.1 Continual learning

Continual Learning (CL), also called incremental or lifelong learning Chen & Liu (2017); Wang et al. (2024), is a machine-learning paradigm that studies systems which receive data sequentially and must acquire new knowledge while retaining what they have previously learned. Unlike conventional offline learning, where the training data are typically assumed to be independent and identically distributed and available simultaneously, CL considers a non-stationary data stream whose underlying distribution may change over time. This setting is particularly relevant for real-world machine-learning systems, where data are collected continuously and retraining a model from scratch on all historical data may be computationally expensive, impractical or infeasible when the data are confidential and can not be stored for privacy reasons.

A main challenge in CL is Catastrophic Forgetting (CF) McCloskey & Cohen (1989), whereby adapting to newly observed data can substantially degrade performance on previously learned domains or tasks. In the sequential setting, let $\mu_1, \mu_2, \ldots, \mu_N$ denote a sequence of datasets that arrive one after another. A model with parameters θ is initially trained on μ_1 and subsequently updated on $\mu_2, \ldots, \mu_N$. Standard empirical risk minimization on only the currently available data,

$$\theta_t = \arg\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mu_t}\big[\ell(f_\theta(x), y)\big]$$

does not constrain the model to preserve its performance on earlier datasets. As a result, the updated model θ_t may drift away from the solutions of $\mu_1, \ldots, \mu_{t-1}$, degrading sharply on them.

CL therefore seeks to balance plasticity, i.e., the ability to acquire information from new data, and stability, i.e., the ability to preserve previously acquired knowledge. This stability-plasticity trade-off is one of the fundamental challenges of continual learning Mermillod et al. (2013)

B.1.1 Continual learning Scenarios

Continual learning problems are commonly organized into three scenarios: task-, domain-, and class-incremental learning, distinguished by whether the task identity is available at test time and, if not, whether it must be inferred van de Ven & Tolias (2019); Wang et al. (2024).

In *task-incremental learning* (TIL), the model encounters a sequence of distinct tasks, typically with different objectives or label spaces, and the task identity is known at inference time. Consequently, the model can use task-specific information, such as a dedicated output head, to determine which prediction problem to solve. In *class-incremental learning* (CIL), new classes are introduced over time, and the model must learn to recognize them while retaining the ability to discriminate among all classes observed so far, including classes never seen together. Crucially, the task identity is unavailable at inference time and must be inferred, which makes CIL substantially more challenging than TIL Belouadah et al. (2021); Zhou et al. (2024). In *domain-incremental learning* (DIL), the prediction problem and label space remain unchanged while the input distribution shifts over time. Formally, the learner observes a sequence of domains $\mu_1, \mu_2, \ldots, \mu_T$ with $\mu_i(X) \neq \mu_j(X)$ for $i \neq j$, so the objective is to learn a single mapping $f : \mathcal{X} \to \mathcal{Y}$ that stays effective across domains.

As in CIL, the task identity is unavailable at test time; unlike TIL, the model maintains no separate task-specific predictors, and unlike CIL, the set of target classes does not grow over time van de Ven & Tolias (2019).

This distinction is particularly relevant for machine learning models learned on EEG data, where signals exhibit substantial variability across recording conditions. In many EEG applications the prediction objective is fixed, for example, classifying motor-imagery or emotions, yet the distribution of the observed signals changes across subjects, sessions, electrode configurations, hardware, and experimental environments. Such changes can be interpreted as domain shifts, and therefore naturally motivate a domain-incremental learning formulation. In our case, we focus on subject variability as the source of domain shift, treating each subject as a distinct domain: the model adapts to subjects arriving sequentially while preserving performance on those seen previously and generalizing to unseen ones.

B.1.2 Continual learning approaches

To mitigate Catastrophic Forgetting (CF), several families of approaches have been proposed Wang et al. (2024): regularization-based, replay-based, optimization-based, representation-based, and architecture-based approaches, though many modern methods combine several mechanisms.

Regularization-based methods. These add an explicit penalty to balance old and new domains, and divide into two main sub-directions Wang et al. (2024). *Weight regularization* constrains changes to parameters according to their importance for previous domains. Elastic Weight Consolidation (EWC) Kirkpatrick et al. (2016) estimates this importance through the Fisher information matrix and penalizes deviations from previously learned values,

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{new}}(\theta) + \frac{\lambda}{2} \sum_k F_k \left(\theta_k - \theta^*_{k,\text{old}}\right)^2, \tag{9}$$

where $\theta^*_{k,\text{old}}$ is the previously learned value of parameter k and F_k its estimated importance; related methods such as Synaptic Intelligence Zenke et al. (2017) differ mainly in how this importance is measured. *Function regularization* instead targets the model's outputs, using the previous model as a teacher via knowledge distillation. Learning without Forgetting (LwF) Li & Hoiem (2016) combines the task loss with a distillation term that keeps the current model's predictions close to those of the frozen previous model,

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}}\big(f_\theta(x), y\big) - \sum_c f_c\big(x; \theta^*\big)\, \log f_c\big(x; \theta\big), \tag{10}$$

where $f(\cdot; \theta^*)$ is the frozen previous model. Regularization methods avoid storing historical examples, but under substantial domain shift these constraints alone may be insufficient to recover information about previous distributions.

Replay-based methods. Replay methods retain or regenerate data representative of previous domains and replay it jointly with current-domain data, letting the model optimize for adaptation and retention simultaneously Wang et al. (2024). Given a memory $\mathbb{M}$ of past examples, the objective combines the current-domain loss with a replay loss over the buffer,

$$\mathcal{L}(\theta) = \mathbb{E}_{(x,y)\sim\mu_t}\big[\ell(f_\theta(x), y)\big] + \lambda\, \mathbb{E}_{(x,y)\sim\mathbb{M}}\big[\ell(f_\theta(x), y)\big], \tag{11}$$

where the first term drives adaptation to the current domain and the second preserves performance on previous ones. Experience Replay stores a limited buffer of past examples that are periodically replayed Chaudhry et al. (2019), and simple strategies can be strikingly effective: GDumb Prabhu et al. (2020) greedily balances a fixed-size memory and retrains from it, while Dark Experience Replay Buzzega et al. (2020) additionally matches the logits of past samples to mitigate forgetting. Random reservoir sampling Vitter (1985) offers a simple mechanism for maintaining a fixed-size memory as new data arrive. Because they provide direct evidence from previous domains rather than relying solely on parameter constraints, replay methods are often particularly effective.

Other approaches. Beyond these two families, Wang et al. (2024) identify three further categories. *Optimization-based* methods manipulate the optimization process directly, for example

through gradient projection or loss-landscape shaping. *Representation-based* methods exploit self-supervised or pretrained representations, including continual pre-training, to obtain features that transfer better across tasks. *Architecture-based* methods dedicate distinct parameters to different domains or expand the network as new domains arrive, as in Progressive Neural Networks Rusu et al. (2016); these resist forgetting well but their cost grows with the number of domains, which may be impractical for long-term EEG settings with many subjects.

B.2 Optimal Transport

B.2.1 Wasserstein distance

Optimal Transport (OT) compares probability measures by quantifying the cost of transporting one into another. We briefly recall the main concepts, following standard references Villani (2008); Peyré & Cuturi (2019a); Peyré (2025). OT is a central tool in modern machine learning, used for distribution alignment and domain adaptation, generative modeling, and comparing complex data distributions Peyré & Cuturi (2019a). Its adoption at scale was enabled by entropic regularization and Sinkhorn's algorithm Cuturi (2013), which we review below.

Let $(\mathcal{X}, |\cdot|)$ be a metric space. In our experiments, $\mathcal{X} = \mathbb{R}^d$ with d the dimension of the flattened EEG, and $|x - x'|$ denotes the Euclidean distance on $\mathbb{R}^d$. Given two probability measures μ and ν on $\mathcal{X}$, and a ground cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the original formulation due to Monge is *map-based*:

$$T^\star \in \operatorname*{arg\,inf}_{T_\# \mu = \nu} \int_{\mathcal{X}} c\big(x, T(x)\big)\, \mathrm{d}\mu(x), \tag{12}$$

where $T : \mathcal{X} \to \mathcal{X}$ is a measurable transport map and the pushforward constraint $T_\#\mu = \nu$ means that T sends μ to ν, formally the pushforward $T_\#\mu$ is the measure defined by $(T_\#\mu)(A) = \mu\big(T^{-1}(A)\big)$ for every measurable set A, equivalently $\int g\, \mathrm{d}(T_\#\mu) = \int g \circ T\, \mathrm{d}\mu$ for all bounded measurable g. This formulation suffers from two problems. First, the constraint set may be empty (e.g., $\mu = \delta_0$, $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$). Second, the problem is non-convex, since the objective depends on T through the ground-cost c.

Kantorovich relaxed this problem by allowing *transport plans*, i.e., joint measures γ on $\mathcal{X} \times \mathcal{X}$ whose marginals are μ and ν, which may split mass:

$$\Gamma(\mu, \nu) = \big\{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \ : \ (\pi_1)_\#\gamma = \mu,\ (\pi_2)_\#\gamma = \nu\big\}. \tag{13}$$

with

$$\pi_1(x, y) = x; \quad \pi_2(x, y) = y \tag{14}$$

The Kantorovich OT problem is then

$$\gamma^\star = \operatorname*{arg\,inf}_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}^2} c(x, y)\, d\gamma(x, y). \tag{15}$$

The Kantorovich formulation has some advantages over the Monge formulation. First, its constraint set $\Gamma(\mu, \nu)$ is non-empty, as it always contains the trivial transport plan $\gamma = \mu \otimes \nu$. Second, its objective is linear with respect to the optimization variable γ, and its constraint set is linear as well. This second advantage defines the Kantorovich problem as a linear program.

In this paper, we assume access to samples from μ and ν, thus approximating these through empirical measures. If μ and ν are given by finite point clouds $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ with weights a_i and b_j, we can write

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{m} b_j \delta_{y_j}, \quad a_i \geq 0,\ b_j \geq 0,\ \sum_i a_i = \sum_j b_j = 1, \tag{16}$$

where δ_{x_i} (resp. y_j) denotes the Dirac measure centered at x_i. A discrete transport plan is then a nonnegative matrix $P \in \mathbb{R}_+^{n \times m}$ such that

$$P\mathbf{1}_m = a, \qquad P^\top \mathbf{1}_n = b, \tag{17}$$

and the discrete OT problem becomes

$$P^\star = \underset{P \in U(a,b)}{\arg\min} \langle P, C \rangle_F \text{ with } U(a,b) = \{P \geq 0 : P\mathbf{1}_m = a,\ P^\top \mathbf{1}_n = b\} \tag{18}$$

where $\langle A, B \rangle_F = A_{ij} B_{ij}$ is the Frobenius inner product. Here, $C_{ij} = c(x_i, y_j)$ denotes the ground-cost matrix. When $c(x, y) = \|x - y\|^p$, the corresponding OT cost defines the p-th power of the p-Wasserstein distance, $W_p(\mu, \nu)^p$. Since equation 18 is a linear program, its computation has complexity $\mathcal{O}(n^3 \log n)$ in the number of samples n Peyré & Cuturi (2019a).

