GeMoT: Geometry and Moment Transfer for Source-Free EEG Domain Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Source-free domain adaptation has garnered increasing attention for cross-subject electroencephalography (EEG) decoding, as it eliminates the need for source-data sharing and labeled target calibration. However, strong source-domain performance does not guarantee transferability, and inter-subject variability can undermine source-model predictions. We propose a Geometry and Moment Transfer (GeMoT) framework that jointly encodes source class knowledge in classifier geometry and feature moments to improve source-model transferability and guide target-domain adaptation. For source-domain model training, we introduce a neural-collapse-inspired rule that replaces class-specific classifier weights with pooled source class prototypes, providing a structured initialization for transfer. Using these prototypes, we construct a memory bank of cross-channel means and variances to preserve complementary class statistics. Target features compensated by these feature moments are matched to corresponding classifier weights via cosine similarity to generate pseudo-labels. We design a neighbourhood attraction loss for local prediction consistency, a progressively weighted compensation-guided supervision loss to train the target-domain model with these pseudo-labels, and an inter-class dispersal loss to reduce prediction overlap across pseudo-classes. The adaptation process also incorporates a label-free early-stopping criterion based on the normalized entropy of the pseudo-label class distribution. Offline leave-one-subject-out evaluation covers three backbones and three EEG datasets spanning P300 detection, motor imagery, and vigilance monitoring. With one fixed backbone per dataset, mean accuracies reach 69.36%, 72.50%, and 82.45% on P300, BNCI2014002, and SADT, respectively. These results demonstrate that coupling classifier geometry with class-specific statistical memory provides a robust, generalizable paradigm for source-free cross-subject EEG decoding.

1 Introduction

Differences in brain anatomy, cognitive state, electrode placement, and recording conditions make EEG decoders difficult to transfer across subjects. This variability affects EEG applications broadly, limiting the reuse of models trained on other individuals. Unsupervised domain adaptation can mitigate distribution shift, but commonly requires access to labeled source recordings during target adaptation. Data-governance and operational constraints can make that access impractical.

Source-free domain adaptation (SFDA) instead adapts a trained source model using only unlabeled target data. To address the limited transferability of source-trained models, existing approaches exploit source decision boundaries, information maximization, and pseudo-label refinement (Liang et al., 2020). EEG methods use augmentation and consistency regularization to improve cross-subject robustness, or model interfaces to transfer knowledge without sharing recordings (Xia et al., 2022; Zhang & Wu, 2023). These strategies convey source knowledge through model parameters or predictions, but make limited explicit use of class-specific statistics to complement the classifier's discriminative structure. This raises a complementary question: how should a source model encode class knowledge to guide adaptation after source data become unavailable?

We address this question through two representations of a shared source class prototype: classifier geometry and feature moments. The alignment between class means and classifier weights

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.

in neural collapse motivates the geometric component (Papyan et al., 2020). Our neural-collapse-guided parameter replacement (NCPR) uses pooled source class prototypes to periodically replace class-specific classifier weights during source training. This provides a structural prior for transfer without assuming that the complete neural-collapse regime is attained. Motivated by the use of moment statistics in EEG representations (Mane et al., 2021), we also retain cross-channel means and variances from the same prototypes. These feature moments summarize class-specific response levels and variability, complementing the classifier prototypes.

During adaptation, the stored feature moments compensate target features, which are matched to the corresponding classifier prototypes through cosine similarity to generate pseudo-labels. Three losses then refine the target model: neighbourhood attraction promotes local consistency, compensation-guided supervision learns from the pseudo-labels, and inter-class dispersal reduces prediction overlap. A normalized entropy criterion monitors the pseudo-label class distribution for label-free early stopping. Together, these operations form the Geometry and Moment Transfer (GeMoT) framework.

Our contributions are:

- We formulate transferred class knowledge through a shared source prototype linking classifier alignment to first- and second-order feature summaries. This formulation connects a neural-collapse-inspired geometric prior with moment-based guidance for source-free adaptation.
- We develop GeMoT, coupling NCPR with a compact feature-moment memory. Moment-compensated cosine matching supplies pseudo-labels for three complementary target losses, with normalized pseudo-label entropy governing label-free early stopping.
- We evaluate the framework with three backbones across three EEG datasets under offline leave-one-subject-out evaluation. With one fixed backbone per dataset, mean accuracies reach 69.36%, 72.50%, and 82.45% on P300, BNCI2014002, and SADT, respectively.

