NEUROACCORD: TIME–SPACE–FREQUENCY GUIDED CONTINUAL LEARNING FOR LARGE EEG MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG foundation models learn transferable representations for diverse decoding tasks. As EEG datasets and paradigms expand, continual learning is needed to acquire new capabilities while retaining performance on earlier tasks. Existing approaches mainly rely on two perspectives as protecting important model parameters and replaying representative historical samples. We observe that protected parameters gradually lose importance for their original tasks as the shared model evolves. Meanwhile, fixed replay buffers cover a diminishing share of historical EEG knowledge amid heterogeneity across tasks and subjects. These observations call for a reference that remains meaningful as parameter roles and sample coverage change. We introduce NeuroAccord, a Time–Space–Frequency (TSF)-guided framework linking physiologically motivated EEG descriptors to foundation-model features. It constructs a TSF basis from current-task descriptors and frozen previous-model features, transfers the teacher's relations among examples, and preserves TSF coupling during adaptation. Across stages, accumulated TSF-related directions constrain feature drift without replaying earlier recordings. In task-incremental experiments spanning nine datasets, six paradigms, and four EEG foundation models, NeuroAccord achieves the highest average anytime accuracy and backward transfer and the lowest forgetting among 18 baselines. These results establish state-of-the-art performance among the evaluated methods and a favorable plasticity–stability balance. Ablation, robustness, and continued-learning analyses further support this balance. More broadly, the findings show the value of grounding evolving model representations in EEG signal structure to accumulate new knowledge while retaining earlier decoding capabilities.

1 INTRODUCTION

EEG foundation models have emerged as an important research direction in brain-signal modeling. Through large-scale pretraining, they learn transferable representations that provide a shared basis for diverse downstream decoding tasks (Wang et al., 2025; Jiang et al., 2024). Knowledge accumulation, however, extends beyond pretraining (Van de Ven et al., 2022). As EEG datasets and task paradigms continue to diversify, there is growing demand for models that can adapt to new signal distributions and decoding objectives while progressively expanding their capabilities. Continual learning provides a framework for acquiring new knowledge while preserving existing capabilities, enabling knowledge gained across successive tasks to accumulate within a shared model (Kirkpatrick et al., 2017).

Two complementary perspectives on preserving prior knowledge in continual learning focus on model parameters and data samples (Zenke et al., 2017). The parameter perspective focuses on how knowledge is encoded in the model, identifying important parameters and constraining them to retain existing capabilities (Rebuffi et al., 2017). The sample perspective focuses on historical learning experience, replaying representative examples to reinforce previous knowledge (Verwimp et al., 2021). Together, these perspectives constrain continual learning to support the integration of new and existing knowledge. Nevertheless, protecting parameters or replaying samples does not guarantee that the model retains its knowledge in EEG foundation model (Aljundi et al., 2018).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

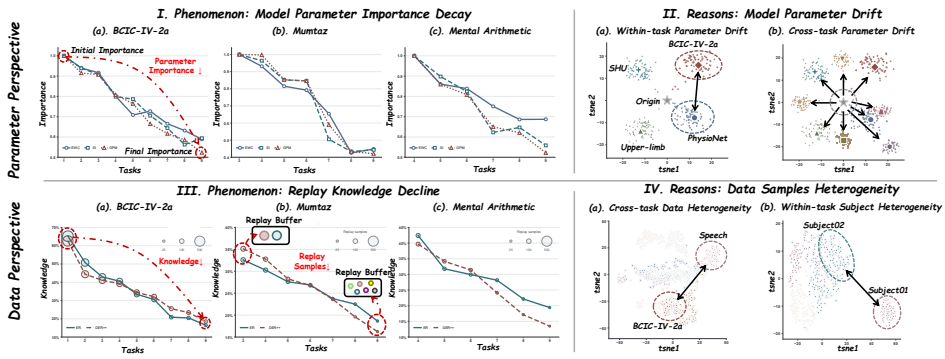


Figure 1: Observations in continual EEG learning. (I) Protected parameters lose importance to earlier tasks. (II) Parameter drift within and across tasks. (III) Historical-knowledge coverage declines as tasks accumulate. (IV) Task and subject heterogeneity hinders representative replay.

From the parameter perspective, protecting parameters identified as important to earlier tasks aims to sustain their contribution to prior decoding capabilities. However, we observe a contrary pattern in EEG foundation models: **key protected parameters gradually become less important to their original tasks** (Fig. 1, I). This suggests that protecting important parameters may not consistently preserve the knowledge supporting earlier EEG tasks (Aljundi et al., 2018). To investigate this phenomenon, we compare models fine-tuned on different datasets and observe parameter drift within and across tasks (Fig. 1, II). We attribute this drift to the joint encoding of task and subject information in EEG foundation models (Benzing, 2022). Across EEG datasets, variation in decoding objectives and subject populations can reshape the shared representations of an EEG foundation model (Fan et al., 2025). As these representations evolve, individual parameters may contribute differently to earlier tasks, leaving historical importance estimates misaligned with the updated model's reliance on them (Lin et al., 2026). Therefore, constraining a fixed set of historically important parameters may be insufficient to preserve prior decoding capabilities.

From the data perspective, replaying historical samples allows the model to revisit past learning experiences (Rolnick et al., 2019). However, we observe the following phenomenon: **retained samples cover a diminishing share of historical knowledge as tasks accumulate** (Fig. 1, III). Under a fixed memory budget, accommodating a growing set of tasks leaves fewer samples to represent each earlier task, and these limited samples may not sufficiently cover its learned knowledge. Further analysis of historical EEG distributions reveals pronounced heterogeneity across task paradigms and across subjects within the same task (Fig. 1, IV). Thus, a buffer spanning all historical EEG tasks can still undersample subject-specific signal patterns within each task (Brain, 2022). As datasets accumulate, a fixed memory must represent both differences across EEG paradigms and variation among subjects within the same decoding task. Sparse replay of these EEG patterns then provides weaker constraints on later updates to the shared backbone, widening the gap between replaying past recordings and preserving the corresponding decoding capabilities (Aljundi et al., 2019).

These two observations raise our central research question: **How can we better organize and preserve EEG knowledge as EEG foundation models learn across tasks?** Rather than relying solely on stage-specific parameter states or sample selections, we turn to the intrinsic structure of EEG as a physiological signal, seeking a common basis for knowledge preservation across tasks.

This perspective is grounded in the link between EEG and neural activity. As scalp recordings of population-level neural electrical activity, EEG signals exhibit temporal dynamics, spatial distributions, and frequency-specific oscillatory activity with a neurophysiological basis (Buzsáki et al., 2012; Buzsaki & Draguhn, 2004). EEG decoding studies illustrate the relevance of these dimensions across tasks. Time-domain waveform statistics have been used for emotional stress detection (Hasan & Kim, 2019). Multichannel covariance patterns support motor imagery discrimination (Ramoser et al., 2000), while responses at stimulus frequencies and their harmonics enable target identification in SSVEP tasks (Chen et al., 2015). These findings motivate a Time–Space–Frequency (TSF) description of EEG. The rules defining TSF descriptors do not change with model weights, and their association with learned features captures how the model represents these signal properties. This relation supplies

a consistent reference as parameter roles evolve, while enabling prior knowledge to be maintained through signal–feature associations without relying solely on a finite replay buffer. TSF thereby bridges EEG data structure and model representations across learning stages.

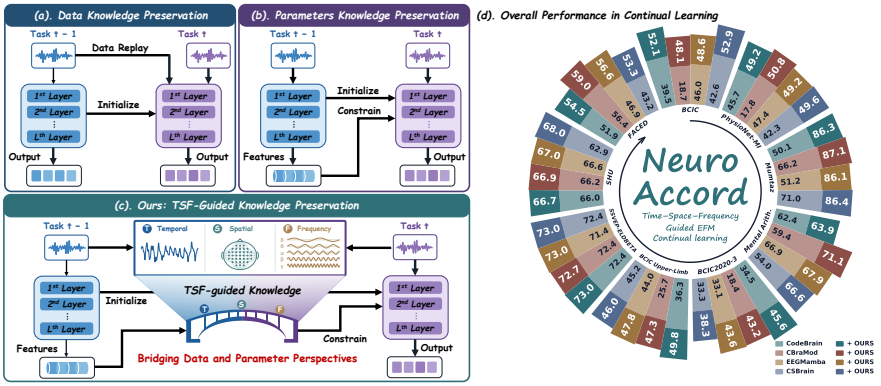


Figure 2: Continual-learning perspectives and the TSF-guided framework. (a) Sample replay revisits historical data; (b) model constraints preserve learned representations. (c) NeuroAccord connects TSF descriptors with model features to guide knowledge transfer and retention. (d) Four foundation models are compared with and without NeuroAccord across nine EEG datasets.

Building on this perspective, we propose NeuroAccord, a continual-learning framework that uses EEG signal structure to guide knowledge accumulation (Fig. 2). We associate EEG Time, Space and Frequency with model features to construct a TSF basis. During new-task learning, it preserves sample relations and TSF coupling to transfer prior knowledge while allowing task-specific adaptation. Across stages, NeuroAccord accumulates TSF-related feature directions and limits drift along them on current-task inputs, retaining earlier knowledge without replaying historical recordings. We evaluate NeuroAccord across four EEG foundation models and nine datasets spanning six decoding paradigms, comparing it with 18 baselines and NeuroAccord achieve the best plasticity and stability.

Our contributions of this work are: **(I)** We reveal the limitations of parameter protection and sample replay in EEG foundation models and introduce TSF as a bridge between EEG data and evolving model representations. **(II)** We develop NeuroAccord around this bridge, combining relational transfer, TSF-coupling preservation, and historical feature anchoring to support adaptation and retention. **(III)** Across nine datasets and four foundation models, NeuroAccord achieves the best performance among 18 baselines, demonstrating a favorable plasticity–stability balance.

2 Preliminaries

2.1 Problem Set Up

We consider a sequence of EEG tasks $\{\mathcal{T}_t\}_{t=1}^{T}$, each with a data distribution $\mathcal{D}_t$ and label space $\mathcal{Y}_t$. At stage t, the ideal objective is

$$\min_{\theta,\omega_t} \mathcal{R}_t(\theta,\omega_t) \quad \text{s.t.} \quad \mathcal{R}_j(\theta,\omega_j) \le \mathcal{R}_j(\theta_{t-1},\omega_j), \quad \forall j < t. \tag{1}$$

Here, $\mathcal{R}_j(\theta,\omega_j)$ is the expected loss on task j. Starting from a pretrained backbone f_{θ_0}, stage t updates its shared parameters from θ_{t-1} to θ_t and trains a task-specific head g_{ω_t} $\mathcal{D}_t^{\text{tr}}$.

2.2 A Common TSF Reference Across EEG Tasks

We formalize the temporal–spatial–frequency (TSF) description that provides a common signal reference across EEG tasks. Each EEG recording x is described through three views,

$$\phi(x) = \big[\phi^T(x), \phi^S(x), \phi^F(x)\big]. \tag{2}$$

The temporal view ϕ^T contains six amplitude and waveform statistics; the spatial view ϕ^S contains five channel-correlation summaries; and the frequency view ϕ^F contains five relative-power measures

in the δ, θ, α, β, and γ bands. Together, these descriptors provide a reference for linking EEG structure to model features. Appendix A details their definitions and interpretations.