B.2.2 Entropic Regularization and Sinkhorn divergence

An alternative to Kantorovich's formulation in equation 15 consists of adding an entropic regularization term,

$$\gamma_\epsilon^\star = \underset{\gamma \in \Gamma(\mu,\nu)}{\arg\inf} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\gamma(x, y) + \varepsilon \mathrm{KL}(\gamma \| \mu \otimes \nu),$$

where KL denotes the Kullback-Leibler divergence, and $\mu \otimes \nu$ denotes the trivial coupling. When $c(x, y) = |x - y|^p$ comes from a metric, this problem defines a *divergence*, known as the *Sinkhorn divergence*,

$$W_{p,\varepsilon}(\mu, \nu)^p = \underset{\gamma \in \Gamma(\mu,\nu)}{\arg\inf} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\gamma(x, y) + \varepsilon \mathrm{KL}(\gamma \| \mu \otimes \nu).$$

When μ and ν are empirical, Cuturi (2013) shows that the computation of the entropic OT plan,

$$P_\epsilon^\star = \underset{P \in U(a,b)}{\arg\min} \Big(\langle P, C \rangle_F + \varepsilon \,\mathrm{KL}(P \,\|\, ab^\top) \Big), \tag{19}$$

can be done via the iterative algorithm of Sinkhorn (1964), that is, Sinkhorn's algorithm. The problem in equation 19 is strictly convex, with a unique minimizer. The optimal coupling takes the multiplicative scaling form

$$P_\epsilon^\star = \mathrm{diag}(u)\, K\, \mathrm{diag}(v), \qquad K = \exp(-C/\varepsilon), \tag{20}$$

for scaling vectors u, v enforcing the marginals.

Sinkhorn iterations arise as alternating Bregman projections in the KL geometry: normalizing P with respect to μ projects onto the affine set $\{P : P\mathbf{1}_m = a\}$, and normalizing with respect to ν projects onto $\{P : P^\top \mathbf{1}_n = b\}$. Repeating these KL-projections converges to the entropically regularized optimum $P^\star$. The projection steps yield the Sinkhorn updates

$$u^{(t+1)} = \frac{a}{K v^{(t)}}, \qquad v^{(t+1)} = \frac{b}{K^\top u^{(t+1)}}, \tag{21}$$

and after L iterations we set $P_\epsilon^{(L)} = \mathrm{diag}(u^{(L)}) K \mathrm{diag}(v^{(L)})$. When $C_{ij} = |\cdot|^p$, the Sinkhorn divergence is $W_{p,\varepsilon}(\mu, \nu)^p = \langle P^{(L)}, C \rangle_F + \varepsilon \mathrm{KL}(P^{(L)} \| ab^\top)$.

A central observation in Genevay et al. (2018) is that the entropic OT cost $W_{p,\varepsilon}(\mu, \nu)$ is a biased approximation of the unregularized OT distance. Because of the entropy term in equation 19, one can show that $W_{p,\varepsilon}(\mu, \mu) > 0$ as soon as c is not identically zero. Intuitively, even when the two measures are identical, the optimal plan $\gamma^\star$ is spread out by the entropy regularization and therefore pays a nonzero transport cost.

To remove this self-bias, Genevay et al. (2018) defines the *Sinkhorn divergence* as a debiased combination of three entropic OT terms:

$$S_{p,\varepsilon}(\mu, \nu)^p = 2W_{p,\varepsilon}(\mu, \nu)^p - W_{p,\varepsilon}(\mu, \mu)^p - W_{p,\varepsilon}(\nu, \nu)^p. \tag{22}$$

By construction, we have

$$S_{p,\varepsilon}(\mu, \mu)^p = 0 \quad \text{for all probability measures } \mu. \tag{23}$$

Moreover, $S_\varepsilon(\mu, \nu)$ is symmetric and finite for any pair of probability measures with finite p-moments. **Henceforth, we use $p = 2$.**

B.2.3 DISPLACEMENT INTERPOLATION

A distinctive feature of OT is its *lifting effect*. Indeed, given a metric space $(\mathcal{X}, |\cdot|)$, the Wasserstein distance W_2 defines a related metric space on the space of probability measures, $(\mathcal{P}_2(\mathcal{X}), W_2)$, where $\mathcal{P}_2$ is the space of probability measures with finite $2-$moments. This induces a Riemannian-like structure Otto (2001), in which geodesics between measures correspond to meaningful interpolations between probability measures. This notion has been formalized through **displacement interpolation**. For $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ with an OT map $T^\star$,

$$\mu_t = ((1-t)\,\mathrm{Id} + tT^\star)_\# \mu, \qquad t \in [0,1],$$

defines a geodesic in $(\mathcal{P}_2(\mathcal{X}), W_2)$. The resulting curve $(\mu_t)_{t\in[0,1]}$ defines a constant-speed geodesic between μ_0 and μ_1 in the Wasserstein space, satisfying

$$\mu = \mu_{t=0}, \qquad \nu = \mu_{t=1}.$$

This interpolation provides a principled way of constructing intermediate measures between two domains. In our setting, it is used to generate synthetic intermediate subjects, enhancing the generalization capabilities of our model.

Concretely, given the two point clouds we compute the optimal plan $\gamma^\star$ by solving the discrete OT problem between μ and ν (Eq. 15). Since $\gamma^\star$ may split the mass of a source point across several targets, it does not directly yield a map; we recover an approximate transport map through its *barycentric projection* Ambrosio et al. (2005), which assigns to each source point x_i the $\gamma^\star$-weighted average of the target points it is coupled to:

$$T_{\gamma^\star}(x_i) = \sum_j \bar{\gamma}^\star_{ij}\, y_j, \qquad \bar{\gamma}^\star_{ij} = \frac{\gamma^\star_{ij}}{\sum_{j'} \gamma^\star_{ij'}}.$$

Applying the displacement interpolation to this map, the interpolated cloud at parameter t is

$$\tilde{x}_i = (1-t)\, x_i + t\, T(x_i),$$

whose empirical measure $\tilde{\mu}_t = \frac{1}{n}\sum_i \delta_{\tilde{x}_i}$ is a barycentric-projection approximation of the McCann interpolant μ_t.

B.2.4 WASSERSTEIN OVER WASSERSTEIN DISTANCE

The Wasserstein distance W_2 is a metric on the space $\mathcal{P}_2(\mathcal{X})$, whose elements are probability measures over points in $\mathcal{X}$. However, in many applications, we are interested in comparing measures of measures. For example, a dataset of EEG recordings from a given subject can be viewed as a collection of class-conditional measures, so that its empirical measure can be regarded as an element of $\mathcal{P}_2(\mathcal{P}_2(\mathcal{X}))$. In this setting, a natural approach is to compare such measures using the Wasserstein distance W_2 as the ground metric between their realizations. This gives rise to the Wasserstein-over-Wasserstein (WoW) framework Nguyen (2016); Bonet et al. (2025), which equips the space of measures over probability measures with a Wasserstein geometry induced by the geometry of $\mathcal{P}_2(\mathcal{X})$. In this appendix, we provide a background on the WoW space and the associated gradient flows on this space, introduced by Bonet et al. (2025).

Wasserstein-over-Wasserstein (WoW) distance The WoW distance Bonet et al. (2025); Nguyen (2016), also known as the hierarchical optimal-transport distance, lifts the squared Wasserstein distance W_2^2 on $\mathcal{P}_2(\mathbb{R}^d)$ to a ground cost and defines an optimal-transport distance between measures over measures: for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$,

$$W^2_{W_2}(\mathbb{P}, \mathbb{Q}) = \inf_{\Gamma \in \Pi(\mathbb{P},\mathbb{Q})} \int W_2^2(\mu, \nu)\, \mathrm{d}\Gamma(\mu, \nu). \tag{24}$$

Here, the transport plan Γ couples measures rather than points, and the cost of moving μ to ν is $W_2^2(\mu, \nu)$. This defines a genuine distance on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ Nguyen (2016).

Riemannian structure and WoW gradient On the metric space $\big(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)), W_{W_2}\big)$, Bonet et al. (2025) develop a corresponding differential structure analogous to Otto's calculus on the Wasserstein space. For a functional $F : \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) \to \mathbb{R}$, its WoW gradient $\nabla_{W_{W_2}} F(\mathbb{P})$ assigns

to $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ a tangent vector,$\nabla_{W_{W_2}} F(\mathbb{P})(\mu) \in T_\mu \mathcal{P}_2(\mathbb{R}^d)$. It is characterized by the first-order expansion, for $\mathbb{Q}$ and $\Gamma \in \exp_{\mathbb{P}}^{-1}(\mathbb{Q})$,

$$F(\mathbb{Q}) = F(\mathbb{P}) + \iint \left\langle \nabla_{W_{W_2}} F(\mathbb{P})\big(\pi^{\mathbb{R}^d}_{\#}\gamma\big)(x), v \right\rangle_x \mathrm{d}\gamma(x,v)\,\mathrm{d}\Gamma(\gamma) + o(W_{W_2}(\mathbb{P},\mathbb{Q}))\,, \tag{25}$$

under the differentiability assumptions Bonet et al. (2025).

Bonet et al. (2025) construct the corresponding WoW gradient flow through the limit of an implicit Jordan-Kinderlehrer-Otto (JKO) scheme. In particular, the discrete scheme is

$$\mathbb{P}_{k+1} \in \operatorname*{argmin}_{\mathbb{P}} \left\{ \frac{1}{2\tau} W^2_{W_2}(\mathbb{P},\mathbb{P}_k) + F(\mathbb{P}) \right\}. \tag{26}$$

Under suitable assumptions on F, including lower semicontinuity, coercivity, properness, and λ-convexity along generalized geodesics, Bonet et al. (2025) establish convergence of the JKO scheme to the corresponding gradient flow.

Forward discretization As an explicit alternative to the implicit JKO scheme, the corresponding Riemannian gradient-descent step can be written formally as

$$\mathbb{P}_{k+1} = \exp_{\mathbb{P}_k}\big(-\tau\,\nabla_{W_{W_2}} F(\mathbb{P}_k)\big), \tag{27}$$

with step size $\tau > 0$. At the level of individual distribution-valued particles, this gives

$$\mu_c^{k+1} = \exp_{\mu_c^k}\big(-\tau\,\nabla_{W_{W_2}} F(\mathbb{P}_k)(\mu_c^k)\big)\,. \tag{28}$$

For the Euclidean base space $\mathbb{R}^d$, the exponential map is $\exp_x(v) = x + v$. Thus, for discrete measures of the form $\mathbb{P}_k = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu_c^k}$, $\quad \mu_c^k = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^{c,k}}$, the update acts on the individual support points as

$$x_i^{c,k+1} = x_i^{c,k} - \tau\,\nabla_{W_{W_2}} F(\mathbb{P}_k)(\mu_c^k)(x_i^{c,k}). \tag{29}$$

This update couples the particles at two levels: an *intra-atom* level through the dependence on the point cloud defining μ_c^k, and an *inter-atom* level through the dependence on the full outer measure $\mathbb{P}_k$ Bonet et al. (2025).