2 Related Work

Source-free domain adaptation. In visual domain adaptation, SHOT freezes the source classifier and adapts the encoder through information maximization and pseudo-label refinement (Liang et al., 2020). For cross-subject EEG domain adaptation, ASFA combines source augmentation with target uncertainty reduction and consistency regularization (Xia et al., 2022), while LSFT transfers knowledge through source-model interfaces (Zhang & Wu, 2023). Black-box adaptation has also been studied for cross-subject drowsiness detection (Yuan et al., 2024). These approaches primarily exploit source models and their predictions, with less explicit use of retained class-specific feature statistics for target feature compensation. GeMoT explores this complementary direction by coupling classifier geometry with feature moments derived from shared source class representations.

Neural collapse and prototype geometry. Neural collapse is characterized by within-class feature collapse, simplex equiangular tight-frame (ETF) geometry, classifier–class-mean alignment, and nearest-class-centre decisions (Papyan et al., 2020). Building on these properties, fixed ETF classifiers have been used to improve imbalanced classification (Yang et al., 2022), while preserving feature–classifier alignment helps mitigate forgetting in few-shot class-incremental learning (Yang et al., 2023). These methods organize representations around prescribed classifier geometry, but do not directly address how learned source class structure can guide unlabeled adaptation under source-free constraints.

EEG feature statistics. Temporal variance pooling summarizes learned EEG responses in FBCNet (Mane et al., 2021), while Euclidean alignment (EA) reduces input covariance shift without labels (He & Wu, 2020). These approaches use statistics for feature characterization or input alignment, with less emphasis on retaining class-specific statistics to guide source-free adaptation. GeMoT retains cross-channel means and variances of learned source class prototypes for target feature compensation, complementing subject-level EA with class-conditioned information in the learned feature space.

3 METHOD

GeMoT couples classifier geometry and feature moments through shared source class representations. During source training, NCPR replaces classifier weights with pooled class prototypes, while a compact moment memory retains complementary statistics. During target adaptation, moment-compensated features are matched to corresponding classifier directions to generate pseudo-labels, and three complementary losses refine the target model.

3.1 PROBLEM FORMULATION

Let $\mathcal{D}_s = \{(X_i^s, y_i^s)\}_{i=1}^{n_s}$ contain labeled EEG trials from source subjects, and $\mathcal{D}_t = \{X_i^t\}_{i=1}^{n_t}$ contain unlabeled trials from a target subject. The domains share C classes. An encoder f_θ produces a structured feature map $M_\theta(X) \in \mathbb{R}^{d\times r}$ and a classifier embedding $z_\theta(X) \in \mathbb{R}^d$. Here, d and r count learned channels and structural coordinates, respectively. θ denotes encoder parameters. A bias-free classifier with class-specific weights $w_c \in \mathbb{R}^d$ produces $p_{\theta,W}(X) = \mathrm{softmax}(W z_\theta(X))$, where $W \in \mathbb{R}^{C\times d}$ has row $w_c^\top$ for class c. After source training, adaptation uses only the trained model, retained class statistics, and unlabeled target trials.

3.2 SOURCE LEARNING AND ALIGNMENT

Before pooling data from different source subjects for supervised learning of the source-domain model, we apply subject-wise EA to reduce differences in input channel covariance (He & Wu, 2020). For subject u with n_u trials, let $\mathrm{Cov}(X_i^u)$ denote the temporally centered channel covariance of trial i. The mean covariance is

$$R_u = \frac{1}{n_u}\sum_{i=1}^{n_u}\mathrm{Cov}(X_i^u), \tag{1}$$

and each trial is aligned as

$$\widetilde{X}_i^u = R_u^{-1/2} X_i^u. \tag{2}$$

When R_u is positive definite, this transform makes the subject's mean channel covariance the identity without requiring identical class-conditional covariances. Source subjects are aligned independently, the same procedure is applied to the target subject using its complete unlabeled set.

The aligned source trials are then pooled to learn class-discriminative representations using cross-entropy with label smoothing. For class c, the smoothed target is $\widetilde{y}_{ic}^s = (1-\epsilon)\mathbb{I}[y_i^s = c] + \epsilon/C$, where $\epsilon \in [0, 1)$ controls smoothing and $\mathbb{I}[\cdot]$ is the indicator function. Let $\mathcal{I}_s$ index the source-training partition and p_{ic}^s denote the predicted class probability. The source objective is

$$\mathcal{L}_s(\theta, W) = -\frac{1}{|\mathcal{I}_s|}\sum_{i\in\mathcal{I}_s}\sum_{c=1}^{C}\widetilde{y}_{ic}^s \log p_{ic}^s, \tag{3}$$

A separate source-validation partition monitors training.