3 NEUROACCORD

3.1 FRAMEWORK OVERVIEW

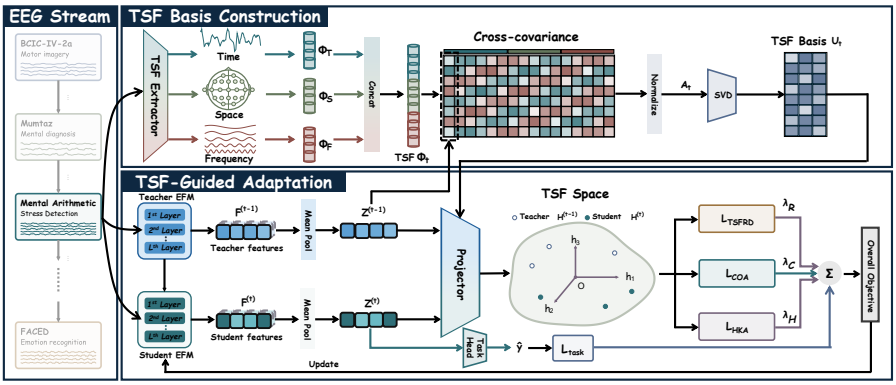


Figure 3: Overview of NeuroAccord.

In this work,we propose NeuroAccord as shown in Fig. 3. NeuroAccord adapts a pretrained EEG foundation model through a shared backbone and task-specific heads. For each subsequent task, TSF descriptors and frozen previous-backbone features of current examples define a TSF basis in feature space. TSFRD transfers the teacher's relations among examples, while COA preserves TSF coupling to guide new-task learning. HKA limits feature drift along accumulated historical TSF directions.

3.2 TSF BASIS CONSTRUCTION

Building on the TSF representation introduced in Section 2.2, we first describe how this structure is incorporated into an EEG foundation model. Let $f_\theta(x) \in \mathbb{R}^{L(x)\times d}$ denote the backbone's feature sequence and $z_\theta(x) = \mathrm{pool}(f_\theta(x))$ its mean-pooled representation, where d is the feature dimension and $L(x)$ is the number of feature tokens for input x. At each stage after the first, we freeze the stage-start backbone as a teacher and denote its pooled representation by $\bar{z}_t(x)$. For the current training dataset $\mathcal{D}_t^{\mathrm{tr}} = \{(x_i, y_i)\}_{i=1}^{n_t}$, we collect teacher features and TSF descriptors as

$$Z_t^0 = \left[\bar{z}_t(x_i)^\top\right]_{i=1}^{n_t} \in \mathbb{R}^{n_t\times d}, \qquad \Phi_t^v = \left[\phi^v(x_i)^\top\right]_{i=1}^{n_t} \in \mathbb{R}^{n_t\times d_v}.$$

Here $\phi^v(x_i)$ is the d_v-dimensional descriptor for view $v \in \mathcal{V} = \{T, S, F\}$. To identify feature directions within the EEG foundation model that are associated with TSF, we measure how teacher features co-vary with TSF descriptors across the current training set. Specifically, let $J_n = I_n - \mathbf{1}\mathbf{1}^\top/n$ be the centering matrix, and let $\mathcal{S}_n$ center descriptor columns and divide by their population standard deviations. We compute cross covariance for each view and concatenate their normalize:

$$C_t^v = \frac{(J_{n_t} Z_t^0)^\top \mathcal{S}_{n_t}(\Phi_t^v)}{n_t - 1}, \qquad A_t = \left[\frac{C_t^v}{\|C_t^v\|_F}\right]_{v\in\mathcal{V}}. \tag{3}$$

Here $C_t^v \in \mathbb{R}^{d\times d_v}$ summarizes linear associations between teacher feature coordinates and the descriptors of view v, which we term TSF coupling. The concatenated matrix $A_t \in \mathbb{R}^{d\times\sum_{v\in\mathcal{V}} d_v}$ represents their joint TSF coupling in the model's feature space. To focus relational constraints on a limited set of model directions with the strongest TSF coupling, we extract a rank-r subspace using singular value decomposition (SVD):

$$A_t = \widetilde{U}_t \Sigma_t V_t^\top, \qquad U_t = [\widetilde{u}_{t,1}, \ldots, \widetilde{u}_{t,r}] \in \mathbb{R}^{d\times r}. \tag{4}$$

Here $\widetilde{U}_t$ and V_t contain the left and right singular vectors, and Σ_t contains the singular values in descending order. Retaining the first r columns $\widetilde{u}_{t,j}$ of $\widetilde{U}_t$ gives the orthonormal TSF basis U_t, which remains fixed during the stage. This basis links the TSF to the EEG foundation model's feature space.

3.3 TSF-GUIDED KNOWLEDGE TRANSFER

Having linked TSF to the model's feature space, we next consider how this connection can support the reuse of accumulated knowledge during new-task learning. Inspired by relational distillation (Park et al., 2019), we view the previous model's relations among current examples as a transferable account of how it organizes the data. Yet these relations do not specify how its features correspond to TSF. We therefore also consider TSF coupling, motivating TSFRD and COA as complementary objectives.

3.3.1 TSF RELATIONAL DISTILLATION

Uniformly matching the previous model's relations among current examples overlooks their structural similarity in the TSF basis. We therefore design TSFRD to weight the distillation of similarities by TSF similarity, prioritizing structurally similar pairs while relaxing constraints on dissimilar ones.

We first compute pairwise cosine similarities using coordinates in the TSF basis. For n current-training examples, let $Z^{(t)}, Z^{(t-1)} \in \mathbb{R}^{n\times d}$ denote pooled features from the model being updated at stage t and the frozen previous-stage model. Their projections are $H^{(t)} = Z^{(t)}U_t$ and $H^{(t-1)} = Z^{(t-1)}U_t$, where $h_i^{(t)}$ denotes the projected feature vector of example i. For sample indices $i, j \in \{1, \dots, n\}$, let $G_{ij}^{(t)}$ denote the cosine similarity between their projected features:

$$G_{ij}^{(t)} = \frac{(h_i^{(t)})^\top h_j^{(t)}}{\|h_i^{(t)}\|_2 \, \|h_j^{(t)}\|_2}. \tag{5}$$

To determine which feature relations receive stronger preservation, we compute TSF-based similarity weights. For view v, let $\widetilde{\Phi}^v = \mathcal{S}_n(\Phi^v)$ be the standardized descriptor matrix and define, for $i \neq j$,

$$K_{ij}^v = \exp\left(-\frac{\|\widetilde{\Phi}_i^v - \widetilde{\Phi}_j^v\|_2^2}{d_v \tau_v}\right), \qquad \tau_v = \mathrm{median}_{p\neq q} \frac{\|\widetilde{\Phi}_p^v - \widetilde{\Phi}_q^v\|_2^2}{d_v}. \tag{6}$$

The bandwidth τ_v captures the typical descriptor separation within each view, yielding larger weights for structurally similar pairs. Let $G^{(t)}, G^{(t-1)} \in \mathbb{R}^{n\times n}$ collect the current and previous models' pairwise cosine similarities. To preserve the previous model's relational structure, we define the TSF relational distillation (TSFRD) loss as follows:

$$\mathcal{L}_{\mathrm{TSFRD}} = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \frac{\sum_{i\neq j} K_{ij}^v [G_{ij}^{(t)} - G_{ij}^{(t-1)}]^2}{\sum_{i\neq j} K_{ij}^v}. \tag{7}$$

TSFRD thus selectively preserves the teacher's projected relational geometry, using TSF to prioritize relations rather than prescribe feature coordinates.

3.3.2 COUPLING ORIENTATION ALIGNMENT

TSFRD preserves pairwise relations among examples, but these relations can remain unchanged as the model's coupling with TSF drifts. We therefore introduce coupling orientation alignment (COA) to preserve the linear associations between projected features and TSF descriptors during adaptation.

Formally, we average products of the centered projected features $J_n H^{(t)}$ and TSF descriptors $J_n \Phi^v$ to obtain their cross-covariance $M^v(H^{(t)})$, which we normalize as follows:

$$M^v(H^{(t)}) = \frac{(J_n H^{(t)})^\top J_n \Phi^v}{n-1}, \qquad \widehat{M}^v(H^{(t)}) = \frac{M^v(H^{(t)})}{\|M^v(H^{(t)})\|_F}. \tag{8}$$

The normalized matrix $\widehat{M}^v(H^{(t)})$ describes the pattern of TSF coupling independently of its overall magnitude. Taking the corresponding matrix $\widehat{M}^v(H^{(t-1)})$ from the frozen previous model as the reference, we define the COA loss:

$$\mathcal{L}_{\mathrm{COA}} = \frac{1}{2|\mathcal{V}|} \sum_{v\in\mathcal{V}} \|\widehat{M}^v(H^{(t)}) - \widehat{M}^v(H^{(t-1)})\|_F^2. \tag{9}$$

COA measures the average cosine distance between normalized cross-covariance matrices, preserving the orientation rather than the magnitude of TSF coupling. It complements TSFRD by penalizing rotations that preserve sample relations but alter TSF coupling, as detailed in Appendix B.

3.4 CROSS-TASK KNOWLEDGE RETENTION

In Section 3.3, we introduced TSFRD and COA to transfer knowledge within the current task's TSF basis, but this basis may not cover directions identified in earlier stages. We therefore introduce historical knowledge anchoring (HKA), which accumulates these directions and constrains feature drift along their joint span on current-task inputs using the frozen previous-stage model as a reference.

Specifically, at the end of task t, we apply the construction in Section 3.2 to the selected backbone and current training data, obtaining the rank-r stage-end TSF basis $E_t \in \mathbb{R}^{d\times r}$. We combine it with the previously retained basis to obtain the accumulated historical TSF basis Q_t of dimension q_t as

$$Q_t = \mathrm{orth}([Q_{t-1}, E_t]) \in \mathbb{R}^{d\times q_t}, \tag{10}$$

where $q_t = \mathrm{rank}(Q_t)$ is the number of independent historical feature directions retained after task t.

During task t, HKA constrains feature drift along Q_{t-1} using n current-training examples. Let $F^{(t)}, F^{(t-1)} \in \mathbb{R}^{m\times d}$ collect their token features from the current model and the frozen previous-stage model, respectively. To measure drift relative to the previous model's feature scale, we normalize the squared discrepancy within Q_{t-1} by its feature variance ν_{t-1} and obtain

$$\nu_{t-1} = \frac{\|J_m F^{(t-1)}\|_F^2}{md}, \qquad \mathcal{L}_{\mathrm{HKA}} = \frac{\|(F^{(t)} - F^{(t-1)})Q_{t-1}\|_F^2}{mq_{t-1}\nu_{t-1}}. \tag{11}$$

Here $m = \sum_{i=1}^{n} L(x_i)$ is the total number of feature tokens,d is the feature dimension, and q_{t-1} is the dimension of the retained historical subspace. HKA restricts feature drift to the historical span, leaving its orthogonal complement unconstrained by this loss. This selective constraint anchors historical knowledge while allowing new-task adaptation. Appendix B gives the gradient derivation.