MMD functionals and kernels A tractable class of objectives on $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ is given by Maximum Mean Discrepancy (MMD) functionals Gretton et al. (2006); Bonet et al. (2025). Given a kernel $K : \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and a fixed target $\mathbb{Q}$, let

$$F(\mathbb{P}) = \frac{1}{2}\,\mathrm{MMD}^2_K(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\iint K(\mu,\nu)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\mu)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\nu). \tag{30}$$

If $K_\nu : \mu \mapsto K(\mu,\nu)$ is Wasserstein differentiable, the WoW gradient is

$$\nabla_{W_{W_2}} F(\mathbb{P})(\mu) = \int \nabla_{W_2} K_\nu(\mu)\,\mathrm{d}(\mathbb{P}-\mathbb{Q})(\nu). \tag{31}$$

Bonet et al. (2025) consider kernels based on the Sliced-Wasserstein distance. In particular, the Gaussian Sliced-Wasserstein kernel is given by

$$K(\mu,\nu) = \exp\left(-\frac{\mathrm{SW}_2^2(\mu,\nu)}{2h}\right), \tag{32}$$

where $h > 0$ is a bandwidth parameter. They also consider the Riesz Sliced-Wasserstein kernel

$$K(\mu,\nu) = -\mathrm{SW}_2(\mu,\nu). \tag{33}$$

The latter is not positive definite and does not require a bandwidth parameter; Bonet et al. (2025) report strong empirical performance for this kernel.

For discrete outer measures $\mathbb{P}_x = \frac{1}{C}\sum_{c=1}^{C} \delta_{\mu_x^c}$, $\quad \mu_x^c = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i^c}$, the resulting objective can be viewed as a finite-dimensional function of the stacked point cloud x. Under the assumptions given by Bonet et al. (2025), the WoW gradient is related to the Euclidean gradient by

$$\nabla_{W_{W_2}} F(\mathbb{P}_x)(\mu_x^c)(x_i^c) = Cn\,\nabla_{x_i^c} F(x). \tag{34}$$

Consequently, the finite-dimensional Euclidean gradient can be used to compute the WoW gradient, with the corresponding normalization factor taken into account.

C Cross-Subject Distribution Shift

To provide empirical evidence of cross-subject distribution shift, we examine how the relationship between spectral features and motor-imagery class varies across subjects in BCI Competition IV-2a. For each subject, we compute a per-electrode correlation between relative band power and class membership, and render it as a scalp topography for each class and frequency band (Figure 5). Concretely, each trial is transformed to the frequency domain with a real FFT, and the periodogram is integrated over five bands (δ: 1-4, θ: 4-8, α: 8-14, β: 14-30, γ: 30-47 Hz). Dividing each band by the total broadband power yields a per-trial, per-electrode *relative band power*. For each of the four classes, we compute the correlation R between relative band power and a binary class label (that class versus the rest) at every electrode and band, across the subject's trials.

The resulting topographies are markedly subject-specific: the sign, magnitude, and spatial location of the discriminative correlations differ across subjects, and even the band carrying the most class information is itself inconsistent. This variability shows that the mapping from spectral features to class is not shared across subjects, motivating modeling at the subject level rather than assuming a common representation.

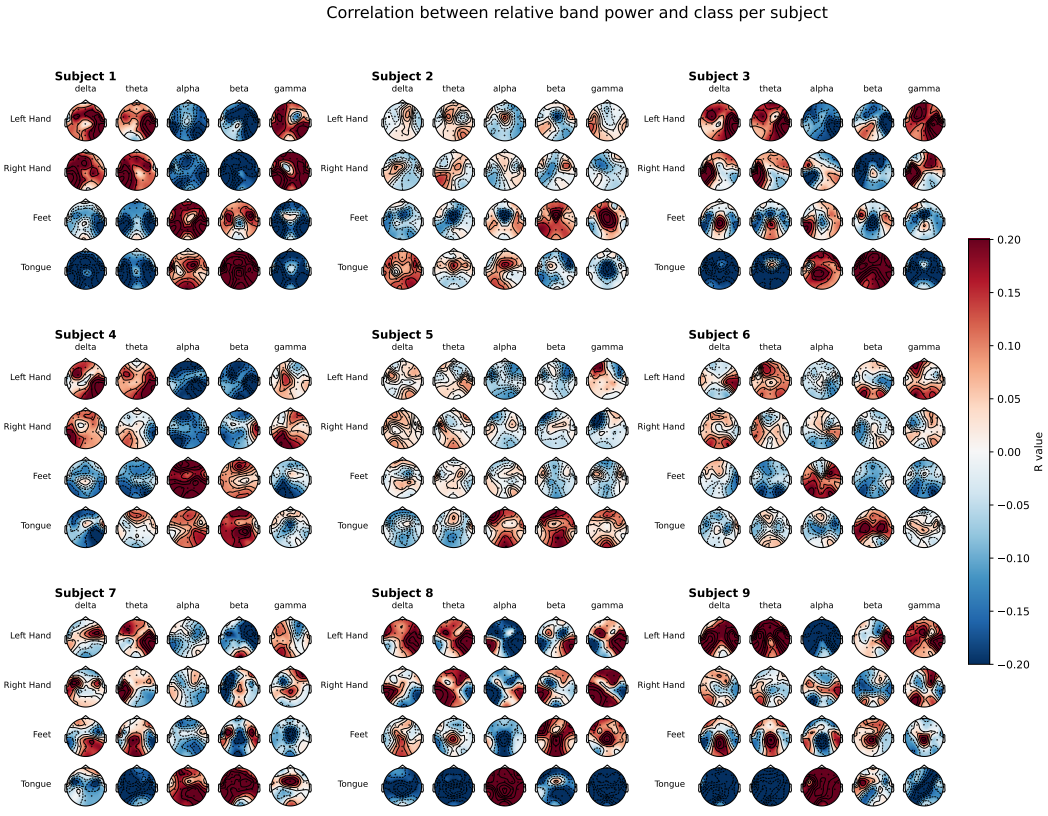


Figure 5: **Subject-specific feature-label relationships on BCI-IV-2a.** Per-electrode correlation R between relative band power and class (one-vs-rest), as scalp topographies per class (rows) and frequency band (columns) for each subject (color scale ± 0.2). The sign, magnitude, and location of the discriminative correlations vary across subjects, showing that the feature-to-class mapping is not shared across subjects.

D SINKHORN MEMORIES

In this section, we consider memories as probability measures over the space $\mathbb{R}^d \times \Delta_C$, meaning, feature $\times$ soft-label joint measures,

$$\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{(\mathbf{x}_i, \mathbf{y}_i)}, \tag{35}$$

where n denotes the number of samples in the support of μ_s. The main difference with respect to our proposed formulation is that Sinkhorn memories are *plain measures*, meaning, they do not have the hierarchical structure we define in equation 1.

Memory Initialization. In this case, we have subject level measures $\{\mu_s\}_{s=1}^{N_S}$ and we need to create an initial measure ν_0. We do so via a debiased Sinkhorn barycenter,

$$\nu_0 = \underset{\nu \in \text{Emp}_n(\mathbb{R}^d \times \Delta_C)}{\arg\min} \frac{1}{N_S}\sum_{s=1}^{N_S} S_\epsilon(\mu_s, \nu),$$

which is a minimization over the support of ν_0. This is done, for instance, with the gradient flow algorithm of Fernandes Montesuma et al. (2026).

Memory Update. For each incoming subject μ_{N_S+t}, we update the memory via,

$$\nu_t = \underset{\nu \in \text{Emp}_n(\mathbb{R}^d \times \Delta_C)}{\arg\min} \beta S_\epsilon(\mu_{N_S+t}, \nu)^2 + (1-\beta) S_\epsilon(\nu_{t-1}, \nu),$$

which approximates the Wasserstein geodesic between the memory ν_{t-1} and the new incoming measure μ_{N_S+t}. $\beta \in [0, 1]$, here, denotes the forgetting rate.

For the sake of comparability, we use $n = B \times m$ samples in the memory, so the WoW memory and the Sinkhorn memory have the same number of samples. The debiased Sinkhorn divergence is fit directly over the pooled (embedding, label) point cloud, with labels encoded as logits on a softmax-parametrized simplex and combined additively with the embedding coordinates into a single product ground metric, so that one optimal-transport plan trades off moving embeddings against relabeling. There is no notion of a subject and no per-class isolation: every replay point competes with every other point within one joint OT problem, and nothing structurally prevents an embedding from drifting toward another class's target if doing so lowers the joint cost.

E EVALUATION

E.1 DATASET SPLITS

Table 8 summarizes the subject-level partition. The source domains $\mathcal{D}_S$ are split into a training subset $\mathcal{D}_S^{\text{train}}$ for supervised pretraining and a held-out test subset $\mathcal{D}_S^{\text{test}}$ (20% of the source samples). Each incremental subject $\mathcal{D}_t$, $t \in \{1, \ldots, N_I\}$, is likewise split into an adaptation subset and a held-out test subset $\mathcal{D}_{\text{test}}^{(t)}$ (20% of the subject's samples). Generalization subjects $\mathcal{D}_G$ are never seen during adaptation and are used only for evaluation.

Table 8: Subject-level dataset partition.

Set	Subjects	Role	Within-subject split
Source subjects	$\sim 20\%$	Source pretraining	train / test (80/20)
Incremental subjects	$\sim 60\%$	Sequential adaptation	adaptation / test (80/20)
Generalization subjects	$\sim 20\%$	Generalization (unseen)	test only

E.2 METRICS

E.2.1 PLASTICITY

For each subject t, we evaluate the pretrained model g_0, the model before adaptation g_{t-1}, and the adapted model g_t on $\mathcal{D}_{\text{test}}^{(t)}$. For $m \in \{\text{BACC}, \text{MF1}\}$,

$$m_0^{(t)} = m(g_0, \mathcal{D}_{\text{test}}^{(t)}), \qquad m_{\text{prev}}^{(t)} = m(g_{t-1}, \mathcal{D}_{\text{test}}^{(t)}), \qquad m_{\text{c}}^{(t)} = m(g_t, \mathcal{D}_{\text{test}}^{(t)}), \tag{36}$$

with averages over the N_I incremental subjects

$$\overline{m}_0 = \frac{1}{N_I}\sum_{t=1}^{N_I} m_0^{(t)}, \qquad \overline{m}_{\text{prev}} = \frac{1}{N_I}\sum_{t=1}^{N_I} m_{\text{prev}}^{(t)}, \qquad \overline{m}_{\text{c}} = \frac{1}{N_I}\sum_{t=1}^{N_I} m_{\text{c}}^{(t)}. \tag{37}$$

In the results table we report the before/after pair $\overline{m}_{\text{prev}}$ and $\overline{m}_{\text{c}}$ as the columns g_{t-1} and g_t.