3.3 NEURAL-COLLAPSE-GUIDED PARAMETER REPLACEMENT

Although the above supervised training regularizes source predictions, it does not explicitly enforce alignment between classifier weights and source class prototypes. We therefore introduce NCPR to encode the learned class structure directly in the classifier weights, providing a geometric reference for subsequent target adaptation.

Let $\mathcal{I}_c = \{i \in \mathcal{I}_s : y_i^s = c\}$ index the source-training trials in class c. For the current encoder, define the class prototype as the empirical conditional mean

$$A_c = \frac{1}{|\mathcal{I}_c|}\sum_{i\in\mathcal{I}_c} M_\theta(\widetilde{X}_i^s). \tag{4}$$

Here, $\widetilde{X}_i^s$ denotes the EA-aligned source trial indexed by i. Adaptive global average pooling (AGAP) averages the r structural coordinates while retaining the d learned channels. It gives the pooled source prototype

$$\pi_c = \mathrm{AGAP}(A_c). \tag{5}$$

Because average pooling is linear, π_c also equals the class mean of individually pooled structured features. It therefore has a least-squares characterization:

$$\pi_c = \arg\min_{v \in \mathbb{R}^d} \sum_{i \in \mathcal{I}_c} \left\| v - \mathrm{AGAP}(M_\theta(\widetilde{X}_i^s)) \right\|_2^2. \tag{6}$$

This elementary identity follows by centering each pooled feature at its class mean. The objective separates into within-class scatter and $|\mathcal{I}_c|\|v - \pi_c\|_2^2$, making the minimizer unique. Thus, prototype replacement transfers a source class centroid in the pooled structured-feature space into the classifier parameters. It does not establish that this centroid minimizes target error.

At each replacement event, the classifier satisfies

$$w_c^+ = \pi_c. \tag{7}$$

The superscript $+$ denotes the state immediately after replacement. Events occur when source-validation accuracy matches or exceeds its previous maximum, after which source optimization resumes. The event schedule and memory updates are detailed in Supplementary Section B.4.

This intervention aligns classifier parameters with pooled structured prototypes at replacement, but subsequent gradients can change them. Moreover, $z_\theta(X)$ need not equal $\mathrm{AGAP}(M_\theta(X))$. NCPR therefore imposes a prototype-based structural prior rather than proving neural-collapse self-duality or a nearest-class-centre decision rule.

As illustrated in Figure 1, the representations of one held-out subject change across source training and target adaptation. Before training (a), trials form a diffuse cloud with substantial overlap between classes. Source training without NCPR (b) produces partial class organization, but the distribution remains dispersed. With NCPR (c), the features occupy a more compact region and show more localized class groupings. Target adaptation (d) further organizes the representation into two elongated groups.

These patterns illustrate the motivation for using two aspects of neural collapse in transfer: within-class compactness and classifier alignment with class means. Compact class structure can make each transferred prototype more representative of its class. Aligning classifier weights with pooled class means supplies the references for cosine-based pseudo-label assignment in Equation 12. Those pseudo-labels supervise target predictions through $\mathcal{L}_{\mathrm{cgs}}$, connecting the source geometric prior to target learning. Together with label smoothing, this geometric constraint aims to produce class structure that remains useful after source data are discarded.

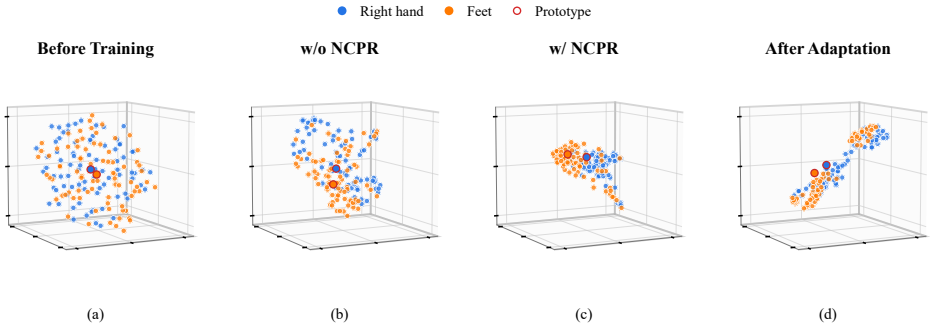


Figure 1: UMAP projections of BNCI2014002 subject 8: (a) before source training, (b) source training without NCPR, (c) with NCPR, and (d) after target adaptation. Blue/orange points denote right-hand/feet trials. Red-outlined markers denote source class prototypes. Panels use identical held-out trials and a shared projection protocol. Labels are used only for retrospective visualization.