3.5 OVERALL OBJECTIVE AND TRAINING

Current-task supervision and the three consistency objectives define the training loss,

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\mathrm{task}} &+ \lambda_R \mathcal{L}_{\mathrm{TSFRD}} \\ &+ \lambda_C \mathcal{L}_{\mathrm{COA}} + \lambda_H \mathcal{L}_{\mathrm{HKA}},\end{aligned} \tag{12}$$

where $\lambda_R, \lambda_C, \lambda_H \geq 0$ weight the three objectives. TSFRD and COA preserve sample relations and TSF coupling in the coordinates defined by U_t, while HKA constrains feature drift within the span of Q_{t-1}. At stages $t > 1$, the frozen model supplies the reference features for all three objectives. Algorithm 1 summarizes this cycle of knowledge reuse and historical maintenance.

Algorithm 1: NeuroAccord continual learning

Input: Pretrained θ_0; task data; TSF extractor.
for $t = 1, \ldots, T$ **do**
  Start from θ_{t-1}; initialize ω_t
  if $t > 1$ **then**
    Freeze teacher from θ_{t-1}
    Construct U_t on $\mathcal{D}_t^{\mathrm{tr}}$, Eqs. (3)–(4)
  for *each scheduled update on* $\mathcal{D}_t^{\mathrm{tr}}$ **do**
    Optimize Eq. (12)
  Select (θ_t, ω_t)
  Compute E_t from θ_t and $\mathcal{D}_t^{\mathrm{tr}}$
  Update Q_t, Eq. (10)
return $\theta_T, \{\omega_t\}_{t=1}^{T}$

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENT SETUP

Datasets. To evaluate NeuroAccord across diverse EEG applications, we use 9 datasets with 6 BCI tasks: (1) motor imagery, (2) emotion recognition, (3) imagined speech, (4) SSVEP, (5) mental disorder diagnosis, and (6) mental stress detection. We order the datasets chronologically by their public release year to define the primary continual-learning sequence, as summarized in Table 1.

Backbones and Baselines. We evaluate NeuroAccord on four publicly available EEG foundation models spanning different architectural designs. Specifically, we consider two Transformer-based backbones, CBraMod (Wang et al., 2025) and CSBrain (Zhou et al., 2026), and two state-space-model (SSM)-style backbones, EEGMamba (Gui et al., 2024) and CodeBrain (Ma et al., 2026). We compare 18 continual-learning baselines from the model parameter and data sample views. Appendix C describes the datasets, backbones, and baseline mechanisms.

Table 1: Dataset statistics in chronological task order.

Order	Dataset	Paradigm	Year	Duration (s)	Input ch.	Labels	Subjects
1	BCIC-IV-2a (Tangermann et al., 2012)	Motor imagery	2008	4	22	4	9
2	PhysioNet-MI (Schalk et al., 2004)	Motor imagery	2009	4	64	4	109
3	Mumtaz (Mumtaz, 2016)	Mental disorder diagnosis	2016	5	19	2	64
4	Mental Arithmetic (Zyma et al., 2019)	Mental stress detection	2018	5	20	2	36
5	BCIC2020-3 (Jeong et al., 2022)	Imagined speech	2020	3	64	5	15
6	BCIC-UpperLimb (Jeong et al., 2022)	Motor imagery	2020	4	75	3	15
7	eldBETA (Liu et al., 2022)	SSVEP	2022	3.2	75	9	100
8	SHU-MI (Ma et al., 2022)	Motor imagery	2022	4	32	2	25
9	FACED (Chen et al., 2023)	Emotion recognition	2023	10	32	9	123

Evaluation Metrics. We report average anytime accuracy (AAA), forgetting (FOR), and backward transfer (BWT). Let a_{tj} denote the balanced accuracy on task j after learning task t:

$$\mathrm{AAA} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{t}\sum_{j=1}^{t} a_{tj},\ \mathrm{FOR} = \frac{1}{T-1}\sum_{j=1}^{T-1}\Big(\max_{j\le s\le T} a_{sj} - a_{Tj}\Big),\ \mathrm{BWT} = \frac{1}{T-1}\sum_{j=1}^{T-1}(a_{Tj} - a_{jj}). \quad (13)$$

4.2 MAIN RESULTS

Table 2: Continual-learning performance across four EEG foundation models (%).

Method	Continual learning view	CBraMod			CSBrain			EEGMamba			CodeBrain		
		AAA↑	**FOR↓**	**BWT↑**	**AAA↑**	**FOR↓**	**BWT↑**	**AAA↑**	**FOR↓**	**BWT↑**	**AAA↑**	**FOR↓**	**BWT↑**
Naive CL	-	**53.17**	**20.95**	**–20.87**	**55.36**	**12.78**	**–12.78**	**53.69**	**6.86**	**–6.86**	**54.76**	**13.66**	**–13.51**
LwF (Li & Hoiem, 2017)	Model parameters	**56.48**	**10.12**	**–9.94**	**58.94**	**4.40**	**–4.26**	**56.00**	**5.04**	**–4.94**	**56.78**	**13.75**	**–13.58**
EWC (Kirkpatrick et al., 2017)		**55.04**	**11.26**	**–11.26**	**57.52**	**7.07**	**–7.07**	**55.46**	**6.37**	**–5.92**	**53.65**	**13.47**	**–13.47**
SI (Zenke et al., 2017)		**57.52**	**9.59**	**–9.59**	**55.75**	**13.15**	**–13.15**	**55.04**	**7.23**	**–7.23**	**56.70**	**13.94**	**–13.77**
GPM (Saha et al., 2021)		**53.91**	**16.33**	**–16.18**	**54.38**	**12.61**	**–12.59**	**54.70**	**6.85**	**–6.68**	**55.50**	**8.96**	**–8.96**
TRGP (Lin et al., 2022)		**57.44**	**10.02**	**–9.98**	**55.03**	**12.50**	**–12.50**	**55.64**	**3.35**	**–3.29**	**55.30**	**12.14**	**–12.14**
SGP (Saha & Roy, 2023)		**53.72**	**15.61**	**–15.59**	**54.39**	**11.90**	**–11.90**	**54.76**	**7.72**	**–7.72**	**56.42**	**8.69**	**–8.69**
DT2W (Qiao et al., 2023)		**58.93**	**6.43**	**–6.04**	**56.71**	**4.33**	**–4.13**	**57.29**	**5.07**	**–5.07**	**59.46**	**6.06**	**–5.81**
O-LoRA (Wang et al., 2023)		**57.52**	**10.41**	**–10.41**	**57.14**	**7.19**	**–6.69**	**56.90**	**2.66**	**–2.49**	**57.35**	**12.11**	**–12.11**
SD-LoRA (Wu et al., 2025)		**55.77**	**10.86**	**–10.58**	**55.87**	**4.48**	**–4.24**	**55.65**	**4.47**	**–4.35**	**57.00**	**11.30**	**–11.30**
BiLoRA (Zhu et al., 2025)		**57.72**	**8.65**	**–8.19**	**57.33**	**4.30**	**–4.26**	**54.37**	**5.31**	**–4.72**	**58.90**	**9.32**	**–9.27**
CL-LoRA (He et al., 2025)		**59.35**	**8.00**	**–7.92**	**57.58**	**8.47**	**–8.47**	**52.23**	**3.77**	**–3.22**	**57.81**	**8.94**	**–8.89**
CMD-PCL (Liu et al., 2026b)		**58.83**	**5.46**	**–5.34**	**57.68**	**9.77**	**–9.77**	**56.16**	**5.76**	**–5.63**	**58.26**	**9.08**	**–9.04**
A-GEM (Chaudhry et al., 2018)	Data samples	**55.88**	**8.75**	**–8.75**	**56.03**	**8.52**	**–8.27**	**55.62**	**5.12**	**–4.67**	**56.64**	**8.82**	**–8.82**
ER (Chaudhry et al., 2019)		**54.69**	**16.70**	**–16.59**	**55.23**	**16.55**	**–16.55**	**54.99**	**12.22**	**–12.22**	**54.63**	**13.19**	**–13.19**
DER (Buzzega et al., 2020)		**54.30**	**13.72**	**–13.64**	**56.20**	**9.56**	**–9.24**	**55.96**	**10.07**	**–10.07**	**56.71**	**8.85**	**–8.81**
DER++ (Buzzega et al., 2020)		**55.50**	**15.54**	**–15.54**	**56.58**	**5.26**	**–5.06**	**55.76**	**7.54**	**–7.54**	**56.92**	**7.82**	**–7.82**
CSReL (Tong et al., 2025)		**58.54**	**12.86**	**–12.50**	**56.40**	**7.20**	**–7.20**	**56.28**	**4.02**	**–3.98**	**58.15**	**8.69**	**–8.21**
OCAR (Urettini & Carta, 2025)		**56.20**	**12.08**	**–12.08**	**55.64**	**7.55**	**–7.30**	**55.71**	**5.08**	**–5.04**	**57.10**	**8.36**	**–8.36**
NeuroAccord	**TSF**	**61.95**	**4.39**	**–4.37**	**60.42**	**4.13**	**–3.04**	**59.01**	**2.43**	**–2.17**	**60.86**	**4.69**	**–4.63**

As shown in Table 2, NeuroAccord achieves the highest AAA and BWT and the lowest FOR on all four backbones. Relative to the strongest baseline for each metric, it improves AAA by 4.38%, 2.51%, 3.00%, and 2.35% and reduces FOR by 19.60%, 3.95%, 8.65%, and 22.61% on CBraMod, CSBrain, EEGMamba, and CodeBrain, respectively. Baselines from both views offer less favorable accuracy–retention trade-offs than NeuroAccord across all four backbones.

4.3 EFFECTIVENESS OF THE TSF REFERENCE

We examine the value of TSF from two perspectives in Fig. 4. Panel (a) shows pairwise distances among nine datasets in TSF descriptor space, illustrating their diversity within a common description. To assess this reference in continual learning, panel (b) replaces TSF with seven alternative spaces derived from model parameters or data samples and compares AAA and FOR on CBraMod and EEGMamba. TSF yields the highest AAA and lowest FOR on both backbones, supporting time, space, and frequency structure as a reference for balancing new-task learning and retention.

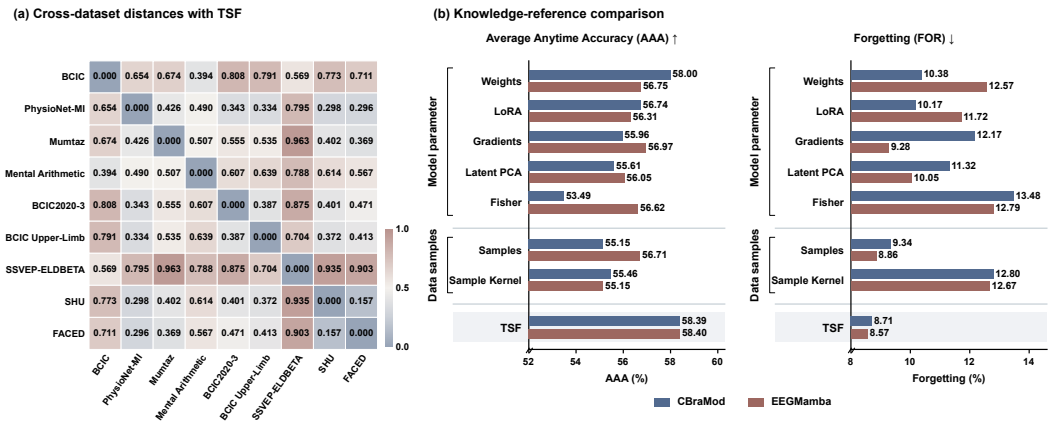


Figure 4: Cross-dataset TSF distances and performance under alternative reference spaces.