E.2.2 STABILITY

Stability measures how well the model retains performance on subjects it has already learned, both the source subjects seen during pretraining and the incremental subjects seen earlier in the stream. For the source subjects, we track performance on $\mathcal{D}_S^{\text{test}}$ relative to the pretrained model,

$$s^{(t)}(m) = m(g_t, \mathcal{D}_S^{\text{test}}), \qquad s^{(0)}(m) = m(g_0, \mathcal{D}_S^{\text{test}}), \tag{38}$$

and report the average retained source performance and the source forgetting at the final step N_I,

$$\text{Ret}_S(m) = s^{(N_I)}(m), \qquad \text{Fgt}_S(m) = s^{(0)}(m) - s^{(N_I)}(m). \tag{39}$$

For the incremental subjects, we re-evaluate the final model g_{N_I} on the held-out test set of every previously seen subject and compare against its performance right after that subject was learned,

$$\text{Ret}_I(m) = \frac{1}{N_I}\sum_{t=1}^{N_I} m(g_{N_I}, \mathcal{D}_{\text{test}}^{(t)}), \qquad \text{Fgt}_I(m) = \frac{1}{N_I}\sum_{t=1}^{N_I} \big(m_{\text{c}}^{(t)} - m(g_{N_I}, \mathcal{D}_{\text{test}}^{(t)})\big). \tag{40}$$

Here Ret denotes average retained performance and Fgt average forgetting (lower is better; negative values indicate positive backward transfer). The Retention and Forgetting columns in the results table report Ret_I and Fgt_I. We reported the stability over source subjects in Appendix I.

E.2.3 GENERALIZATION

After each adaptation step t, we evaluate g_t on the fixed unseen-subject set $\mathcal{D}_G$:

$$g^{(t)}(m) = m(g_t, \mathcal{D}_G), \qquad g^{(0)}(m) = m(g_0, \mathcal{D}_G). \tag{41}$$

Following the anytime evaluation protocol of BrainUICL Zhou et al. (2025a), we summarize generalization by the average anytime score and report its gain over the pretrained model,

$$\text{AA}_{N_I}(m) = \frac{1}{N_I}\sum_{t=1}^{N_I} g^{(t)}(m), \qquad \Delta_{\text{gen}}(m) = \text{AA}_{N_I}(m) - g^{(0)}(m). \tag{42}$$

We denote $\text{AA}_{N_I}(\text{BACC})$ by AAA and $\text{AA}_{N_I}(\text{MF1})$ by AAF1 in the results table.

Source-only and Oracle each correspond to a single, fixed model, g_0 for source-only and the jointly-trained model for Oracle, rather than a sequence of per-task checkpoints $g_1, \ldots, g_{N_I}$. Because Stability (Retention/Forgetting) is defined over consecutive updates and because the Plasticity gain is defined relative to g_0, neither is applicable to these two baselines and both are marked "$-$". For these rows we instead report, under Plasticity, the single model's mean balanced accuracy / MF1 evaluated once over the incremental (continual) subjects' test sets, placed in the g_t slot with g_{t-1} left as "$-$"; and under Generalization (g_{N_I}), its mean balanced accuracy / MF1 over the held-out generalization subjects.

E.3 COMPUTATIONAL COST

Compared with random replay, the main extra cost of WAVE is the memory compression (Section 2.1). We solve it with a WoW gradient flow of N_{flow} steps, where each step runs Sinkhorn for N_{sink} iterations (both fixed; see Table 10). Each flow step compares $\mathcal{O}(B^2)$ pairs of atoms. The kernel never mixes classes, so each pair splits into C small per-class Sinkhorn problems. With balanced classes, one pair costs $\mathcal{O}\big(\frac{m^2}{C}(d + N_{\text{sink}})\big)$, which is roughly C times cheaper than one joint $m \times m$ problem. Compressing one new subject therefore costs

$$\mathcal{O}\Big(N_{\text{flow}}\, B^2\, \tfrac{m^2}{C}\, (d + N_{\text{sink}})\Big). \tag{43}$$

The augmentation is much cheaper: it finds the K nearest atoms using class means only, then solves K exact transport problems. The memory holds at most B atoms of m samples, so the cost per subject stops growing once the memory is full, and the memory is never larger than the random-replay buffer. WAVE therefore costs extra *compute*, not extra *storage*. Table 9 reports the runtime of WAVE against random replay.

Table 9: Runtime of WAVE vs. random replay, averaged over runs. Memory update is the average time of one WAVE UPDATEMEMORY call (under 0.1 s for random replay).

Dataset	NVIDIA GPU	WAVE	Random replay	Memory update (s)
BCI-IV-2a	L4	11 min	7 min	38
PhysioNet-MI	L4	2 h 18 min	1 h 09 min	50
SHU-MI	L4	35 min	21 min	39
FACED	L4	3 h 36 min	2 h 20 min	58
ISRUC	L40S	3 h 25 min	2 h 22 min	79

F HYPERPARAMETER SETTINGS

Table 10 lists the training and memory-compression settings, and Table 11 the search grids explored for each method. Loss weights refer to the classification, replay, and augmentation terms $\lambda_c, \lambda_r, \lambda_a$.

G ADDITIONAL ABLATION STUDIES

G.1 LOSS WEIGHTS

We study the effect of the weighting coefficients λ_{c}, λ_{r}, and λ_{a}. This analysis evaluates the relative importance of each loss term to the final performance. In particular, for a subset of configurations we set $\lambda_{\text{a}} = 0$ (removing the virtual-subject, augmentation-based term) to isolate the contribution of data augmentation. Tables 12, 13, 14, 15 and 16 report this sweep on BCI-IV-2a, PhysioNet-MI, SHU-MI, FACED and ISRUC respectively.

The effect of the virtual-subject term is dataset-dependent. On FACED and BCI-IV-2a, the virtual-subject term is the decisive factor for cross-subject generalization. On BCI-IV-2a it is especially clear: without it, generalization (AAA/AAF1) falls below the source-only baseline g_0, whereas enabling it lifts both back above the baseline, at the cost of only slightly higher forgetting; the best overall score is obtained by our setting $\lambda_{\text{c}} = 0.5,\ \lambda_{\text{r}} = \lambda_{\text{a}} = 0.25$ (58.38). FACED shows the same pattern more mildly, with the virtual term improving generalization and plasticity and yielding the lowest forgetting. On PhysioNet-MI it gives the best generalization and Score, but only by a small margin within the run-to-run variance, so the gain is minor. On SHU-MI it makes little difference, as all configurations fall within each other's variance. On ISRUC the term instead hurts, so the best configuration disables it. The best configuration per dataset, identified in this ablation, is the one reported in the main results (Tables 2, 3, 4, 5 and 6). Overall, the virtual-subject term mainly affects generalization and its benefit is dataset-dependent, while λ_{c} and λ_{r} primarily trade plasticity against stability.

Table 10: Training and memory-compression settings.

Group	Setting	Value
Common	Optimizer	AdamW
	Scheduler	Cosine (min LR 10^{-8})
	Weight decay	5×10^{-2}
	Gradient clipping	1.0
	Batch size	64
	Label smoothing	0.1
Backbone	Encoder	CBraMod
	Layers	12
	Hidden dimension D	200
	Heads	8
Classifier h_ϕ	MLP dimensions	$D \to D \to C$
	Activation / dropout	ELU / 0.1
Source phase	Learning rate	10^{-4}
	Epochs	50
Incremental phase	Learning rate	10^{-5}
	Epochs (per step)	30
Memory budget (B, m)	BCI-IV	$B = 4,\ m = 200$
	FACED	$B = 40,\ m = 50$
	ISRUC	$B = 30,\ m = 300$
	PhysioNet-MI	$B = 30,\ m = 50$
	SHU-MI	$B = 8,\ m = 160$
number of virtual subjects (K)	BCI-IV	$K = 2$
	FACED	$K = 10$
	ISRUC	$K = 10$
	PhysioNet-MI	$K = 10$
	SHU-MI	$K = 4$
Flow	Step size τ	100 (BCI-IV), 300 (FACED & ISRUC), 50 (Physio& SHU)
	Flow steps	200
	Momentum	0.9
	Kernel	Riesz
Sinkhorn solver	Entropic reg. ε	10^{-3}
	Max iterations	200
	Tolerance	10^{-6}
Protocol	Seeds	$\{0,\ 42,\ 123,\ 2026\}$

Table 11: Hyperparameter search grids per method. For our method, the four configurations of $(\lambda_c, \lambda_r, \lambda_a)$ are evaluated.

Method	Search grid
EWC	$\lambda_{\text{EWC}} \in \{10,\ 100,\ 400,\ 1000,\ 4000\}$
LwF	$\lambda_{\text{LwF}} \in \{0.1,\ 0.5,\ 1,\ 5,\ 10\}$
Random replay	$\lambda_c \in \{0.5,\ 1\},\ \lambda_r = 0.5$
WAVE and Sinkhorn ablation	$\lambda_c = 0.5,\ \lambda_r = 0.5,\ \lambda_a = 0$
	$\lambda_c = 1,\ \ \lambda_r = 0.5,\ \lambda_a = 0$
	$\lambda_c = 0.5,\ \lambda_r = 0.25,\ \lambda_a = 0.25$
	$\lambda_c = 0.5,\ \lambda_r = 0.5,\ \lambda_a = 0.5$

G.2 Sensitivity to Subject Arrival Order

Since the order in which subjects arrive is uncontrolled in real-world applications, a continual-learning pipeline should be robust to it. To assess this, we randomly perturb the arrival order of the incremental subjects and repeat the continual-learning process for our framework and the continual-learning baselines: random replay, naive fine-tuning, EWC, and LwF, over four permutations of the

Table 12: Sweep over the loss weights $\lambda_c, \lambda_r, \lambda_a$ on **BCI-IV-2a**, mean $\pm$ std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_f}	g_{N_f}	Score
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	**47.16 ± 1.65**	70.56 ± 0.78 (↑29.02)	41.87 ± 2.42	70.30 ± 0.70 (↑34.20)	65.01 ± 1.33	5.55 ± 1.29	65.12 ± 1.47	5.18 ± 1.27	45.40 ± 0.81 (↓0.15)	43.72 ± 0.87 (↓1.58)	57.67 ± 0.91
$\lambda_c = \lambda_r = 0.5$ (no virtual)	46.93 ± 1.73	69.96 ± 0.50 (↑28.42)	41.59 ± 2.71	69.64 ± 0.36 (↑33.54)	**65.83 ± 1.80**	**4.13 ± 1.77**	**65.92 ± 1.86**	**3.71 ± 1.81**	45.08 ± 0.30 (↓0.47)	43.36 ± 0.43 (↓1.94)	57.55 ± 0.61
$\lambda_c = \lambda_r = \lambda_a = 0.5$	47.04 ± 1.61	69.72 ± 1.19 (↑28.18)	**42.48 ± 1.61**	69.37 ± 1.41 (↑33.27)	64.91 ± 1.52	4.81 ± 0.88	64.90 ± 1.64	4.47 ± 1.02	46.49 ± 1.01 (↑0.94)	45.47 ± 0.94 (↑0.18)	58.20 ± 0.53
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	46.69 ± 1.58	**70.75 ± 1.06 (↑29.21)**	41.75 ± 1.71	**70.49 ± 1.22 (↑34.39)**	64.39 ± 1.17	6.36 ± 1.05	64.45 ± 1.31	6.04 ± 1.01	**46.63 ± 0.98 (↑1.08)**	**45.54 ± 0.92 (↑0.24)**	**58.38 ± 0.78**