3.4 Class-Specific Feature Moments

For class c, $A_{c,kq}$ denotes the retained prototype at learned channel k and structural coordinate q. Its cross-channel mean $\mu_{c,q}$ and sample variance $\nu_{c,q}$ are

$$\mu_{c,q} = \frac{1}{d} \sum_{k=1}^{d} A_{c,kq}. \tag{8}$$

$$\nu_{c,q} = \frac{1}{d-1}\sum_{k=1}^{d}(A_{c,kq} - \mu_{c,q})^2. \tag{9}$$

Collecting the r coordinates gives $\mu_c = (\mu_{c,1}, \ldots, \mu_{c,r})^\top$ and $\nu_c = (\nu_{c,1}, \ldots, \nu_{c,r})^\top$ in $\mathbb{R}^r$. The variance requires $d > 1$ and uses the sample-variance denominator $d - 1$, without assuming independent channels. The fixed memory $\mathcal{P}_s = \{(\mu_c, \nu_c)\}_{c=1}^{C}$ stores $2Cr$ scalars, independently of the source trial count. These summarize class-averaged feature maps, rather than across-trial moments or spectral power.

3.5 Moment-Compensated Pseudo-Labeling

For target trial i, let $m_i = \mathrm{AGAP}(M_\theta(\widetilde{X}_i^t)) \in \mathbb{R}^d$. We combine the pooled source moments into a class-specific compensation coefficient:

$$a_c = \frac{1}{r}\sum_{q=1}^{r}\big[\lambda\mu_{c,q} + (1-\lambda)\nu_{c,q}\big]. \tag{10}$$

The parameter $\lambda \in [0, 1]$ balances the two moments. The compensated representation used for matching is

$$h_{ic} = \mathcal{N}_b(a_c m_i), \tag{11}$$

where $\mathcal{N}_b$ is the normalization operator used during assignment and b identifies the backbone (Supplementary Section B.5). The corresponding classifier prototype w_c provides the class reference:

$$s_{ic} = \frac{h_{ic}^\top w_c}{\|h_{ic}\|_2 \|w_c\|_2}. \tag{12}$$

For nonzero representations and classifier weights, the pseudo-label is

$$\widehat{y}_i = \arg\max_{c\in\{1,\ldots,C\}} s_{ic}. \tag{13}$$

The construction preserves class correspondence between feature moments and classifier prototypes. It performs scalar conditioning rather than coordinate-wise moment matching. Normalization is retained explicitly because a positive scalar alone cancels in cosine similarity. Its role and scale-invariance boundary are discussed in the supplement. Pseudo-labels are refreshed periodically using the complete unlabeled target set.

3.6 Target Adaptation Objectives

Neighbourhood attraction loss. Write $z_i^t = z_\theta(\widetilde{X}_i^t)$ and $p_i = p_{\theta,W}(\widetilde{X}_i^t)$ for the target embedding and prediction. Let $\mathcal{N}_K(i)$ index the K most similar target-memory embeddings under cosine similarity, excluding trial i. With stored neighbour predictions $\bar{p}_k$ held fixed during each update, local consistency is encouraged by

$$\mathcal{L}_{\mathrm{sim}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k\in\mathcal{N}_K(i)} p_i^\top \bar{p}_k, \tag{14}$$

where B is the mini-batch size. Neighbours are selected by feature similarity without pseudo-class filtering. This form omits a term constant with respect to the current update, preserving the parameter gradients. The supplement gives the derivation.

Compensation-guided supervision loss. Moment-compensated pseudo-labels connect retained source knowledge to target learning through cross-entropy:

$$\mathcal{L}_{\mathrm{cgs}} = -\frac{1}{B}\sum_{i=1}^{B}\log p_{i,\widehat{y}_i}. \tag{15}$$

To gradually strengthen this supervision, let $j \in \{0, \ldots, J-1\}$ index target updates within budget J. Its weight increases according to

$$\alpha(j) = 1 - \exp(-\gamma j/J), \tag{16}$$

where $\gamma > 0$ controls the growth rate. Early updates thus emphasize local target structure before compensation-derived supervision becomes stronger.

Inter-class dispersal loss. For pairs assigned to different pseudo-classes, we penalize the overlap between their prediction vectors:

$$\mathcal{L}_{\text{dis}} = \frac{1}{B}\sum_{i=1}^{B}\sum_{\ell=1}^{B}\mathbb{I}[\widehat{y}_i \neq \widehat{y}_\ell]p_i^\top p_\ell. \tag{17}$$

The indicator selects different-class pairs, complementing neighbourhood attraction with separation between pseudo-classes. The sum includes ordered pairs and is scaled by the mini-batch size.