4.4 ABLATION STUDY

We examine the contributions of the three consistency objectives and the TSF views on CBraMod and EEGMamba in Tables 3 and 4, respectively. Table 3 shows that the full objective achieves the highest AAA and BWT and the lowest FOR on both backbones. Removing TSFRD or COA reduces AAA and increases FOR, indicating that sample-relation transfer and TSF coupling provide complementary guidance during new-task learning. HKA contributes most strongly to retention among the three objectives: omitting it raises FOR from 4.39% to 8.65% on CBraMod and from 2.43% to 4.28% on EEGMamba, the largest increases among single-objective ablations. Removing any two objectives causes further deterioration, consistent with the three losses preserving prior knowledge.

Table 3: Loss ablation (%).

Method	ℓ_{TSFRD}	ℓ_{COA}	ℓ_{HKA}	CBraMod			EEGMamba		
				AAA	**FOR**	**BWT**	**AAA**	**FOR**	**BWT**
Ours (full)	✓	✓	✓	**61.95**	**4.39**	**−4.37**	**59.01**	**2.43**	**−2.17**
w/o ℓ_{TSFRD}	✗	✓	✓	**61.11**	**4.81**	**−4.79**	**58.35**	**3.50**	**−2.84**
w/o ℓ_{COA}	✓	✗	✓	**60.36**	**5.41**	**−5.36**	**57.95**	**4.10**	**−4.07**
w/o ℓ_{HKA}	✓	✓	✗	**59.61**	**8.65**	**−8.57**	**57.26**	**4.28**	**−4.21**
w/o ℓ_{TSFRD},ℓ_{COA}	✗	✗	✓	**57.91**	**9.00**	**−9.00**	**54.69**	**6.31**	**−5.81**
w/o ℓ_{TSFRD},ℓ_{HKA}	✗	✓	✗	**57.32**	**12.09**	**−12.09**	**54.82**	**6.36**	**−6.10**
w/o ℓ_{COA},ℓ_{HKA}	✓	✗	✗	**56.12**	**11.81**	**−11.81**	**54.38**	**6.62**	**−6.46**

Table 4: TSF-view ablation (%).

Method	T	S	F	CBraMod			EEGMamba		
				AAA	**FOR**	**BWT**	**AAA**	**FOR**	**BWT**
Ours (full)	✓	✓	✓	**61.95**	**4.39**	**−4.37**	**59.01**	**2.43**	**−2.17**
w/o T	✗	✓	✓	**60.70**	**7.23**	**−7.19**	**57.99**	**3.72**	**−3.05**
w/o S	✓	✗	✓	**61.10**	**6.14**	**−6.09**	**58.06**	**4.28**	**−4.21**
w/o F	✓	✓	✗	**60.84**	**7.32**	**−7.28**	**57.92**	**3.77**	**−3.29**
w/o T+S	✗	✗	✓	**60.16**	**8.89**	**−8.67**	**57.09**	**3.50**	**−2.77**
w/o T+F	✗	✓	✗	**60.07**	**7.84**	**−7.84**	**56.75**	**3.89**	**−3.28**
w/o S+F	✓	✗	✗	**59.68**	**8.35**	**−8.17**	**57.24**	**4.26**	**−3.35**

Table 4 compares the contributions of the time, space, and frequency views. The complete TSF reference achieves higher AAA and BWT and lower FOR than every single- or two-view configuration on both backbones. Omitting any one view lowers AAA and increases FOR, although the magnitude varies between backbones. Specifically, the increase in FOR ranges from 1.75 %to 2.93 % on CBraMod and from 1.29 %to 1.85 % on EEGMamba. No individual view or pair of views recovers the full model's performance, supporting their complementary roles in guiding knowledge transfer.

4.5 SENSITIVITY AND ROBUSTNESS ANALYSIS

To assess sensitivity to objective weighting, we vary the TSFRD, COA, and HKA coefficients individually on CBraMod and EEGMamba while holding the others fixed. We also evaluate NeuroAccord across five random seeds and 20 task orders on all four backbones (Fig. 5). Intermediate weights generally provide higher AAA and lower FOR than extreme settings, although the preferred values differ by backbone. This pattern suggests that the three consistency objectives benefit from balanced weighting. Across random seeds and task orders, AAA standard deviations range from 0.38–0.68% and 0.93–1.32%, respectively. Task order introduces more variation than random initialization, but the AAA and FOR results together indicate robust performance across the evaluated sequences.

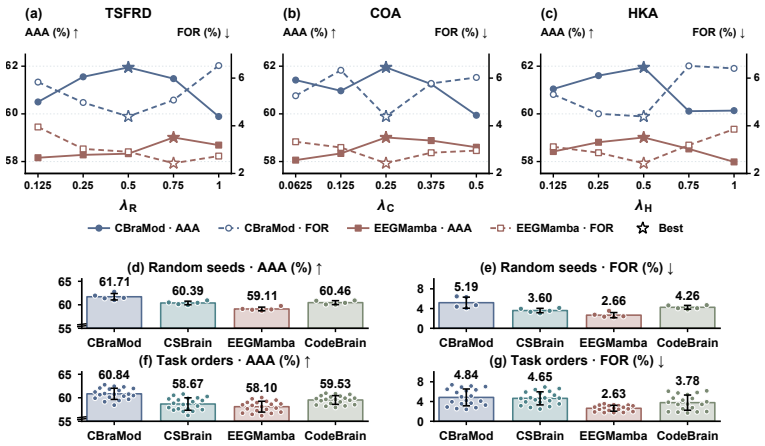


Figure 5: Loss-weight sensitivity (a–c) and robustness across random seeds and task orders (d–g).

4.6 ADAPTATION TO UNSEEN DATASETS

Table 5: Continued learning on the unseen datasets HMC and TUAB.

Method	HMC			TUAB		
	BAcc	κ	W-F1	BAcc	AUROC	PR-AUC
CBraMod (Park et al., 2019)	0.7225	0.7079	0.7743	0.7759	0.8604	0.8683
CSBrain (Zhou et al., 2026)	0.7245	0.6893	0.7657	0.7837	0.8661	0.8663
EEGMamba (Gui et al., 2024)	0.7200	0.7058	0.7721	0.7557	0.8539	0.8569
CodeBrain (Ma et al., 2026)	0.7239	0.6844	0.7586	0.7831	0.8634	0.8692
NeuroLM (Jiang et al., 2025)	0.6998	0.6716	0.7520	0.7609	0.8408	0.8514
Uni-NTFM (Chen et al., 2026)	0.7327	0.7181	0.7829	0.7842	0.8548	0.8317
ECHO (Liu et al., 2026a)	0.7363	0.7164	0.7861	0.7865	0.8582	0.8545
NeuroAccord (CBraMod)	0.7454	0.7207	0.7911	0.7863	0.8579	0.8499
NeuroAccord (CSBrain)	**0.7506**	**0.7296**	**0.7967**	**0.7983**	**0.8720**	0.8553
NeuroAccord (EEGMamba)	0.7346	0.7224	0.7880	0.7855	0.8607	0.8439
NeuroAccord (CodeBrain)	0.7464	0.7245	0.7939	0.7906	0.8672	**0.8731**

We further evaluate NeuroAccord on HMC (Alvarez-Estevez & Rijsman, 2021) and TUAB (de Diego, 2017), introduced as new tasks after the nine-task continual-learning sequence. Table 5 shows consistent improvements in balanced accuracy over each corresponding backbone on both datasets. The results support continued knowledge acquisition on datasets across distinct decoding objectives.

5 RELATED WORK

Continual-learning methods preserve past knowledge differently. EWC and SI constrain important parameters, whereas LwF distills previous predictions (Kirkpatrick et al., 2017; Zenke et al., 2017; Li & Hoiem, 2017). GPM and SGP regulate updates through historical subspaces, while O-LoRA separates low-rank task updates (Saha et al., 2021; Saha & Roy, 2023; Wang et al., 2023). ER and DER replay examples, with DER also retaining logits (Chaudhry et al., 2019; Buzzega et al., 2020).

6 CONCLUSION

In this study, we presented NeuroAccord, a Time–Space–Frequency (TSF)-guided framework linking EEG structure with model parameters. Its relational transfer, TSF coupling, and historical anchoring coordinate adaptation and retention. Across nine datasets, six paradigms, and four backbones, NeuroAccord improves new-task learning and reduces forgetting relative to 18 baselines. More broadly, these findings highlight the value of grounding evolving model representations in physiologically meaningful EEG structure, which provides a common reference across tasks for integrating new knowledge and preserving the signal relationships underlying earlier decoding capabilities.

AI USE STATEMENT

We used LLM to assist with polishing manuscript text, refining scientific figures and organizing the paper. The authors take full responsibility for the accuracy of all claims, citations, figures, and experimental results in the final manuscript, including material prepared with AI assistance.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European conference on computer vision*, pp. 144–161. Springer, 2018.

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Diego Alvarez-Estevez and Roselyne M Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PloS one*, 16(8):e0256111, 2021.

Frederik Benzing. Unifying importance based regularisation methods for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2372–2396. PMLR, 2022.

EBS Brain. Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: A review. *Inter-and Intra-subject Variability in Brain Imaging and Decoding*, 8722, 2022.

Gyorgy Buzsaki and Andreas Draguhn. Neuronal oscillations in cortical networks. *science*, 304 (5679):1926–1929, 2004.

György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420, 2012.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

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

Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences*, 112(44):E6058–E6067, 2015.

Zhisheng Chen, Yingwei Zhang, Qizhen Lan, Tianyu Liu, Huacan Wang, Yi Ding, Ziyu Jia, Ronghao Chen, Kun Wang, and Xinliang Zhou. Uni-ntfm: A unified foundation model for eeg signal representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 109178–109210, 2026.

Silvia López de Diego. *Automated interpretation of abnormal adult electroencephalograms*. Temple University, 2017.

Kejia Fan, Yajiang Huang, Jingyu He, Feijiang Han, Jianheng Tang, Huiping Zhuang, Anfeng Liu, Tian Wang, Mianxiong Dong, Houbing Herbert Song, et al. Calm: A ubiquitous crowdsourced analytic learning mechanism for continual service construction with data privacy preservation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(2):1–30, 2025.

Yiyu Gui, MingZhi Chen, Yuqi Su, Guibo Luo, and Yuchao Yang. Eegmamba: Bidirectional state space model with mixture of experts for eeg multi-task classification. *arXiv preprint arXiv:2407.20254*, 2024.

Md Junayed Hasan and Jong-Myon Kim. A hybrid feature pool-based emotional stress state detection algorithm using eeg signals. *Brain sciences*, 9(12):376, 2019.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30534–30544. IEEE, 2025.