Table 13: Sweep over the loss weights $\lambda_c, \lambda_r, \lambda_a$ on **PhysioNet-MI**, mean $\pm$ std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_f}	g_{N_f}	Score
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	**47.36 ± 1.14**	56.65 ± 2.12 (↑15.72)	**44.44 ± 0.88**	55.72 ± 1.55 (↑18.61)	53.60 ± 1.21	**3.05 ± 1.49**	53.46 ± 1.18	**2.26 ± 1.04**	44.47 ± 0.89 (↑5.38)	44.07 ± 0.87 (↑4.94)	50.78 ± 1.16
$\lambda_c = \lambda_r = 0.5$ (no virtual)	46.28 ± 0.99	56.85 ± 2.05 (↑15.91)	43.16 ± 0.74	55.81 ± 1.44 (↑18.70)	**53.65 ± 1.00**	3.20 ± 1.35	**53.49 ± 0.96**	2.32 ± 0.74	44.17 ± 0.84 (↑5.08)	43.73 ± 0.82 (↑4.61)	50.70 ± 1.07
$\lambda_c = \lambda_r = \lambda_a = 0.5$	46.24 ± 1.60	**57.36 ± 2.43 (↑16.43)**	42.91 ± 1.49	**56.18 ± 1.99 (↑19.07)**	53.37 ± 1.02	3.99 ± 1.95	53.27 ± 1.21	2.91 ± 1.41	44.23 ± 0.87 (↑5.14)	43.81 ± 0.84 (↑4.68)	50.77 ± 1.19
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	47.14 ± 0.82	57.01 ± 1.75 (↑16.07)	44.23 ± 0.10	55.98 ± 1.10 (↑18.87)	53.49 ± 1.37	3.52 ± 1.56	53.36 ± 1.42	2.63 ± 1.15	**44.73 ± 1.01 (↑5.64)**	**44.30 ± 1.00 (↑5.18)**	**50.94 ± 0.98**

Table 14: Sweep over the loss weights $\lambda_c, \lambda_r, \lambda_a$ on **SHU-MI**, mean $\pm$ std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_f}	g_{N_f}	Score
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	57.64 ± 0.66	65.07 ± 1.54 (↑7.50)	57.13 ± 0.97	64.87 ± 1.57 (↑8.01)	59.51 ± 1.21	5.57 ± 1.04	59.15 ± 0.92	5.72 ± 1.43	57.08 ± 1.09 (↓0.65)	56.80 ± 1.14 (↓0.74)	60.22 ± 1.02
$\lambda_c = \lambda_r = 0.5$ (no virtual) (WAVE)	57.40 ± 0.96	65.10 ± 1.37 (↑7.53)	56.81 ± 1.27	64.89 ± 1.39 (↑8.04)	**60.21 ± 1.03**	**4.89 ± 0.51**	**59.81 ± 0.66**	**5.08 ± 0.99**	**57.20 ± 1.05 (↓0.53)**	**56.83 ± 1.09 (↓0.71)**	**60.49 ± 0.85**
$\lambda_c = \lambda_r = \lambda_a = 0.5$	**58.52 ± 0.70**	**65.30 ± 0.32 (↑7.73)**	**57.63 ± 1.05**	**64.96 ± 0.46 (↑8.10)**	59.60 ± 1.25	5.70 ± 1.01	59.35 ± 0.79	5.61 ± 0.55	56.35 ± 1.60 (↓1.38)	55.68 ± 2.21 (↓1.85)	59.96 ± 0.95
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$	57.91 ± 0.81	64.90 ± 0.37 (↑7.33)	57.21 ± 0.94	64.64 ± 0.35 (↑7.78)	59.50 ± 0.95	5.41 ± 0.87	59.29 ± 0.66	5.35 ± 0.58	56.53 ± 1.66 (↓1.20)	55.99 ± 2.16 (↓1.55)	59.92 ± 0.77

Table 15: Sweep over the loss weights $\lambda_c, \lambda_r, \lambda_a$ on **FACED**, mean $\pm$ std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_f}	g_{N_f}	Score
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	42.80 ± 0.97	45.35 ± 1.29 (↑18.23)	46.44 ± 1.25	49.54 ± 1.56 (↑21.70)	58.44 ± 1.24	-13.10 ± 0.19	57.87 ± 1.48	-8.32 ± 0.48	37.53 ± 0.78 (↑10.84)	37.00 ± 0.75 (↑10.84)	46.05 ± 1.09
$\lambda_c = \lambda_r = 0.5$ (no virtual)	43.03 ± 1.12	44.97 ± 1.24 (↑17.85)	46.83 ± 1.32	49.28 ± 1.29 (↑21.43)	59.06 ± 1.40	**-14.10 ± 0.31**	58.45 ± 1.60	**-9.17 ± 0.67**	37.66 ± 0.73 (↑10.97)	37.13 ± 0.70 (↑10.96)	46.14 ± 1.05
$\lambda_c = \lambda_r = \lambda_a = 0.5$	**43.35 ± 2.18**	45.45 ± 1.34 (↑18.33)	**47.28 ± 2.53**	50.01 ± 1.02 (↑22.17)	**59.78 ± 1.77**	-14.33 ± 0.69	**59.14 ± 1.97**	-9.12 ± 1.02	**38.01 ± 0.75 (↑11.32)**	**37.52 ± 0.82 (↑11.35)**	46.66 ± 1.11
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	43.31 ± 1.83	**45.78 ± 1.70 (↑18.66)**	47.25 ± 2.22	**50.57 ± 1.60 (↑22.73)**	59.30 ± 1.70	-13.52 ± 0.51	58.62 ± 1.92	-8.05 ± 0.83	37.98 ± 0.76 (↑11.29)	37.50 ± 0.82 (↑11.34)	**46.69 ± 1.26**

Table 16: Sweep over the loss weights $\lambda_c, \lambda_r, \lambda_a$ on **ISRUC**, mean $\pm$ std over 4 seeds.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Configuration	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_f}	g_{N_f}	Score
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	66.10 ± 0.86	75.87 ± 0.37 (↑15.47)	71.01 ± 0.53	**81.80 ± 0.55 (↑21.01)**	68.74 ± 0.42	7.12 ± 0.71	71.59 ± 0.45	10.21 ± 0.80	72.41 ± 0.14 (↓2.50)	75.37 ± 0.29 (↓2.05)	74.12 ± 0.22
$\lambda_c = \lambda_r = 0.5$ (no virtual) (WAVE)	**66.38 ± 0.92**	**75.89 ± 0.36 (↑15.50)**	**71.07 ± 0.56**	81.76 ± 0.41 (↑20.97)	**68.82 ± 1.07**	**7.07 ± 0.89**	**71.67 ± 0.97**	**10.09 ± 1.22**	**72.48 ± 0.53 (↓2.43)**	**75.42 ± 0.43 (↓1.99)**	**74.17 ± 0.52**
$\lambda_c = \lambda_r = \lambda_a = 0.5$	64.45 ± 0.53	74.35 ± 0.39 (↑13.95)	68.38 ± 0.52	80.49 ± 0.50 (↑19.70)	60.86 ± 3.62	13.49 ± 3.80	63.47 ± 2.77	17.02 ± 2.93	71.23 ± 0.20 (↓3.68)	74.23 ± 0.21 (↓3.18)	70.13 ± 1.26
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$	64.55 ± 0.67	74.54 ± 0.37 (↑14.14)	68.50 ± 0.63	80.65 ± 0.53 (↑19.85)	60.77 ± 3.44	13.76 ± 3.38	63.48 ± 2.74	17.16 ± 2.70	71.21 ± 0.18 (↓3.69)	74.29 ± 0.18 (↓3.12)	70.17 ± 1.31

subject order. We use for each method the same set of hyperparameters used in Section 3. We use a fixed training seed to isolate the effect of subject order, and report the mean $\pm$ standard deviation across these runs in tables 17, 18, 19, 20 and 21.

Across all the datasets, the ranking of methods is largely preserved under permutation, confirming that our approach is robust to subject arrival order. On BCI-IV-2a, our method outperforms naive fine-tuning, EWC, and LwF on stability and generalization, and remains second only to random replay, which stays a hard baseline to beat. On PhysioNet-MI all methods remain very close (within about 0.5 Score). This reinforces our hypothesis that PhysioNet-MI subjects are relatively similar: when subjects differ little, the choice of continual-learning strategy matters less, consistent with the near-ties in the main results and the strong performance of the flat distilled memory in Ablation I, Section 3. On FACED, `WAVE` ranks first, reproducing the ordering of the main results (Table 5), with markedly smaller standard deviations than BCI-IV; this is expected given its far larger number of incremental subjects (80 vs. 5), over which permutation effects average out. On SHU-MI and ISRUC our method ranks second behind random replay, on par with or ahead of the other baselines on stability and generalization, again with small standard deviations, indicating limited sensitivity to order.