Joint objective. The target model minimizes

$$\mathcal{L}_t = \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} + \alpha(j)\mathcal{L}_{\text{cgs}} + \lambda_{\text{dis}}\mathcal{L}_{\text{dis}}. \tag{18}$$

Both the encoder and classifier are updated, while the source moment memory remains fixed. The weights λ_{sim} and λ_{dis} control local agreement and inter-class separation. Training settings and the complete procedure appear in Supplementary Section B and Algorithm S1.

3.7 Label-Free Early Stopping

At each evaluation update j, the pseudo-label class proportions are

$$\rho_c(j) = \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{I}[\widehat{y}_i(j) = c]. \tag{19}$$

Their normalized entropy, denoted h-norm, is

$$\mathcal{H}(j) = -\frac{1}{\log C}\sum_{c=1}^{C}\rho_c(j)\log\rho_c(j), \tag{20}$$

with $0\log 0 = 0$ and $C > 1$. Thus $\mathcal{H}(j) \in [0, 1]$, and low values indicate concentration on fewer classes. The first evaluation at or below a dataset–backbone threshold initiates early stopping after a fixed grace period, subject to the update budget. The model with maximum observed h-norm is restored for inference. The model-selection rule is detailed in Supplementary Section B.5.

This criterion uses no target labels, but measures assignment balance rather than correctness. Its interpretation assumes that broad class coverage is desirable. Genuinely imbalanced or missing-class target populations require a different prior.

4 Experiments

4.1 Datasets and Preprocessing

Table 1 summarizes the three public EEG datasets, P300 event-related responses (Ramele et al., 2019), BNCI2014002 motor imagery (BNCI Horizon 2020, 2026; Steyrl et al., 2016), and SADT vigilance monitoring (Cao et al., 2019; Cui et al., 2022), used to evaluate the framework. They span different class semantics, recording layouts, and temporal scales. Each experiment transfers between subjects within a dataset. It does not transfer between different tasks or label spaces.

Table 1: EEG datasets after preprocessing. ERP: event-related potential. MI: motor imagery. Trial counts are per subject.

Dataset	Paradigm	Subjects	EEG ch.	Hz	Classes	Window	Trials/subject
P300	ERP	8	8	250	2	0–1 s	1396–1400
BNCI2014002	MI	14	15	512	2	3–8 s	160
SADT	Vigilance	11	30	128	2	3 s	102–314

P300 trials undergo 1–12 Hz filtering and 0–1 s epoching, with deterministic within-character sampling to balance target and non-target events. BNCI2014002 combines both sessions, uses 3–8 s epochs, and applies 8–40 Hz filtering, common-average referencing, and within-trial channel-wise normalization. SADT uses the released three-second epochs without additional signal transforms. All subjects then undergo EA independently. Supplementary Section B.1 provides the complete preprocessing specifications.

4.2 Evaluation Protocol

We use offline, transductive leave-one-subject-out (LOSO) evaluation. Each fold uses one subject as the unlabeled target and pools the remaining subjects as labeled sources. Following subject-wise EA, source trials are split into 70% training and 30% validation. Source validation controls prototype replacement and contains no target trials. Adaptation uses the complete unlabeled target set after source trials are discarded. Target labels are used to evaluate accuracy, they do not enter adaptation gradients, checkpoint selection, or stopping decisions.

Within GeMoT, we evaluate CNN, EEGNet (Lawhern et al., 2018), and SincNet-inspired (Ravanelli & Bengio, 2018) encoders. Source learning and target adaptation share the optimizer and learning-rate settings. Architectures, training procedures, and hyperparameters are specified in Supplementary Sections B.3–B.5.

Repeated-run accuracies are averaged within each held-out subject. Dataset summaries give the mean and sample standard deviation across these subject-level averages. In the main comparison, GeMoT uses EEGNet for P300, CNN for BNCI2014002, and SincNet for SADT, without switching backbones between subjects. Baselines retain their method-specific architectures, training procedures, and source-access regimes. The comparison therefore measures complete method configurations.

We apply two-sided paired Student's t-tests to matched subject accuracies, giving 8, 14, and 11 pairs for P300, BNCI2014002, and SADT. Holm correction covers the nine baseline comparisons separately within each dataset, with adjusted $p < 0.05$ defining significance. Supplementary Section D.2 presents paired differences, confidence intervals, raw and adjusted p-values, and inferential limitations.