Ji-Hoon Jeong, Jeong-Hyun Cho, Young-Eun Lee, Seo-Hyun Lee, Gi-Hwan Shin, Young-Seok Kweon, José del R Millán, Klaus-Robert Müller, and Seong-Whan Lee. 2020 international brain–computer interface competition: A review. *Frontiers in human neuroscience*, 16:898300, 2022.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Jun-You Lin, Ying Choon Wu, and Tzyy-Ping Jung. The identity trap in eeg foundation models: A diagnostic audit. *arXiv preprint arXiv:2606.06647*, 2026.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. *arXiv preprint arXiv:2202.02931*, 2022.

Bingchuan Liu, Yijun Wang, Xiaorong Gao, and Xiaogang Chen. eldbeta: A large eldercare-oriented benchmark database of ssvep-bci for the aging population. *Scientific data*, 9(1):252, 2022.

Chenyu Liu, Yuqiu Deng, Tianyu Liu, Jinan Zhou, Xinliang Zhou, Ziyu Jia, and Yi Ding. Echo: Toward contextual seq2seq paradigms in large eeg models. In *International Conference on Learning Representations*, volume 2026, pp. 153809–153842, 2026a.

Yu Liu, Haoqin Yang, Jinping Sui, Hui Wang, Haipeng Li, Weimin Wang, and Qi Jia. Time series class-incremental learning via confidence-guided mask distillation and prototype-guided contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23917–23925, 2026b.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. Codebrain: Bridging decoupled tokenizer and multi-scale architecture for eeg foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 83526–83574, 2026.

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.

Wajid Mumtaz. Mdd patients and healthy controls eeg data (new). figshare. dataset, 2016.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Zhongzheng Qiao, Minghui Hu, Xudong Jiang, Ponnuthurai Nagaratnam Suganthan, and Ramasamy Savitha. Class-incremental learning on multivariate time series via shape-aligned temporal distillation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Herbert Ramoser, Johannes Muller-Gerking, and Gert Pfurtscheller. Optimal spatial filtering of single trial eeg during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4): 441–446, 2000.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Gobinda Saha and Kaushik Roy. Continual learning with scaled gradient projection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 9677–9685, 2023.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51(6):1034–1043, 2004.

M Tangermann et al. Review of the bci competition iv. frontiers in neuroscience. *Benchmarking study on BCI algorithms*, 2012.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *International Conference on Learning Representations*, volume 2025, pp. 13964–13999, 2025.

Edoardo Urettini and Antonio Carta. Online curvature-aware replay: Leveraging $\mathbf{2^{nd}}$ order information for online continual learning, 2025.

Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9365–9374. IEEE, 2021.

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. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. In *International Conference on Learning Representations*, volume 2025, pp. 58337–58353, 2025.

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

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *Advances in Neural Information Processing Systems*, 38: 87150–87195, 2026.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25613–25622. IEEE, 2025.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A DETAILED CONSTRUCTION OF TSF

Time–Space–Frequency (TSF) describes EEG through temporal dynamics, multichannel relationships, and spectral composition. These complementary properties support electrophysiological analysis and task decoding. Below, we define the sixteen coordinates and connect their construction to representative studies.

Let $x \in \mathbb{R}^{C\times P\times S}$ be a supplied EEG example, where C denotes the number of channels, P the number of consecutive patches, and S the number of samples per patch. The value x_{cps} is the sample at channel c, patch p, and within-patch position s, and f_s is the sampling rate in Hz. We assume $C, P, S \geq 2$ and a sampling rate sufficient to resolve the frequency bands below. The resulting descriptor is

$$\phi(x) = \big[\phi^T(x), \phi^S(x), \phi^F(x)\big], \qquad (d_T, d_S, d_F) = (6, 5, 5). \tag{14}$$

Each example is summarized independently using its signal values alone. The formulas below are written for nonzero normalization denominators. The implementation floors amplitude-scale denominators at 10^{-8} and eigenvalue- and band-power normalization denominators at 10^{-12}.

A.1 TIME VIEW

The time view separates amplitude-distribution properties from sample-order information. This distinction has an electrophysiological basis. Simulations and EEG analysis show that temporal offsets between neuronal sources alter the waveform and harmonic content of their mixture. We summarize the observed dynamics through amplitude concentration, higher-order moments, lagged differences, and sign alternation.

For a quantity a defined on the input tensor, write $\langle a\rangle = (CPS)^{-1}\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{s=1}^{S} a_{cps}$. Define the pooled mean, standard deviation, and standardized samples as

$$\mu = \langle x\rangle, \qquad \sigma = \sqrt{\langle (x-\mu)^2\rangle}, \qquad u_{cps} = \frac{x_{cps}-\mu}{\sigma}. \tag{15}$$

Both statistics are pooled over the complete example.

Amplitude concentration. The shape factor T_1 compares root-mean-square and mean absolute amplitudes to describe how strongly large excursions contribute to the recording,

$$T_1 = \frac{\sqrt{\langle x^2\rangle}}{\langle |x|\rangle}. \tag{16}$$

The ratio is invariant to a common nonzero amplitude gain and satisfies $T_1 \geq 1$, with equality for constant absolute amplitude. In an EEG application, Hasan et al. combine shape factor, skewness, and kurtosis with wavelet features, followed by feature selection and nearest-neighbor classification of DEAP calm/stress labels derived from valence and arousal ratings. This motivates retaining amplitude-distribution shape in the time view.

Amplitude asymmetry and large excursions. To complement the shape factor with signed asymmetry and sensitivity to large deviations from the mean, we retain the standardized third and fourth moments,

$$T_2 = \langle u^3\rangle, \qquad T_3 = \langle u^4\rangle. \tag{17}$$

Skewness T_2 distinguishes positive and negative amplitude asymmetry, while kurtosis T_3 emphasizes large standardized excursions. We use the unadjusted fourth standardized moment, giving a Gaussian population reference of three. These statistics distinguish amplitude distributions that have similar variance, including differences associated with transient deflections superimposed on ongoing activity. Related features occur in affective-state and sleep-stage analysis. Together, T_2 and T_3 characterize asymmetry and tail weight in the pooled amplitude distribution.

Patch-lag variation. Temporal ordering complements amplitude-distribution information. Its physiological relevance is illustrated by Sherman et al., who identify brief human MEG beta events and link their waveform to coordinated dendritic inputs using biophysical modeling and animal

laminar recordings. Motivated by temporal structure beyond average power, we measure normalized variation at a patch lag,

$$T_4 = \frac{1}{\sigma}\sqrt{\frac{\sum_{c=1}^{C}\sum_{p=2}^{P}\sum_{s=1}^{S}(x_{cps} - x_{c,p-1,s})^2}{C(P-1)S}}. \tag{18}$$

The coordinate compares corresponding positions in neighboring patches, separated by S/f_s seconds. Larger values indicate greater change at this lag relative to overall amplitude variation, providing a summary of temporal change at the supplied patch lag.

Local variation. To complement this longer-lag summary, the local variation T_5 measures changes between successive samples within each patch,

$$T_5 = \frac{1}{\sigma}\sqrt{\frac{\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{s=2}^{S}(x_{cps} - x_{c,p,s-1})^2}{CP(S-1)}}. \tag{19}$$

This measures local roughness relative to amplitude dispersion, following the derivative-sensitive view of EEG dynamics. For example, Vidaurre et al. evaluate log-variance features of the signal and its derivatives on motor-imagery EEG from 75 participants, using spatial filtering and regularized linear discriminant analysis. TSF captures local variation through a normalized first-difference summary. The dimensionless T_5 measures change per sample and complements the longer-lag variation in T_4.

Sign alternation. The final time-view coordinate is the zero-crossing fraction,

$$T_6 = \frac{1}{CP(S-1)}\sum_{c=1}^{C}\sum_{p=1}^{P}\sum_{s=2}^{S}\mathbf{1}\{x_{cps}x_{c,p,s-1} < 0\}. \tag{20}$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function. Each strictly opposite-sign pair contributes equally to T_6. Period-based EEG analysis relates crossing information to spectral moments, and zero-crossing features also appear in sleep-stage classification. Crossings are measured relative to zero in the original supplied samples. Pairs containing zero contribute zero to the numerator, and patch boundaries are excluded. Thus, T_6 is the fraction of within-patch sample pairs with opposite signs.

Together, $\phi^T(x) = (T_1, \ldots, T_6)$ separates amplitude-distribution properties from temporal variation, allowing recordings with similar overall amplitude to differ in their time-view descriptors.

A.2 Space view

Multichannel covariance captures task-relevant relationships between EEG channels. For example, Ramoser et al. learn common spatial patterns from left- and right-hand motor-imagery covariance matrices and classify their contrasting projected variances. TSF summarizes second-order spatial structure through two eigenvalue statistics for concentration across spatial modes and three correlation quantiles for pairwise dependence.

Write each channel as $w_c \in \mathbb{R}^N$, where $N = PS$ and a indexes samples. Channel-wise centering and amplitude normalization give

$$\bar{w}_c = \frac{1}{N}\sum_{a=1}^{N} w_{ca}, \qquad \sigma_c = \sqrt{\frac{1}{N}\sum_{a=1}^{N}(w_{ca} - \bar{w}_c)^2}, \qquad v_{ca} = \frac{w_{ca} - \bar{w}_c}{\sigma_c}. \tag{21}$$

For $V = [v_{ca}] \in \mathbb{R}^{C\times N}$, the correlation matrix and normalized eigenvalues are

$$R = \frac{VV^\top}{N}, \qquad \pi_c = \frac{\lambda_c}{\sum_{j=1}^{C}\lambda_j}, \qquad 0 \le \lambda_1 \le \cdots \le \lambda_C. \tag{22}$$

Here λ_c are the eigenvalues of R, a Pearson correlation matrix for nonconstant channels. Its entries describe zero-lag linear dependence, and its eigenvalues describe variance along spatial modes. Channel-wise normalization emphasizes relative dependence, consistent with the use of correlation in EEG spatial-complexity research.

Spatial complexity. The normalized effective rank S_1 measures how broadly standardized channel variance is distributed across spatial modes,

$$S_1 = \frac{1}{C} \exp\left(-\sum_{c=1}^{C} \pi_c \log \pi_c \right), \qquad 0 \log 0 = 0. \tag{23}$$

This follows the eigenvalue-entropy principle of Omega complexity, applied to correlation and divided by C. A single mode gives $S_1 = 1/C$, whereas equal variance across all C modes gives $S_1 = 1$. The resulting value summarizes the effective dimensionality of recorded channel activity. In an overnight EEG example, Wackermann et al. compare segment-wise Omega complexity with scored sleep stages and observe lower complexity during slow-wave sleep. In a visual Go/Nogo task, Jia et al. associate higher single-trial complexity with shorter N1/N2 latencies and smaller N1/N2/P3 amplitudes. These examples connect spatial-mode diversity to brain state and task-evoked processing.

Leading-mode dominance. The leading variance fraction measures the dominance of the strongest channel combination,

$$S_2 = \pi_C. \tag{24}$$

Larger S_2 indicates stronger concentration in one mode. It complements S_1, which also depends on the distribution of the remaining variance. In an in-vitro hippocampal seizure model, Li et al. use correlation eigendecomposition and surrogate comparisons to identify synchronized clusters and increasing collective activity. In TSF, the leading variance fraction summarizes the dominance of shared recorded activity.