H Broader Impacts

The proposed method advances neural decoding of EEG signals across subjects in a continual setting, which is realistic for EEG applications where subjects arrive sequentially. Beyond improving generalization to unseen subjects, it facilitates the broader adoption of neural interfaces by enabling

Table 17: Subject-order ablation results on **BCI-IV-2a**, mean ± std over 4 subject-order permutations. Parenthesized gains are measured against this ablation's own source-only baseline g_0: BACC = 40.91, MF1 = 35.95, AAA = 46.79, AAF1 = 46.57.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Method	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Naive fine-tuning	42.38 ± 2.70	**70.45 ± 1.73 (↑29.54)**	37.65 ± 4.47	**70.02 ± 1.55 (↑34.07)**	56.27 ± 1.99	14.18 ± 3.37	54.38 ± 2.49	15.64 ± 3.63	42.49 ± 1.21 (↓4.30)	38.03 ± 2.30 (↓8.54)	52.43 ± 1.18
Random Replay ($\lambda_r = 0.5, \lambda_c = 1$)	**46.07 ± 2.47**	69.21 ± 2.45 (↑28.30)	**42.38 ± 2.15**	68.78 ± 2.23 (↑32.83)	**66.88 ± 0.73**	**2.33 ± 2.11**	**66.62 ± 0.68**	**2.16 ± 1.87**	**48.22 ± 1.40 (↑1.43)**	**47.72 ± 1.47 (↑1.16)**	**59.60 ± 1.17**
EWC ($\lambda = 10$)	41.67 ± 2.49	70.13 ± 1.07 (↑29.22)	37.00 ± 4.53	69.68 ± 0.80 (↑33.73)	55.70 ± 2.37	14.43 ± 2.82	53.74 ± 2.54	15.94 ± 2.77	42.38 ± 1.43 (↓4.41)	37.76 ± 2.76 (↓8.81)	52.07 ± 1.59
LwF ($w = 1$)	43.47 ± 2.83	68.11 ± 1.27 (↑27.20)	38.78 ± 4.79	67.42 ± 0.80 (↑31.47)	58.68 ± 2.67	9.43 ± 3.21	57.39 ± 2.94	10.03 ± 3.18	44.50 ± 1.11 (↓2.29)	41.30 ± 1.88 (↓5.26)	54.20 ± 0.95
WAVE	45.11 ± 1.93	69.62 ± 1.66 (↑28.71)	41.28 ± 2.94	69.30 ± 1.25 (↑33.35)	65.91 ± 2.89	3.72 ± 3.13	65.65 ± 2.93	3.65 ± 2.87	47.56 ± 1.10 (↑0.77)	46.55 ± 1.40 (↓0.02)	58.98 ± 1.54

Table 18: Subject-order ablation results on **PhysioNet-MI**, mean ± std over 4 subject-order permutations. Parenthesized gains are measured against this ablation's own source-only baseline g_0: BACC = 39.57, MF1 = 36.08, AAA = 39.11, AAF1 = 39.04.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Method	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Naive fine-tuning	46.76 ± 1.00	56.04 ± 1.52 (↑16.47)	43.80 ± 0.84	54.87 ± 1.34 (↑18.79)	53.33 ± 2.16	2.72 ± 1.98	53.08 ± 2.52	1.79 ± 1.70	44.85 ± 0.06 (↑5.74)	44.50 ± 0.10 (↑5.47)	50.65 ± 0.98
Random Replay ($\lambda_c = 1, \lambda_r = 0.5$)	**47.35 ± 0.79**	54.71 ± 1.97 (↑15.14)	**45.14 ± 0.55**	53.52 ± 1.79 (↑17.44)	**54.86 ± 0.93**	**-0.15 ± 1.11**	**54.94 ± 0.98**	**-1.42 ± 0.91**	45.07 ± 0.42 (↑5.95)	**44.99 ± 0.39 (↑5.95)**	**50.92 ± 0.80**
EWC ($\lambda = 10$)	46.55 ± 0.90	55.70 ± 1.62 (↑16.13)	43.70 ± 0.86	54.36 ± 1.46 (↑18.28)	52.98 ± 1.80	2.72 ± 1.71	52.76 ± 2.30	1.60 ± 1.57	44.68 ± 0.13 (↑5.57)	44.34 ± 0.16 (↑5.31)	50.36 ± 0.91
LwF ($w = 0.5$)	46.72 ± 0.77	56.06 ± 1.37 (↑16.49)	43.92 ± 0.64	54.98 ± 1.36 (↑18.91)	53.54 ± 1.80	2.52 ± 1.79	53.33 ± 2.07	1.66 ± 1.36	**45.13 ± 0.05 (↑6.01)**	44.86 ± 0.07 (↑5.82)	50.88 ± 0.84
WAVE	46.94 ± 0.53	**56.93 ± 1.76 (↑17.36)**	43.92 ± 0.95	**55.10 ± 1.27 (↑19.02)**	52.31 ± 2.07	4.62 ± 2.78	52.13 ± 2.31	2.96 ± 2.20	44.76 ± 0.24 (↑5.64)	44.49 ± 0.27 (↑5.45)	50.47 ± 0.83

Table 19: Subject-order ablation results on **SHU-MI**, mean ± std over 4 subject-order permutations. Parenthesized gains are measured against this ablation's own source-only baseline g_0: BACC = 56.81, MF1 = 55.72, AAA = 57.71, AAF1 = 57.63.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Method	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Naive fine-tuning	56.46 ± 0.81	**65.28 ± 0.90 (↑8.46)**	55.87 ± 1.30	**65.15 ± 1.02 (↑9.43)**	59.10 ± 1.00	6.18 ± 0.81	59.04 ± 1.03	6.11 ± 0.83	57.12 ± 0.55 (↓0.59)	57.00 ± 0.54 (↓0.63)	60.25 ± 0.75
Random Replay ($\lambda_c = 0.5$)	**57.77 ± 1.10**	64.64 ± 0.43 (↑7.82)	**57.31 ± 1.57**	64.55 ± 0.27 (↑8.83)	**64.45 ± 1.32**	**0.19 ± 1.63**	**64.45 ± 1.33**	**0.10 ± 1.43**	**59.53 ± 1.33 (↑1.82)**	**59.48 ± 1.32 (↑1.85)**	**62.75 ± 0.67**
EWC ($\lambda = 4000$)	57.05 ± 1.80	64.43 ± 1.18 (↑7.61)	56.54 ± 2.31	64.33 ± 1.04 (↑8.61)	60.15 ± 2.48	4.28 ± 2.41	60.06 ± 2.36	4.27 ± 2.14	57.53 ± 1.03 (↓0.17)	57.44 ± 1.06 (↓0.19)	60.50 ± 1.07
LwF ($w = 5$)	57.62 ± 1.12	63.10 ± 0.54 (↑6.29)	57.17 ± 1.24	63.01 ± 0.68 (↑7.29)	61.03 ± 2.21	2.07 ± 2.34	60.88 ± 2.27	2.13 ± 1.97	58.80 ± 0.77 (↑1.10)	58.76 ± 0.78 (↑1.13)	60.86 ± 0.86
WAVE	57.16 ± 1.37	64.89 ± 0.56 (↑8.08)	56.44 ± 1.69	64.70 ± 0.67 (↑8.98)	59.88 ± 0.66	5.02 ± 0.43	59.28 ± 0.40	5.42 ± 0.36	58.64 ± 0.75 (↑0.93)	58.48 ± 0.81 (↑0.85)	60.86 ± 0.61

Table 20: Subject-order ablation results on **FACED**, mean ± std over 4 subject-order permutations. Parenthesized gains are measured against this ablation's own source-only baseline g_0: BACC = 27.13, MF1 = 26.64, AAA = 25.56, AAF1 = 24.89.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Method	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Naive fine-tuning	39.42 ± 1.03	42.64 ± 0.87 (↑15.50)	42.31 ± 0.79	46.03 ± 1.01 (↑19.40)	54.27 ± 1.78	-11.64 ± 2.14	53.88 ± 2.06	-7.84 ± 2.76	35.06 ± 0.22 (↑9.50)	34.76 ± 0.27 (↑9.87)	43.01 ± 0.32
Random Replay ($\lambda_c = \lambda_r = 0.5$)	41.03 ± 0.25	44.04 ± 0.70 (↑16.90)	44.05 ± 0.70	47.51 ± 0.63 (↑20.87)	57.19 ± 1.02	**-13.16 ± 1.08**	56.66 ± 1.08	**-9.15 ± 1.46**	36.66 ± 0.24 (↑11.09)	36.32 ± 0.35 (↑11.43)	44.87 ± 0.31
EWC ($\lambda = 10$)	39.96 ± 0.66	43.02 ± 0.35 (↑15.88)	42.60 ± 0.49	46.58 ± 0.82 (↑19.94)	54.22 ± 1.39	-11.20 ± 1.29	53.77 ± 1.76	-7.20 ± 2.10	34.95 ± 0.15 (↑9.38)	34.63 ± 0.21 (↑9.74)	43.08 ± 0.40
LwF ($w = 1$)	40.18 ± 0.43	43.46 ± 0.38 (↑16.32)	42.71 ± 0.87	46.78 ± 0.95 (↑20.14)	55.84 ± 1.76	-12.38 ± 2.11	55.55 ± 2.18	-8.77 ± 2.76	35.49 ± 0.30 (↑9.93)	35.25 ± 0.42 (↑10.36)	43.84 ± 0.41
WAVE	**41.38 ± 0.55**	**44.58 ± 0.94 (↑17.45)**	**44.57 ± 0.66**	**48.21 ± 0.87 (↑21.57)**	**57.47 ± 0.78**	-12.89 ± 0.50	**56.81 ± 0.90**	-8.60 ± 0.83	**36.74 ± 0.26 (↑11.17)**	**36.42 ± 0.30 (↑11.53)**	**45.16 ± 0.30**

Table 21: Subject-order ablation results on **ISRUC**, mean ± std over 4 subject-order permutations. Parenthesized gains are measured against this ablation's own source-only baseline g_0: BACC = 61.07, MF1 = 61.95, AAA = 74.76, AAF1 = 77.31.

	Plasticity				Stability				Generalization		
	BACC		MF1		BACC		MF1		AAA	AAF1	
Method	g_{t-1}	g_t	g_{t-1}	g_t	Retention	Forgetting	Retention	Forgetting	g_{N_I}	g_{N_I}	Score
Naive fine-tuning	59.97 ± 0.84	75.94 ± 0.21 (↑14.87)	63.12 ± 1.53	**81.84 ± 0.21 (↑19.89)**	61.09 ± 2.48	14.84 ± 2.39	64.60 ± 3.11	17.24 ± 3.04	66.92 ± 0.91 (↓7.84)	69.24 ± 1.25 (↓8.07)	69.30 ± 1.53
Random Replay ($\lambda_c = \lambda_r = 0.5$)	**68.60 ± 0.22**	75.97 ± 0.30 (↑14.90)	**73.19 ± 0.49**	81.48 ± 0.38 (↑19.53)	**75.07 ± 0.41**	**0.90 ± 0.38**	**78.51 ± 0.33**	**2.97 ± 0.27**	**74.57 ± 0.27 (↓0.19)**	**77.48 ± 0.20 (↑0.17)**	**77.16 ± 0.24**
EWC ($\lambda = 4000$)	64.81 ± 0.84	76.07 ± 0.28 (↑15.00)	68.60 ± 0.74	81.53 ± 0.25 (↑19.59)	68.08 ± 1.00	7.99 ± 1.07	71.33 ± 0.67	10.20 ± 0.65	71.45 ± 1.64 (↓3.31)	73.58 ± 1.84 (↓3.73)	73.47 ± 0.63
LwF ($w = 1$)	61.56 ± 0.77	75.37 ± 0.25 (↑14.30)	64.81 ± 1.28	80.95 ± 0.34 (↑19.01)	61.94 ± 2.17	13.43 ± 2.36	65.34 ± 2.62	15.61 ± 2.81	69.02 ± 0.73 (↓5.74)	71.45 ± 1.10 (↓5.87)	70.16 ± 1.22
WAVE	66.94 ± 0.82	**76.18 ± 0.48 (↑15.11)**	71.51 ± 0.74	81.71 ± 0.48 (↑19.76)	71.20 ± 1.34	4.98 ± 1.37	74.66 ± 1.64	7.06 ± 1.64	72.46 ± 0.45 (↓2.30)	75.40 ± 0.32 (↓1.91)	75.17 ± 0.60

models to adapt to new subjects over time, and it points future models toward capturing subject-level distributional structure rather than assuming a shared representation. In addition, by storing only synthetic, compressed representations rather than raw EEG, the method reduces the amount of sensitive neural data that must be retained, which is beneficial for privacy in medical and BCIs settings. By improving continual cross-subject EEG decoding, `WAVE` could, if misused, ease inference of private mental states from neural signals without consent. Our synthetic replay memory mitigates this by avoiding raw EEG storage, though without formal guarantees; responsible deployment still requires regulation and privacy safeguards.