4.3 Comparison with Existing Methods

As shown in Table 2, GeMoT achieves the highest mean accuracy among all compared methods on all three datasets. It reaches 69.36% on P300, 72.50% on BNCI2014002, and 82.45% on SADT. These results exceed the strongest baseline means by 1.95, 3.39, and 4.23 percentage points, respectively. Subject-level results are provided in Supplementary Tables S2–S4.

Table 2: LOSO accuracy (%) across three EEG datasets. The Classical group comprises CSP (CSP+LDA), EA (EA followed by CSP+LDA), and standalone EEGNet. Entries show cross-subject means and sample standard deviations. Boldface marks column maxima. For GeMoT, the backbones are EEGNet, CNN, and SincNet on P300, BNCI2014002, and SADT, respectively.

Category	Method	P300	BNCI2014002	SADT
Classical	CSP	$51.69_{\pm 2.14}$	$64.69_{\pm 15.38}$	$66.88_{\pm 17.17}$
	EA	$53.20_{\pm 1.77}$	$64.73_{\pm 14.34}$	$74.54_{\pm 7.61}$
	EEGNet	$61.76_{\pm 4.60}$	$54.12_{\pm 9.55}$	$69.91_{\pm 7.42}$
UDA	DAN	$67.15_{\pm 4.45}$	$68.97_{\pm 11.14}$	$78.22_{\pm 7.96}$
	DANN	$61.95_{\pm 5.84}$	$68.62_{\pm 10.93}$	$62.81_{\pm 4.45}$
	CDAN	$67.41_{\pm 4.34}$	$66.72_{\pm 13.45}$	$78.21_{\pm 8.46}$
SFDA	SHOT	$63.00_{\pm 5.90}$	$69.11_{\pm 17.25}$	$76.36_{\pm 8.98}$
	ASFA	$52.94_{\pm 2.28}$	$64.24_{\pm 15.15}$	$73.66_{\pm 12.68}$
	LSFT	$53.90_{\pm 2.20}$	$68.26_{\pm 15.26}$	$78.04_{\pm 12.01}$
	GeMoT	$\mathbf{69.36}_{\pm 5.04}$	$\mathbf{72.50}_{\pm 12.32}$	$\mathbf{82.45}_{\pm 9.92}$

Subject-paired tests further assess these differences. After within-dataset Holm correction over the nine baseline comparisons, GeMoT significantly outperforms all nine baselines on P300, six on BNCI2014002, and three on SADT. The remaining comparisons have positive mean differences but do not meet the adjusted significance threshold. Paired effect estimates, confidence intervals, and adjusted p-values are provided in Supplementary Tables S5–S7.

With the complete framework, EEGNet leads on P300, CNN on BNCI2014002, and SincNet on SADT, as shown in Table 3. The same prototype construction, moment memory, and target objec-

tives remain applicable across these architectures. Performance nevertheless depends on the encoder, supporting portability of the adaptation interface rather than backbone-invariant accuracy.

Table 3: LOSO accuracy (%) of GeMoT with three backbones. Entries are subject means and sample standard deviations. Boldface marks the highest mean per dataset.

Backbone	P300	BNCI2014002	SADT
CNN	$64.31_{\pm 4.89}$	$\mathbf{72.50}_{\pm 12.32}$	$82.02_{\pm 7.37}$
EEGNet	$\mathbf{69.36}_{\pm 5.04}$	$66.43_{\pm 13.01}$	$79.76_{\pm 7.94}$
SincNet	$52.58_{\pm 1.88}$	$68.66_{\pm 15.95}$	$\mathbf{82.45}_{\pm 9.92}$

4.4 Component and Loss Ablations

Transferred class knowledge. Table 4 summarizes the contributions of prototype replacement and feature moments to classification accuracy. NCPR improves mean accuracy over source-only training by 2.31, 4.76, and 14.89 points on P300, BNCI2014002, and SADT. The complete framework has the highest mean in every column, although individual moment effects remain small and dataset dependent.

Table 4: Ablation of transferred class knowledge. Entries are subject mean accuracies (%) and sample standard deviations. The w/o mean and w/o variance variants retain only the variance and mean moments with NCPR, respectively. The final row is the complete method.