Pairwise dependence. To summarize dependence strengths across channel pairs, define the multiset $\mathcal{O} = \{|R_{cc'}| : c \neq c'\}$ and its three quantiles,

$$S_3 = \mathrm{Quantile}_{0.50}(\mathcal{O}), \qquad S_4 = \mathrm{Quantile}_{0.75}(\mathcal{O}), \qquad S_5 = \mathrm{Quantile}_{0.90}(\mathcal{O}). \tag{25}$$

The median S_3 summarizes typical dependence, S_4 its upper-middle range, and S_5 its upper tail. These selected quantiles provide a compact summary of the distribution of sensor-level dependence. Absolute values retain correlation strength across both polarities, and diagonal exclusion removes self-correlations. Linear interpolation uses all $C(C-1)$ off-diagonal entries, including both orientations of each pair.

The fixed-width vector $\phi^S(x) = (S_1, \ldots, S_5)$ is invariant to reordering the same channels, which leaves the eigenvalues and $\mathcal{O}$ unchanged. These descriptors summarize sensor-level dependence, which is shaped by electrode coverage, reference choice, and volume conduction. Numerically, we symmetrize R, clip roundoff-induced negative eigenvalues to zero, and floor the logarithm argument at 10^{-12}.

A.3 Frequency view

The frequency view describes the allocation of EEG power across physiologically relevant timescales. Spectral structure supports diverse decoding tasks. For example, Chen et al. identify attended characters from SSVEP responses to frequency–phase-coded visual stimuli. TSF summarizes spectral composition through five relative-power coordinates.

For each channel and patch, compute the nonnegative-frequency Fourier coefficients

$$X_{cpk} = \sum_{s=0}^{S-1} x_{cp,s+1} e^{-2\pi \mathrm{i} k s / S}, \qquad f_k = \frac{k f_s}{S}, \qquad k = 0, \ldots, \lfloor S/2 \rfloor, \tag{26}$$

where i is the imaginary unit and f_k is the frequency of bin k. The transform is applied directly to each supplied patch, without additional detrending or tapering. We use $\mathcal{B}_\delta = [1, 4)$, $\mathcal{B}_\theta = [4, 8)$, $\mathcal{B}_\alpha = [8, 13)$, $\mathcal{B}_\beta = [13, 30)$, and $\mathcal{B}_\gamma = [30, 40]$, all in Hz. With $\mathcal{B} = \{\delta, \theta, \alpha, \beta, \gamma\}$, define

$$P_b = \sum_{c=1}^{C} \sum_{p=1}^{P} \sum_{k: f_k \in \mathcal{B}_b} |X_{cpk}|^2, \qquad F_b = \frac{P_b}{\sum_{b' \in \mathcal{B}} P_{b'}}, \qquad b \in \mathcal{B}. \tag{27}$$

Power is pooled across channels and patches before normalization over 1–40 Hz. Thus $\phi^F(x) = (F_\delta, F_\theta, F_\alpha, F_\beta, F_\gamma)$ describes relative spectral composition, with nonnegative entries summing to one when the power denominator is nonzero and its numerical floor is inactive.

Delta activity. F_δ retains the slow contribution in 1–4 Hz. Harmony et al. compare mental calculation with a visually matched control and observe increased low-frequency power during calculation. Delta activity also increases with short-term-memory load. These contrasts motivate retaining slow spectral content associated with internal processing.

Theta activity. F_θ measures relative power in 4–8 Hz. In verbal and spatial working-memory tasks, Gevins et al. use high-resolution EEG to show that greater memory load accompanies increased frontal-midline theta and reduced parietocentral alpha. Their different responses motivate retaining distinct coordinates for the two bands. Pooling yields an example-level summary of their relative spectral contributions.

Alpha-range activity. F_α covers 8–13 Hz, including posterior alpha and overlapping sensorimotor mu activity. Their interpretation depends on recording location and task. A physiological link appears in Haegens et al., who record local field potentials and spikes in a monkey performing vibrotactile discrimination. Spiking varies with alpha phase, and lower alpha power accompanies higher firing rates and better discrimination. These findings motivate retaining alpha-range activity in the pooled spectral description.

Beta activity. F_β measures relative power in 13–30 Hz. Pfurtscheller et al. observe reduced mu and beta power over the contralateral sensorimotor region during imagined hand movement. This event-related desynchronization links band-power modulation to motor imagery. Retaining both F_α and F_β describes their relative contributions and captures task-dependent shifts in spectral composition.

Lower-gamma activity. F_γ retains the 30–40 Hz contribution. Tallon-Baudry et al. use time–frequency analysis to identify stronger non-phase-locked responses near 40 Hz for coherent than noncoherent visual stimuli. In emotion recognition, Zheng et al. classify positive, neutral, and negative states using bandwise differential entropy, finding beta and gamma informative and the five-band combination most effective in their study. These examples motivate frequency-resolved descriptions. TSF uses relative power and a narrower gamma interval than their 31–50 Hz emotion features.

Relative-power normalization removes a common amplitude gain, and each coordinate depends on both its own band and the normalization total. Band power also includes periodic and aperiodic contributions. The frequency view therefore summarizes their combined spectral composition. Together, the three views retain complementary properties of EEG dynamics, multichannel organization, and spectral content.

B Mathematical analysis of the consistency objectives

We analyze the complementary constraints imposed by TSFRD, COA, and HKA using the notation of Section 3. The previous model, TSF descriptors, and selected TSF bases are fixed during each stage. The derivations concern the main-text objectives with $n \geq 2$, nonzero normalization denominators, and, for HKA, $q_{t-1} > 0$ and $\nu_{t-1} > 0$.

B.1 Relational geometry and TSF coupling

TSFRD and COA retain two aspects of the previous model's representations. The former preserves relations among examples, while the latter preserves how feature coordinates co-vary with the shared TSF reference.

TSF-weighted sample geometry. TSFRD preserves the angular organization of examples within the span of the TSF basis. To make the role of TSF explicit, combine the view-wise normalized kernels into

$$W_{ij} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{K^v_{ij}}{\sum_{p \neq q} K^v_{pq}} \quad (i \neq j), \qquad W_{ii} = 0. \tag{28}$$

These weights sum to one and assign equal total mass to each view. Writing $\overline{H}^{(s)}$ for the row-normalized $H^{(s)}$, the relation matrix is $G^{(s)} = \overline{H}^{(s)}(\overline{H}^{(s)})^\top$ for $s \in \{t-1, t\}$. Equation (7) becomes

$$\mathcal{L}_{\mathrm{TSFRD}} = \left\| W^{\odot 1/2} \odot \left(G^{(t)} - G^{(t-1)} \right) \right\|_F^2, \tag{29}$$

where $\odot$ and $W^{\odot 1/2}$ denote elementwise multiplication and square root. The previous model supplies the geometry to preserve, while TSF controls the relative penalty on each relation. Positive kernel weights give $\mathcal{L}_{\mathrm{TSFRD}} = 0$ exactly when the cosine matrices agree. Their magnitudes prioritize discrepancies during optimization.

The objective is invariant to positive row scaling and common orthogonal transformations. For any positive diagonal $D \in \mathbb{R}^{n \times n}$ and orthogonal $O \in \mathbb{R}^{r \times r}$, $G(DHO) = G(H)$. Consequently, identical sample relations can coexist with different feature orientations relative to TSF. The next objective constrains this reference-dependent information.

Directional TSF coupling. The cross-covariance in Eq. (8) measures how selected feature coordinates vary with TSF descriptors across aligned examples. Write $M_t^v = M^v(H^{(t)})$ and $M_{t-1}^v = M^v(H^{(t-1)})$, and use hats for their Frobenius-normalized forms. With $\langle A, B \rangle_F = \mathrm{tr}(A^\top B)$,

$$\begin{aligned} \frac{1}{2}\|\widehat{M}_t^v - \widehat{M}_{t-1}^v\|_F^2 &= 1 - \frac{\langle M_t^v, M_{t-1}^v \rangle_F}{\|M_t^v\|_F \, \|M_{t-1}^v\|_F} \\ &= 1 - \cos\big(\mathrm{vec}(M_t^v), \mathrm{vec}(M_{t-1}^v)\big), \end{aligned} \tag{30}$$

where vec stacks matrix entries. Thus, the factor $1/2$ makes each view's contribution a cosine distance, and COA averages directional disagreement across views. Its zero-loss condition is

$$\mathcal{L}_{\mathrm{COA}} = 0 \quad \Longleftrightarrow \quad M_t^v = a_v M_{t-1}^v \quad \text{for some } a_v > 0 \text{ for each } v \in \mathcal{V}. \tag{31}$$

The objective preserves the relative pattern and signs of linear associations with TSF while allowing each view's overall coupling strength to change. This separates the association structure used for knowledge transfer from its magnitude.

B.2 Complementarity and reference-dependent orientation

Reference-dependent orientation. Consider $H^{(t)} = H^{(t-1)}O$ with $O^\top O = I_r$ and the same basis U_t and descriptors for both models. TSFRD is zero because orthogonal transformations preserve sample cosine relations. In contrast,

$$M_t^v = O^\top M_{t-1}^v, \qquad \widehat{M}_t^v = O^\top \widehat{M}_{t-1}^v, \tag{32}$$

because centering commutes with right multiplication and orthogonal transformations preserve the Frobenius norm. To characterize the resulting constraint, define the coupling matrix $S = |\mathcal{V}|^{-1} \sum_v \widehat{M}_{t-1}^v (\widehat{M}_{t-1}^v)^\top \in \mathbb{R}^{r \times r}$. It is positive semidefinite with $\mathrm{tr}(S) = 1$ and summarizes how normalized TSF coupling is distributed across selected feature directions. Let $S = \sum_{a=1}^r \mu_a u_a u_a^\top$ be its eigendecomposition, with orthonormal u_a and eigenvalues $\mu_a \geq 0$. Substitution into COA yields

$$\begin{aligned} \mathcal{L}_{\mathrm{COA}} &= \frac{1}{2|\mathcal{V}|} \sum_v \|(O^\top - I_r)\widehat{M}_{t-1}^v\|_F^2 \\ &= \frac{1}{2}\,\mathrm{tr}\big[(O^\top - I_r) S (O - I_r)\big] \\ &= \frac{1}{2} \sum_{a=1}^r \mu_a \|(O^\top - I_r)u_a\|_2^2. \end{aligned} \tag{33}$$

Since $\sum_a \|(O^\top - I_r)u_a\|_2^2 = \|O - I_r\|_F^2$, the smallest and largest eigenvalues, $\mu_{\min}$ and $\mu_{\max}$, bound the orientation penalty as

$$\frac{\mu_{\min}}{2}\|O - I_r\|_F^2 \leq \mathcal{L}_{\mathrm{COA}} \leq \frac{\mu_{\max}}{2}\|O - I_r\|_F^2. \tag{34}$$

These expressions quantify the information added beyond sample geometry. COA penalizes orientation changes according to their measured TSF coupling. If S is positive definite, zero COA requires $O = I_r$ within this transformation family. More generally, the penalty vanishes exactly when $O^\top u = u$ for every $u \in \mathrm{range}(S)$. The reference therefore anchors the coupled directions while preserving freedom in uncoupled directions.