I ADDITIONAL STABILITY RESULTS

Here we extend the stability analysis to the source subjects, using the metric defined in Eq. 39, and restate the incremental-subject results from all the experiments above. Across nearly all settings, source stability is broadly consistent with incremental stability in terms of ranking: a method's ability to retain source knowledge and its ability to retain knowledge of previously seen incremental subjects are closely aligned, so the conclusions drawn from incremental stability also hold for the source domain. We report the incremental-versus-source breakdown for the main results (Tables 22, 23, 24, 25 and 26), the Sinkhorn ablation (Table 27), the loss-weight ablation (Tables 28, 29, 30 and 31), and the subject-order ablation (Tables 32, 33, 34, 35 and 36).

Table 22: Stability breakdown (incremental vs. source) of Table 2 on **BCI-IV-2a**. **Stability Incremental**: retention on previously seen incremental subjects, measured by the final model g_{N_I}. **Stability Source**: retention on the held-out source-domain test split, relative to g_0. **Bold** marks the best result among adapting methods (Oracle excluded), mean $\pm$ std over 4 seeds.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	53.26 ± 2.04	18.40 ± 2.40	51.51 ± 2.34	20.02 ± 2.54	52.61 ± 3.11	17.56 ± 4.29	47.98 ± 4.37	22.02 ± 5.76
Random Replay ($\lambda_r = 0.5, \lambda_c = 1$)	**66.66 ± 1.76**	**3.39 ± 1.33**	**66.83 ± 1.85**	**3.01 ± 1.21**	**71.11 ± 2.21**	**-0.94 ± 2.99**	**71.09 ± 2.18**	**-1.10 ± 2.80**
EWC ($\lambda = 10$)	53.75 ± 1.77	16.30 ± 1.73	52.03 ± 2.24	17.54 ± 2.39	53.28 ± 2.03	16.89 ± 2.80	48.44 ± 3.46	21.56 ± 4.34
LwF ($w = 1$)	56.73 ± 1.30	11.41 ± 1.41	56.14 ± 1.92	10.78 ± 2.08	58.83 ± 1.66	11.34 ± 1.83	56.21 ± 2.96	13.79 ± 3.33
WAVE	64.39 ± 1.17	6.36 ± 1.05	64.45 ± 1.31	6.04 ± 1.01	69.61 ± 1.97	0.57 ± 2.65	69.76 ± 2.08	0.24 ± 2.83

Table 23: Stability breakdown (incremental vs. source) of Table 3 on **PhysioNet-MI**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	53.93 ± 1.92	1.80 ± 1.15	53.98 ± 2.07	0.66 ± 0.80	46.53 ± 1.05	1.05 ± 2.91	45.99 ± 1.17	1.15 ± 2.71
Random Replay ($\lambda_c = 1, \lambda_r = 0.5$)	**54.52 ± 0.66**	**0.60 ± 1.46**	**54.68 ± 0.79**	**-0.73 ± 1.12**	**47.94 ± 1.45**	**-0.36 ± 1.98**	**47.27 ± 2.05**	**-0.13 ± 1.95**
EWC ($\lambda = 10$)	53.62 ± 1.29	1.89 ± 1.24	53.73 ± 1.37	0.52 ± 1.13	47.51 ± 0.78	0.07 ± 3.07	47.02 ± 1.19	0.12 ± 3.18
LwF ($w = 0.5$)	54.41 ± 1.27	1.06 ± 0.89	54.32 ± 1.40	0.11 ± 1.06	46.55 ± 1.40	1.02 ± 2.46	45.74 ± 1.65	1.40 ± 2.29
WAVE	53.49 ± 1.37	3.52 ± 1.56	53.36 ± 1.42	2.63 ± 1.15	45.30 ± 2.09	2.28 ± 1.26	43.95 ± 2.28	3.19 ± 2.22

Table 24: Stability breakdown (incremental vs. source) of Table 4 on **SHU-MI**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	58.73 ± 3.08	6.37 ± 3.30	58.28 ± 3.73	6.60 ± 3.88	66.19 ± 6.30	10.07 ± 5.83	66.24 ± 6.37	10.01 ± 5.98
Random Replay ($\lambda_c = 0.5$)	**63.87 ± 0.85**	**0.12 ± 0.92**	**63.85 ± 0.87**	**0.01 ± 0.91**	**76.89 ± 1.65**	**-0.63 ± 1.63**	**76.95 ± 1.71**	**-0.70 ± 1.61**
EWC ($\lambda = 4000$)	61.23 ± 1.54	2.62 ± 1.01	61.04 ± 1.48	2.56 ± 0.76	73.49 ± 1.85	2.77 ± 1.58	73.59 ± 2.02	2.66 ± 1.66
LwF ($w = 5$)	60.88 ± 2.42	1.80 ± 2.05	60.85 ± 2.49	1.67 ± 1.89	72.98 ± 3.61	3.28 ± 2.67	72.75 ± 3.79	3.50 ± 2.90
WAVE	60.21 ± 1.03	4.89 ± 0.51	59.81 ± 0.66	5.08 ± 0.99	71.31 ± 0.76	4.95 ± 1.50	71.52 ± 0.64	4.73 ± 1.34

Table 25: Stability breakdown (incremental vs. source) of Table 5 on **FACED**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	55.00 ± 2.71	-10.34 ± 1.73	54.43 ± 2.93	-5.62 ± 1.57	44.99 ± 1.19	-10.94 ± 1.84	43.91 ± 0.85	-10.24 ± 1.34
Random Replay ($\lambda_c = \lambda_r = 0.5$)	58.66 ± 2.47	**-13.52 ± 1.92**	58.21 ± 2.63	**-8.62 ± 2.22**	47.11 ± 0.25	-13.07 ± 0.90	46.28 ± 0.64	-12.61 ± 0.60
EWC ($\lambda = 10$)	54.05 ± 1.76	-9.99 ± 0.59	53.37 ± 1.92	-4.86 ± 0.52	44.73 ± 1.00	-10.69 ± 1.61	43.56 ± 1.00	-9.89 ± 1.26
LwF ($w = 1$)	56.27 ± 1.96	-11.60 ± 1.44	55.69 ± 2.29	-7.03 ± 1.81	46.04 ± 1.05	-12.00 ± 1.83	44.99 ± 0.60	-11.32 ± 1.43
WAVE	**59.30 ± 1.70**	-13.52 ± 0.51	**58.62 ± 1.92**	-8.05 ± 0.83	**47.50 ± 2.88**	**-13.45 ± 2.06**	**46.34 ± 2.84**	**-12.68 ± 2.19**

Table 26: Stability breakdown (incremental vs. source) of Table 6 on **ISRUC**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	57.74 ± 0.21	17.57 ± 0.36	60.37 ± 0.28	21.23 ± 0.25	72.59 ± 1.38	7.24 ± 0.74	73.08 ± 1.29	8.11 ± 0.87
Random Replay ($\lambda_c = \lambda_r = 0.5$)	**75.22 ± 0.22**	**0.69 ± 0.29**	**78.45 ± 0.32**	**3.17 ± 0.48**	76.96 ± 1.10	2.88 ± 0.57	**78.33 ± 1.04**	**2.85 ± 0.72**
EWC ($\lambda = 4000$)	68.57 ± 1.41	6.92 ± 1.39	70.64 ± 1.48	10.67 ± 1.44	**77.23 ± 0.66**	**2.61 ± 0.76**	78.13 ± 0.54	3.06 ± 0.52
LwF ($w = 1$)	58.50 ± 0.95	17.02 ± 1.03	61.11 ± 0.72	20.11 ± 1.02	74.80 ± 1.32	5.04 ± 0.73	75.32 ± 1.33	5.87 ± 0.74
WAVE	68.82 ± 1.07	7.07 ± 0.89	71.67 ± 0.97	10.09 ± 1.22	74.63 ± 1.21	5.21 ± 0.81	75.94 ± 0.83	5.25 ± 0.63

Table 27: Stability breakdown (incremental vs. source) of Table 7 for **Sinkhorn** on all the datasets.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Dataset	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
BCI-IV-2a	56.08 ± 2.43	14.58 ± 2.39	54.74 ± 2.61	15.60 ± 2.54	59.66 ± 0.87	10.51 ± 1.68	57.60 ± 1.41	12.40 ± 2.27
PhysioNet-MI	54.73 ± 0.90	1.71 ± 2.35	54.77 ± 0.97	0.83 ± 1.49	48.47 ± 0.88	-0.89 ± 3.02	47.80 ± 1.45	-0.66 ± 3.33
SHU-MI	59.03 ± 2.93	5.71 ± 3.00	58.48 ± 3.69	6.09 ± 3.70	68.54 ± 5.92	7.72 ± 5.04	68.63 ± 5.90	7.62 ± 5.04
FACED	55.96 ± 2.31	-11.18 ± 1.12	55.30 ± 2.42	-6.48 ± 1.15	44.91 ± 1.41	-10.87 ± 2.07	43.75 ± 1.11	-10.08 ± 1.64
ISRUC	67.30 ± 0.15	8.24 ± 0.14	71.40 ± 0.24	9.99 ± 0.29	71.64 ± 0.73	8.19 ± 0.62	74.01 ± 0.81	7.18 ± 0.85

Table 28: Stability breakdown (incremental vs. source) of the loss-weight sweep on **BCI-IV-2a**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	65.01 ± 1.33	5.55 ± 1.29	65.12 ± 1.47	5.18 ± 1.27	68.97 ± 3.04	1.20 ± 4.01	68.71 ± 3.33	1.29 ± 4.16
$\lambda_c = \lambda_r = 0.5$ (no virtual)	**65.83 ± 1.80**	**4.13 ± 1.77**	**65.92 ± 1.86**	**3.71 ± 1.81**	**69.71 ± 2.97**	**0.46 ± 3.34**	69.51 ± 3.26	0.49 ± 3.43
$\lambda_c = \lambda_r = \lambda_a = 0.5$	64.91 ± 1.52	4.81 ± 0.88	64.90 ± 1.64	4.47 ± 1.02	69.70 ± 2.34	0.47 ± 1.60	**69.84 ± 2.56**	**0.16 ± 1.79**
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	64.39 ± 1.17	6.36 ± 1.05	64.45 ± 1.31	6.04 ± 1.01	69.61 ± 1.97	0.57 ± 2.65	69.76 ± 2.08	0.24 ± 2.83