Configuration	P300	BNCI2014002	SADT
Source only	$65.42_{\pm 6.01}$	$66.12_{\pm 14.40}$	$66.94_{\pm 7.61}$
w/ NCPR	$67.73_{\pm 4.76}$	$70.88_{\pm 14.37}$	$81.83_{\pm 9.06}$
w/o mean moment	$67.67_{\pm 5.02}$	$70.62_{\pm 14.03}$	$81.68_{\pm 8.96}$
w/o variance moment	$67.93_{\pm 4.67}$	$71.00_{\pm 13.62}$	$82.05_{\pm 8.97}$
GeMoT	$\mathbf{69.36}_{\pm 5.04}$	$\mathbf{72.50}_{\pm 12.32}$	$\mathbf{82.45}_{\pm 9.92}$

Target objectives. Table 5 shows target-domain accuracy when the three losses are used individually and in combination. The strongest reduced objective changes across datasets: attraction with supervision on P300, dispersal alone on BNCI2014002, and attraction with dispersal on SADT. Combining all three losses yields the highest mean throughout, exceeding the strongest reduced configuration by 0.77, 0.47, and 0.16 points, respectively. Supplementary Tables S11–S13 show subject-level variation.

Table 5: Target-loss ablation: neighbourhood attraction ($\mathcal{L}_{\text{sim}}$), compensation-guided supervision ($\mathcal{L}_{\text{cgs}}$), and inter-class dispersal ($\mathcal{L}_{\text{dis}}$). Entries are subject mean accuracies (%) and sample standard deviations. Boldface marks column maxima.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cgs}}$	$\mathcal{L}_{\text{dis}}$	P300	BNCI2014002	SADT
✓	×	×	$67.70_{\pm 6.02}$	$66.58_{\pm 16.75}$	$67.48_{\pm 14.90}$
×	✓	×	$65.17_{\pm 6.28}$	$70.29_{\pm 12.02}$	$72.68_{\pm 9.50}$
×	×	✓	$67.33_{\pm 5.39}$	$72.03_{\pm 13.50}$	$80.90_{\pm 9.22}$
✓	✓	×	$68.59_{\pm 5.73}$	$66.22_{\pm 15.56}$	$66.07_{\pm 12.62}$
✓	×	✓	$67.92_{\pm 5.93}$	$71.45_{\pm 13.86}$	$82.29_{\pm 8.97}$
×	✓	✓	$66.79_{\pm 5.87}$	$71.36_{\pm 12.98}$	$77.23_{\pm 9.20}$
✓	✓	✓	$\mathbf{69.36}_{\pm 5.04}$	$\mathbf{72.50}_{\pm 12.32}$	$\mathbf{82.45}_{\pm 9.92}$

4.5 Hyperparameter Sensitivity

P300 accuracy remains within 68.68–70.07%, whereas BNCI2014002 and SADT show larger declines at extreme parameter values, as shown in Figure 2. We vary K from 1 to 10, λ_{sim} from 0.4 to 4.0, and λ_{dis} from 0.1 to 1.0. BNCI2014002 peaks at the configured values and retains similar accuracy at neighbouring moderate settings. Larger neighbourhoods or attraction weights can dilute

local class structure, consistent with the declines at these extremes. SADT exhibits broad plateaus, particularly for $\lambda_{\text{dis}} \geq 0.3$.

The largest declines on BNCI2014002 and SADT occur with very weak dispersal, suggesting that class separation is especially important when local agreement alone is insufficient. These results support a useful region around the configured values rather than dependence on an isolated setting.

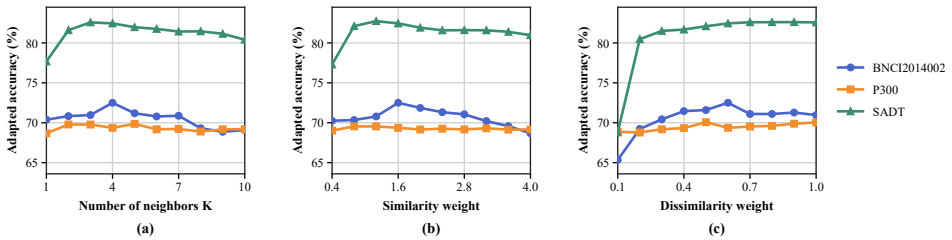


Figure 2: One-factor sensitivity of (a) neighbour count K, (b) attraction weight λ_{sim}, and (c) dispersal weight λ_{dis}. Other parameters remain fixed. Curves show subject-mean LOSO accuracy of GeMoT after h-norm selection, using EEGNet/CNN/SincNet backbones for P300/BNCI2014002/SADT. Dataset-specific views appear in Supplementary Figures S1–S3.