Nonredundancy of the two objectives. The distinction operates in both directions. Taking $O = -I_r$ gives $\mathcal{L}_{\mathrm{TSFRD}} = 0$ and $\mathcal{L}_{\mathrm{COA}} = 2$. Conversely, a common shift $H^{(t)} = H^{(t-1)} + \mathbf{1}a^\top$ leaves $M_t^v = M_{t-1}^v$ because $J_n\mathbf{1} = 0$. COA is therefore zero while cosine relations can change. For an explicit single-view example, let the rows of $H^{(t-1)} = \Phi^v$ be $e_1^\top, -e_1^\top, e_2^\top, -e_2^\top$, where e_1, e_2 are the standard basis of $\mathbb{R}^2$. Choosing $a = e_2/2$ changes G_{12} from -1 to $-3/5$, giving $\mathcal{L}_{\mathrm{TSFRD}} \geq (4/25)(W_{12} + W_{21}) > 0$ with $\mathcal{L}_{\mathrm{COA}} = 0$. Thus, neither zero-loss condition implies the other. Together, the objectives preserve both sample geometry and TSF coupling, while retaining a common positive feature-scaling freedom.

B.3 HKA AS A SELECTIVE QUADRATIC CONSTRAINT

Selective feature discrepancy. Let $P_{t-1} = Q_{t-1}Q_{t-1}^\top$ be the orthogonal projector onto the span of the accumulated historical TSF basis. Since Q_{t-1} has orthonormal columns, the norm of the projected coordinates equals that of the projected feature error. HKA therefore admits the equivalent form

$$\mathcal{L}_{\mathrm{HKA}} = \frac{\|(F^{(t)} - F^{(t-1)})P_{t-1}\|_F^2}{mq_{t-1}\nu_{t-1}}. \tag{35}$$

The orthogonal decomposition of feature drift then gives

$$\|F^{(t)} - F^{(t-1)}\|_F^2 = mq_{t-1}\nu_{t-1}\mathcal{L}_{\mathrm{HKA}} + \|(F^{(t)} - F^{(t-1)})(I_d - P_{t-1})\|_F^2. \tag{36}$$

This decomposition specifies the scope of preservation. HKA penalizes the historical component of feature drift, leaving the complementary component outside its direct constraint. Its value depends on the retained span, independently of the orthonormal basis used to represent it. The denominator averages drift over tokens and retained coordinates and expresses it relative to the previous model's feature variance.

Selective feature gradient. The teacher variance ν_{t-1} is fixed with respect to $F^{(t)}$. Differentiating the projected quadratic gives

$$\begin{aligned}
\mathrm{d}\mathcal{L}_{\mathrm{HKA}} &= \frac{1}{mq_{t-1}\nu_{t-1}}\,\mathrm{d}\left\|(F^{(t)} - F^{(t-1)})Q_{t-1}\right\|_F^2 \\
&= \frac{1}{mq_{t-1}\nu_{t-1}}\,\mathrm{d}\,\mathrm{tr}\left[(F^{(t)} - F^{(t-1)})P_{t-1}(F^{(t)} - F^{(t-1)})^\top\right] \\
&= \frac{2}{mq_{t-1}\nu_{t-1}}\left\langle (F^{(t)} - F^{(t-1)})P_{t-1}, \mathrm{d}F^{(t)}\right\rangle_F .
\end{aligned} \tag{37}$$

Identifying the gradient and using $P_{t-1}(I_d - P_{t-1}) = 0$ gives

$$\nabla_{F^{(t)}}\mathcal{L}_{\mathrm{HKA}} = \frac{2(F^{(t)} - F^{(t-1)})P_{t-1}}{mq_{t-1}\nu_{t-1}}, \qquad (\nabla_{F^{(t)}}\mathcal{L}_{\mathrm{HKA}})(I_d - P_{t-1}) = 0. \tag{38}$$

This establishes Eq. (**??**) and shows how HKA restricts its direct feature gradient to historical directions. As Q_{t-1} accumulates directions across tasks, the same objective extends this selective consistency to the learning history. These identities describe the current-token feature objective; its realization through shared-parameter updates also depends on the model Jacobian, the optimizer, and the other training objectives.

B.4 WHEN DOES CURRENT-TASK ANCHORING LIMIT FORGETTING?

The preceding identities alone do not bound performance on an earlier task. Indeed, if the current and historical input supports do not overlap, a model can agree with its teacher on every current input while changing its predictions arbitrarily on historical inputs. Then HKA is zero but forgetting can be complete. We therefore state explicit conditions under which current-task anchoring transfers to historical performance.

Fix an earlier task $j < t$. Let $\mathcal{D}_j^{\mathrm{bal}}$ assign equal probability to each class of task j, and let μ_j be its input marginal. Let μ_t be the current-task training input distribution. The fixed task-j head produces

logit vector $s_j^{(u)}(x)$ from the backbone at stage u, and $R_j^{(u)}$ denotes its balanced classification error. Write $P = P_{t-1}$ and define the root-mean-square historical-direction drift on one input by

$$D_t(x) = \left[\frac{\|(F^{(t)}(x) - F^{(t-1)}(x))P\|_F^2}{L(x)} \right]^{1/2}, \qquad \mathcal{E}_t = \mathbb{E}_{x\sim\mu_t} D_t(x)^2. \tag{39}$$

Here $L(x)$ is the number of feature tokens. The quantity $\mathcal{E}_t$ is the population counterpart of the projected numerator in HKA, averaged per token. When token counts are fixed within a task, as in our task-specific inputs, its minibatch counterpart equals $q_{t-1}\nu_{t-1}\mathcal{L}_{\mathrm{HKA}}$.

We make two dependencies explicit. First, suppose the old head's logit change obeys

$$\|s_j^{(t)}(x) - s_j^{(t-1)}(x)\|_\infty \le L_j D_t(x) + \varepsilon_{j,t}^{\perp} \quad \text{for } \mu_j\text{-almost every } x. \tag{40}$$

The constant L_j describes sensitivity to historical directions, while $\varepsilon_{j,t}^{\perp}$ bounds logit change not explained by them. For a fixed linear head on mean-pooled features, these terms can be bounded using its operator norms on the ranges of P and $I - P$, respectively. Second, let h_t be a K_t-Lipschitz function of the common TSF descriptor $\phi(x)$, and define its approximation error for projected drift as

$$\delta_{j,t} = \mathbb{E}_{\mu_j}|D_t(x) - h_t(\phi(x))| + \mathbb{E}_{\mu_t}|D_t(x) - h_t(\phi(x))|. \tag{41}$$

The residual is small only when TSF describes the aspects of the signals that govern projected feature drift; this is not implied merely by constructing a TSF basis.

Proposition (conditional forgetting bound). Let $M_{j,t-1}(x,y) = s_{j,y}^{(t-1)}(x) - \max_{c\neq y} s_{j,c}^{(t-1)}(x)$ be the previous model's correct-class margin, and set $p_{j,t}(\gamma) = \Pr_{\mathcal{D}_j^{\mathrm{bal}}}\{0 \le M_{j,t-1}(x,y) \le \gamma\}$. For any $\gamma > 0$ and finite first moments of the TSF distributions, the one-stage increase in balanced error satisfies

$$\begin{aligned} R_j^{(t)} - R_j^{(t-1)} \le\ & p_{j,t}(\gamma) \\ & + \frac{2}{\gamma}\left[L_j\left(\sqrt{\mathcal{E}_t} + K_t W_1(\phi_\sharp\mu_j, \phi_\sharp\mu_t) + \delta_{j,t} \right) + \varepsilon_{j,t}^{\perp} \right]. \end{aligned} \tag{42}$$

Here W_1 is the 1-Wasserstein distance between the two distributions of TSF descriptors. To verify the bound, an example that the previous model classifies correctly with margin greater than γ remains correct whenever its logit change is below $\gamma/2$. Markov's inequality thus bounds the additional error by $p_{j,t}(\gamma) + (2/\gamma)\mathbb{E}_{\mu_j}\|s_j^{(t)} - s_j^{(t-1)}\|_\infty$. Equation (40) controls the logit term. For any coupling of $\phi_\sharp\mu_j$ and $\phi_\sharp\mu_t$, the Lipschitz property of h_t and Eq. (41) give

$$\mathbb{E}_{\mu_j} D_t \le \mathbb{E}_{\mu_t} D_t + K_t W_1(\phi_\sharp\mu_j, \phi_\sharp\mu_t) + \delta_{j,t} \le \sqrt{\mathcal{E}_t} + K_t W_1(\phi_\sharp\mu_j, \phi_\sharp\mu_t) + \delta_{j,t}, \tag{43}$$

which completes the argument.

Let $b_{j,t}(\gamma)$ denote the right-hand side of Eq. (42). Since population balanced accuracy is $1 - R_j^{(t)}$, telescoping the one-stage inequality yields an upper bound for the population analogue of the paper's forgetting metric:

$$\mathrm{FOR}_{\mathrm{pop}} \le \frac{1}{T-1} \sum_{j=1}^{T-1} \sum_{t=j+1}^{T} b_{j,t}(\gamma_{j,t}) \qquad (\gamma_{j,t} > 0). \tag{44}$$

This is a conditional, not an unconditional, guarantee. HKA directly reduces the current-input term $\mathcal{E}_t$, whereas large cross-task TSF shift, weak TSF prediction of projected drift, off-subspace head dependence, or small old-task margins can make the bound loose. The reported validation FOR additionally has finite-sample variation, which is not included in this population bound. The HKA ablation in Table 3 provides empirical evidence of retention; the bound clarifies the conditions under which its current-task constraint can explain that effect.

C Datasets, backbones, and baseline methods

C.1 Datasets

The benchmark comprises nine datasets covering six EEG decoding paradigms, with each dataset forming one task in the continual-learning sequence. The sequence spans motor imagery, mental

disorder diagnosis, mental stress detection, imagined speech, SSVEP, and emotion recognition. Table 1 summarizes the task order and model-input configurations. Below, we describe the source datasets and the decoding problems used in our study.

BCIC-IV-2a. BCI Competition IV Dataset 2a contains recordings from nine participants performing four motor-imagery tasks involving the left hand, right hand, both feet, and tongue. The acquisition uses 22 EEG channels over two sessions on different days. We retain the four-class formulation and use four-second inputs, providing a multi-class motor-imagery task at the beginning of the sequence.

PhysioNet-MI. The PhysioNet EEG Motor Movement/Imagery Dataset contains 64-channel recordings from 109 participants. Its recording runs include executed and imagined movements involving unilateral fists or bilateral hands and feet. Our benchmark uses a four-class motor-imagery configuration with four-second inputs. This task introduces a different montage and participant population while retaining motor imagery as the decoding paradigm.

Mumtaz. The Mumtaz dataset provides resting-state EEG from participants with major depressive disorder and healthy controls, with 64 participants in the source population. We use 19-channel, five-second inputs for binary classification. In contrast to the preceding movement-imagery tasks, the prediction target describes clinical group membership rather than an instructed mental action.