Table 29: Stability breakdown (incremental vs. source) of the loss-weight sweep on **PhysioNet-MI**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	53.60 ± 1.21	**3.05 ± 1.49**	53.46 ± 1.18	**2.26 ± 1.04**	45.57 ± 1.36	2.01 ± 2.18	44.49 ± 2.00	2.65 ± 1.91
$\lambda_c = \lambda_r = 0.5$ (no virtual)	**53.65 ± 1.00**	3.20 ± 1.35	**53.49 ± 0.96**	2.32 ± 0.74	45.95 ± 1.54	1.63 ± 2.31	44.84 ± 2.03	2.30 ± 2.52
$\lambda_c = \lambda_r = \lambda_a = 0.5$	53.37 ± 1.02	3.99 ± 1.95	53.27 ± 1.21	2.91 ± 1.41	**46.64 ± 2.05**	**0.94 ± 3.70**	**45.67 ± 3.01**	**1.47 ± 4.21**
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	53.49 ± 1.37	3.52 ± 1.56	53.36 ± 1.42	2.63 ± 1.15	45.30 ± 2.09	2.28 ± 1.26	43.95 ± 2.28	3.19 ± 2.22

Table 30: Stability breakdown (incremental vs. source) of the loss-weight sweep on **FACED**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	58.44 ± 1.24	-13.10 ± 0.19	57.87 ± 1.48	-8.32 ± 0.48	46.60 ± 1.20	-12.55 ± 0.65	45.86 ± 1.19	-12.19 ± 1.15
$\lambda_c = \lambda_r = 0.5$ (no virtual)	59.06 ± 1.40	**-14.10 ± 0.31**	58.45 ± 1.60	**-9.17 ± 0.67**	47.24 ± 1.64	-13.20 ± 0.89	**46.57 ± 1.73**	**-12.91 ± 1.42**
$\lambda_c = \lambda_r = \lambda_a = 0.5$	**59.78 ± 1.77**	-14.33 ± 0.69	**59.14 ± 1.97**	-9.12 ± 1.02	47.38 ± 2.51	-13.34 ± 1.67	46.25 ± 2.37	-12.59 ± 1.93
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$ (WAVE)	59.30 ± 1.70	-13.52 ± 0.51	58.62 ± 1.92	-8.05 ± 0.83	**47.50 ± 2.88**	**-13.45 ± 2.06**	46.34 ± 2.84	-12.68 ± 2.19

Table 31: Stability breakdown (incremental vs. source) of the loss-weight sweep on **ISRUC**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Configuration	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
$\lambda_c = 1,\ \lambda_r = 0.5$ (no virtual)	68.74 ± 0.42	7.12 ± 0.71	71.59 ± 0.45	10.21 ± 0.80	74.35 ± 0.85	5.49 ± 0.41	75.70 ± 0.43	5.49 ± 0.61
$\lambda_c = \lambda_r = 0.5$ (no virtual) (WAVE)	**68.82 ± 1.07**	**7.07 ± 0.89**	**71.67 ± 0.97**	**10.09 ± 1.22**	**74.63 ± 1.21**	**5.21 ± 0.81**	**75.94 ± 0.83**	**5.25 ± 0.63**
$\lambda_c = \lambda_r = \lambda_a = 0.5$	60.86 ± 3.62	13.49 ± 3.80	63.47 ± 2.77	17.02 ± 2.93	71.27 ± 0.75	8.56 ± 0.60	72.51 ± 0.59	8.68 ± 0.69
$\lambda_c = 0.5,\ \lambda_r = \lambda_a = 0.25$	60.77 ± 3.44	13.76 ± 3.38	63.48 ± 2.74	17.16 ± 2.70	71.33 ± 0.54	8.51 ± 0.34	72.56 ± 0.35	8.63 ± 0.47

Table 32: Stability breakdown (incremental vs. source) of the subject-order ablation on **BCI-IV-2a**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Method	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	56.27 ± 1.99	14.18 ± 3.37	54.38 ± 2.49	15.64 ± 3.63	54.49 ± 3.75	14.61 ± 3.75	52.80 ± 5.79	16.66 ± 5.79
Random Replay ($\lambda_r = 0.5, \lambda_c = 1$)	**66.88 ± 0.73**	**2.33 ± 2.11**	**66.62 ± 0.68**	**2.16 ± 1.87**	**70.17 ± 3.70**	**-1.07 ± 3.70**	**69.97 ± 4.06**	**-0.51 ± 4.06**
EWC ($\lambda = 10$)	55.70 ± 2.37	14.43 ± 2.82	53.74 ± 2.54	15.94 ± 2.77	54.86 ± 3.56	14.24 ± 3.56	52.87 ± 6.17	16.59 ± 6.17
LwF ($w = 1$)	58.68 ± 2.67	9.43 ± 3.21	57.39 ± 2.94	10.03 ± 3.18	57.02 ± 4.33	12.08 ± 4.33	55.49 ± 5.41	13.97 ± 5.41
WAVE	65.91 ± 2.89	3.72 ± 3.13	65.65 ± 2.93	3.65 ± 2.87	68.69 ± 0.71	0.41 ± 0.71	68.60 ± 0.61	0.86 ± 0.61

Table 33: Stability breakdown (incremental vs. source) of the subject-order ablation on **PhysioNet-MI**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Method	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	53.33 ± 2.16	2.72 ± 1.98	53.08 ± 2.52	1.79 ± 1.70	45.08 ± 1.28	3.18 ± 1.28	45.14 ± 1.65	3.01 ± 1.65
Random Replay ($\lambda_c = 1, \lambda_r = 0.5$)	**54.86 ± 0.93**	**-0.15 ± 1.11**	**54.94 ± 0.98**	**-1.42 ± 0.91**	**49.07 ± 3.02**	**-0.82 ± 3.02**	**49.65 ± 2.91**	**-1.49 ± 2.91**
EWC ($\lambda = 10$)	52.98 ± 1.80	2.72 ± 1.71	52.76 ± 2.30	1.60 ± 1.57	46.16 ± 1.65	2.09 ± 1.65	46.39 ± 1.84	1.77 ± 1.84
LwF ($w = 0.5$)	53.54 ± 1.80	2.52 ± 1.79	53.33 ± 2.07	1.66 ± 1.36	45.21 ± 1.18	3.04 ± 1.18	45.14 ± 1.38	3.01 ± 1.38
WAVE	52.31 ± 2.07	4.62 ± 2.78	52.13 ± 2.31	2.96 ± 2.20	44.85 ± 2.32	3.41 ± 2.32	44.51 ± 2.55	3.65 ± 2.55

Table 34: Stability breakdown (incremental vs. source) of the subject-order ablation on **SHU-MI**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Method	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	59.10 ± 1.00	6.18 ± 0.81	59.04 ± 1.03	6.11 ± 0.83	70.54 ± 2.47	7.25 ± 2.47	70.66 ± 2.34	6.99 ± 2.34
Random Replay ($\lambda_c = 0.5$)	**64.45 ± 1.32**	**0.19 ± 1.63**	**64.45 ± 1.33**	**0.10 ± 1.43**	**77.85 ± 1.47**	**-0.07 ± 1.47**	**77.84 ± 1.47**	**-0.20 ± 1.47**
EWC ($\lambda = 4000$)	60.15 ± 2.48	4.28 ± 2.41	60.06 ± 2.36	4.27 ± 2.14	74.70 ± 2.71	3.08 ± 2.71	74.89 ± 2.62	2.75 ± 2.62
LwF ($w = 5$)	61.03 ± 2.21	2.07 ± 2.34	60.88 ± 2.27	2.13 ± 1.97	75.34 ± 2.42	2.44 ± 2.42	75.39 ± 2.23	2.25 ± 2.23
WAVE	59.88 ± 0.66	5.02 ± 0.43	59.28 ± 0.40	5.42 ± 0.36	73.45 ± 1.12	4.33 ± 1.12	73.53 ± 1.11	4.11 ± 1.11

Table 35: Stability breakdown (incremental vs. source) of the subject-order ablation on **FACED**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Method	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	54.27 ± 1.78	-11.64 ± 2.14	53.88 ± 2.06	-7.84 ± 2.76	47.17 ± 1.04	-13.10 ± 1.04	47.07 ± 1.93	-13.27 ± 1.93
Random Replay ($\lambda_c = \lambda_r = 0.5$)	57.19 ± 1.02	**-13.16 ± 1.08**	56.66 ± 1.08	**-9.15 ± 1.46**	48.07 ± 2.08	-14.00 ± 2.08	47.55 ± 1.86	-13.75 ± 1.86
EWC ($\lambda = 10$)	54.22 ± 1.39	-11.20 ± 1.29	53.77 ± 1.76	-7.20 ± 2.10	47.56 ± 1.92	-13.50 ± 1.92	47.33 ± 2.15	-13.53 ± 2.15
LwF ($w = 1$)	55.84 ± 1.76	-12.38 ± 2.11	55.55 ± 2.18	-8.77 ± 2.76	47.36 ± 1.39	-13.30 ± 1.39	47.36 ± 2.05	-13.56 ± 2.05
WAVE	**57.47 ± 0.78**	-12.89 ± 0.50	**56.81 ± 0.90**	-8.60 ± 0.83	**49.24 ± 1.86**	**-15.18 ± 1.86**	**48.55 ± 2.42**	**-14.75 ± 2.42**

Table 36: Stability breakdown (incremental vs. source) of the subject-order ablation on **ISRUC**.

	Stability Incremental				Stability Source			
	BACC		MF1		BACC		MF1	
Method	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting	Retention	Forgetting
Naive fine-tuning	61.09 ± 2.48	14.84 ± 2.39	64.60 ± 3.11	17.24 ± 3.04	70.98 ± 4.39	9.15 ± 4.39	73.24 ± 3.73	8.08 ± 3.73
Random Replay ($\lambda_c = \lambda_r = 0.5$)	**75.07 ± 0.41**	**0.90 ± 0.38**	**78.51 ± 0.33**	**2.97 ± 0.27**	**77.85 ± 0.40**	**2.28 ± 0.40**	**79.42 ± 0.45**	**1.90 ± 0.45**
EWC ($\lambda = 4000$)	68.08 ± 1.00	7.99 ± 1.07	71.33 ± 0.67	10.20 ± 0.65	75.39 ± 1.80	4.74 ± 1.80	77.50 ± 1.31	3.82 ± 1.31
LwF ($w = 1$)	61.94 ± 2.17	13.43 ± 2.36	65.34 ± 2.62	15.61 ± 2.81	72.62 ± 6.36	7.51 ± 6.36	74.59 ± 5.31	6.73 ± 5.31
WAVE	71.20 ± 1.34	4.98 ± 1.37	74.66 ± 1.64	7.06 ± 1.64	75.30 ± 0.95	4.83 ± 0.95	76.72 ± 0.88	4.60 ± 0.88