Mental Arithmetic. This dataset records EEG before and during serial subtraction in 36 participants. The arithmetic condition introduces sustained cognitive demand relative to the resting condition. Our benchmark uses five-second inputs in a binary mental-stress detection setting, adding a task concerned with cognitive state rather than movement intention or clinical diagnosis.

BCIC2020-3. Track 3 of the 2020 International BCI Competition provides imagined-speech recordings from 15 participants. The five labels correspond to the imagined expressions Hello, Help me, Stop, Thank you, and Yes. We use 64-channel, three-second inputs to distinguish these commands, extending the sequence to covert language-related activity.

BCIC-UpperLimb. Track 4 of the same competition provides recordings associated with three upper-limb grasping actions, namely cylindrical, spherical, and lumbrical grasping, from 15 participants. The source protocol includes both movement execution and imagery. Our configuration extracts the four-second imagery interval from each trial and evaluates three-class grasp-imagery recognition. This distinction specifies the task used here without treating the entire source collection as imagery-only data.

eldBETA. eldBETA is an SSVEP benchmark with 100 participants from an aging population and nine visual targets. Target-specific periodic stimulation induces frequency-locked responses for classification. We retain the nine-target setting and use 3.2-second inputs. This task introduces externally driven visual responses alongside the self-generated imagery and state-related activity represented by the other datasets.

SHU-MI. SHU-MI contains left-hand and right-hand motor-imagery recordings from 25 participants, with five recording sessions on separate days and 32 EEG channels. We use four-second inputs for binary classification. Its repeated-session acquisition adds temporal recording variability within a motor-imagery paradigm already represented elsewhere in the sequence.

FACED. The Finer-grained Affective Computing EEG Dataset records 32-channel EEG from 123 participants watching 28 emotion-eliciting video clips. Its nine categories comprise amusement, inspiration, joy, tenderness, anger, fear, disgust, sadness, and neutrality. We use ten-second inputs and retain all nine labels, introducing finer-grained affective decoding as the final task.

Together, these datasets vary in decoding targets, recording layouts, participant populations, and input duration. They therefore support continual adaptation across heterogeneous EEG tasks rather than only incremental expansion of one shared label space. Each task retains its own prediction head, selected by task identity at evaluation.

C.2 EEG FOUNDATION-MODEL BACKBONES

We consider four pretrained EEG foundation models spanning two architectural families. CBraMod and CSBrain use Transformer-based representations, whereas EEGMamba and CodeBrain use SSM-style sequence modeling. CodeBrain combines state-space-based global processing with local attention. This selection covers different mechanisms for integrating spatial and temporal information without requiring a common internal architecture.

CBraMod. CBraMod uses criss-cross attention to model spatial and temporal dependencies through parallel attention paths. An asymmetric conditional positional encoding represents the distinct structures of the channel and time axes. Masked EEG-patch reconstruction provides its pretraining objective, yielding a shared backbone for downstream decoding.

CSBrain. CSBrain organizes EEG information at multiple spatial and temporal scales. Its cross-scale spatiotemporal tokenization captures local patterns within temporal windows and anatomical regions, while structured sparse attention integrates information across windows and regions. Alternating these operations supports the progressive combination of local and broader-context representations.

EEGMamba. EEGMamba uses bidirectional Mamba blocks to encode EEG-patch sequences. Its selective state-space processing offers an alternative to global self-attention for aggregating contextual information. The model learns general EEG representations through masked-patch reconstruction and transfers the pretrained backbone to downstream tasks.

CodeBrain. CodeBrain combines a decoupled temporal–frequency tokenizer with a multi-scale EEG architecture. The tokenizer represents temporal and spectral information through separate discrete components. Its EEG-SSM backbone couples structured global convolution with sliding-window attention, integrating long-range dependencies and local context within a hybrid architecture.

For NeuroAccord, the common interface is the backbone feature sequence and its pooled representation defined in Section 3. TSF associations and consistency objectives operate on these features, with dimensions determined by the selected backbone. The construction introduces neither a shared architectural block across backbones nor a trainable descriptor encoder.

C.3 CONTINUAL-LEARNING BASELINES

The comparison set comprises 18 methods, grouped according to model-parameter and data-sample views of continual learning. Twelve model-side methods retain model parameters or model-derived information, including importance estimates, teacher responses, activation bases, and feature prototypes. Six sample-side methods retain historical examples, sometimes together with model outputs or curvature information. The distinction concerns the information used to constrain later learning, rather than whether a method uses regularization or distillation. The following descriptions specify the published mechanisms underlying the task-incremental baseline adaptations.

C.3.1 MODEL-SIDE APPROACHES

EWC. Elastic weight consolidation estimates parameter importance using the diagonal Fisher information from previous learning. A weighted quadratic penalty discourages changes to important parameters relative to their retained values. Historical knowledge is thus represented by a parameter reference and its coordinate-wise importance.

SI. Synaptic intelligence accumulates parameter importance from each parameter's contribution to loss reduction along the optimization trajectory. Subsequent tasks penalize movement away from the consolidated parameter values in proportion to this importance. Unlike EWC, its importance estimate follows the learning path rather than a post-training Fisher calculation.

LwF. Learning without forgetting retains the previous model as a teacher and distills its old-task predictions on current-task inputs. The new model jointly learns the current task and preserves

these teacher responses, without requiring stored old training examples. The retained knowledge is expressed through the teacher's behavior rather than coordinate-wise parameter importance.

CMD-PCL. CMD-PCL combines confidence-guided mask distillation with prototype-guided contrastive learning for time-series continual learning. The former modulates distillation using prediction confidence, while the latter uses retained class prototypes to structure the separation of old and new classes. It preserves prior knowledge using both the previous model and model-derived feature summaries.

DT2W. DT2W transfers knowledge through shape-aligned temporal distillation. Soft dynamic time warping aligns feature sequences before their discrepancy is penalized, accommodating temporal misalignment between representations. The original class-incremental formulation also uses prototype augmentation to support old-class recognition without storing raw exemplars.

GPM. Gradient projection memory extracts principal directions from layer-input activations using singular value decomposition and accumulates them in a memory basis. Later gradients are projected onto the complement of the retained activation span. The stored basis summarizes directions associated with previous computation rather than historical input records themselves.

SGP. Scaled gradient projection assigns importance to retained representation directions and scales gradient components accordingly. This relaxes uniform exclusion of the historical span, permitting different degrees of adaptation along directions with different importance. Historical knowledge is carried by the basis together with its directional weights.

TRGP. Trust region gradient projection identifies related previous tasks at each layer and forms a trust region for knowledge transfer. Learnable scaling reuses knowledge from those tasks, while gradient projection regulates interference with retained subspaces. Its design couples historical representation memory with selective reuse rather than treating every previous task identically.

O-LoRA. O-LoRA represents task adaptation through low-rank parameter updates and encourages orthogonality between the subspaces associated with different tasks. Previously learned adapters carry historical knowledge, while new adapters learn additional directions. The protection mechanism therefore operates on the geometry of parameter updates.

SD-LoRA. SD-LoRA decouples the direction and magnitude of low-rank adaptation. It retains previously learned directions while adjusting their contributions and learning directions for new tasks. This separates the preservation of adaptation directions from the allocation of their strength during subsequent learning.

BiLoRA. BiLoRA parameterizes adaptation within almost-orthogonal task-specific parameter spaces using fixed orthogonal bases and learnable low-rank components. The construction separates updates associated with different tasks while retaining parameter-efficient adaptation. Previously acquired knowledge is represented through task-specific parameters.

CL-LoRA. CL-LoRA combines task-shared and task-specific low-rank adapters. The shared branch uses orthogonal construction and distillation to preserve transferable information, while task-specific components support new-task adaptation. Block-wise learnable weights regulate their contributions, dividing parameter capacity between shared and specialized knowledge.

C.3.2 SAMPLE-SIDE APPROACHES

ER. Experience replay maintains a finite buffer of labeled historical examples. Training interleaves these examples with current-task data, so earlier tasks continue to contribute supervised updates. The replay set directly represents past experience through selected input–label pairs.

DER. Dark experience replay stores historical examples together with model logits recorded during learning. Later training penalizes discrepancies between current outputs and these stored logits.

The replayed inputs provide access to past records, while the logits preserve information about the model's earlier predictions beyond the hard labels.

DER++. DER++ extends DER with a supervised loss on replayed examples and their ground-truth labels. It combines consistency with stored logits and renewed classification supervision on historical data. Both terms depend on the retained sample buffer, placing DER++ on the sample side despite its distillation component.

A-GEM. Averaged gradient episodic memory computes a reference gradient from examples sampled from episodic memory. When the current gradient conflicts with this reference, it applies a projection that removes the conflicting component under an averaged first-order constraint. Although its correction acts in parameter space, the historical constraint is supplied by retained examples.

OCAR. Online curvature-aware replay combines episodic replay with second-order information to shape online updates. A Kronecker-factored Fisher approximation supports curvature-aware preconditioning under a constraint on predictive-distribution change. Historical samples remain part of the learning process, with curvature information supplementing their contribution to update control.

CSReL. Coreset selection via reducible loss selects historical examples according to the portion of their loss that further learning can reduce. It estimates this quantity by comparing losses from models trained on a coreset and on the available full data. The resulting criterion favors informative, learnable examples over records whose high loss also persists in the reference model, directing the use of limited replay capacity.

C.4 Procedures for the experimental analyses

Main results and plasticity–stability analysis. For each backbone, all methods learn the same nine tasks in the prescribed order. After each stage, we evaluate balanced accuracy on every task learned so far. The resulting stage-by-task matrix yields AAA, FOR, and BWT as defined in Eq. (13).

TSF-reference comparison. For Fig. 4(a), we extract Time–Space–Frequency descriptors from the nine datasets and assemble their pairwise dataset distances into a symmetric matrix. For Fig. 4(b), we hold the backbone, task sequence, data splits, and training and evaluation procedures fixed, changing only the reference used by the framework. The seven alternatives draw on weights, LoRA, gradients, Fisher information, latent PCA, samples, or a sample kernel. We run each configuration through the full task sequence on CBraMod and EEGMamba and compute AAA and FOR from its stage-wise evaluations. The distance matrix describes dataset variation, whereas the controlled comparison tests the effect of the reference choice on continual-learning performance.

Ablations. On CBraMod and EEGMamba, we retrain after removing one or two of TSFRD, COA, and HKA to isolate their contributions. In a separate set of runs, we retain only the indicated time, space, and frequency descriptor views when constructing the TSF reference. Both ablations use the original task sequence and report AAA, FOR, and BWT.

Sensitivity and robustness. We vary one loss weight at a time on CBraMod and EEGMamba while keeping the other two at their default values. We separately repeat the continual-learning evaluation under five random seeds and 20 task orders on all four backbones. The sensitivity panels report AAA and FOR for each weight choice; the robustness panels summarize the run-level AAA and FOR values.

Adaptation to unseen datasets. HMC and TUAB are excluded from the original nine-task sequence. After completing that sequence, we continue learning with these datasets as additional tasks and evaluate the resulting models on their respective decoding objectives. For HMC, we report balanced accuracy, Cohen's κ, and weighted F1; for TUAB, we report balanced accuracy, AUROC, and PR-AUC.