5 Discussion and Conclusion

GeMoT achieves the highest mean accuracy among the compared methods on P300, BNCI2014002, and SADT, without accessing source trials or target labels during adaptation. These results span event-related responses, motor imagery, and vigilance monitoring, supporting the value of jointly transferring classifier geometry and class-specific statistics for cross-subject EEG decoding. The framework connects source knowledge retention with target adaptation through a shared representation of class structure.

NCPR embeds source class prototypes in the classifier, providing explicit class references for transfer. Moment memory retains response levels and variability from the same prototypes and supplies class-specific information for target feature compensation. Their common origin preserves class correspondence between geometric references and statistical memory during pseudo-label generation. The complete framework leads the module ablations on all three datasets, supporting the combined design of prototype geometry, moment memory, and target learning. Together, these components provide a way to preserve and reuse source class knowledge after the original recordings become unavailable.

The target objectives combine source class guidance with the structure of unlabeled target data. Neighbourhood attraction promotes local consistency, compensation-guided supervision anchors learning to the transferred class references, and inter-class dispersal encourages separation between pseudo-classes. Their combination achieves the highest mean accuracy in the loss ablations on each dataset, supporting complementary roles in target adaptation. The same prototype construction, moment memory, and objectives also accommodate CNN, EEGNet, and SincNet, allowing encoder choice to reflect the EEG task. Overall, GeMoT provides an effective, reusable framework that links structured source knowledge to unsupervised target learning for cross-subject EEG decoding.

6 Limitations

The framework depends on the source model learning reliable class structure. Weak representations can produce unreliable classifier prototypes and feature moments, while large target shifts can weaken their correspondence with target features. Without source trials or target labels during adaptation, diagnosing and correcting these failures remains difficult. Stronger source representations, uncertainty-aware memories, and detection of out-of-support targets are therefore relevant extensions.

Ethics and Reproducibility Statement

This study uses public EEG datasets without recruiting new subjects. The original publications document consent and ethics procedures. Withholding source trials during adaptation does not provide a formal privacy guarantee, including resistance to reconstruction or membership inference. The evaluation protocol and Supplementary Sections B and C describe preprocessing, architectures, training, model selection, and aggregation across repeated runs.

AI Assistance Statement

Generative AI tools assisted with language polishing, Chinese–English translation, and checks of numerical, terminological, mathematical, and presentational consistency. The authors proposed the core scientific ideas, designed the method and experiments, interpreted the results, and determined the conclusions. The authors retain responsibility for reviewing and verifying all AI-assisted text, code, and calculations.

References

Benjamin Blankertz, Ryota Tomioka, Steven Lemm, Motoaki Kawanabe, and Klaus-Robert Müller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):41–56, 2008. doi: 10.1109/MSP.2008.4408441.

BNCI Horizon 2020. The BNCI Horizon 2020 database: Data sets. https://bnci-horizon-2020.eu/database/data-sets, 2026. Accessed September 4, 2026.

Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific Data*, 6:19, 2019. doi: 10.1038/s41597-019-0027-4.

Jian Cui. EEG driver drowsiness dataset. figshare, version 3, 2021. URL https://figshare.com/articles/dataset/EEG_driver_drowsiness_dataset/14273687.

Jian Cui, Zirui Lan, Yisi Liu, Ruilin Li, Fan Li, Olga Sourina, and Wolfgang Müller-Wittig. A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods*, 202:173–184, 2022. doi: 10.1016/j.ymeth.2021.04.017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. URL https://www.jmlr.org/papers/v17/15-239.html.

Peiliang Gong, Han Zhang, Zhen Jiang, Chenyu Liu, Ziyu Jia, Xinliang Zhou, Daoqiang Zhang, and Xiaoli Li. Foundation model guided dual-branch co-adaptation for source-free EEG decoding, 2026. URL https://arxiv.org/abs/2605.00857.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. doi: 10.1109/TBME.2019.2913914.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105. PMLR, 2015. URL https://proceedings.mlr.press/v37/long15.html.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/ab88b15733f543179858600245108dd8-Abstract.html.

Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A. P. Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain–computer interface, 2021. URL https://arxiv.org/abs/2104.01233.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Rodrigo Ramele, Ana Julia Villar, and Juan Miguel Santos. Histogram of gradient orientations of signal plots applied to P300 detection. *Frontiers in Computational Neuroscience*, 13:43, 2019. doi: 10.3389/fncom.2019.00043.

Mirco Ravanelli and Yoshua Bengio. Speaker recognition from raw waveform with SincNet, 2018. URL https://arxiv.org/abs/1808.00158.

David Steyrl, Reinhold Scherer, Josef Faller, and Gernot R. Müller-Putz. Random forests in non-invasive sensorimotor rhythm brain–computer interfaces: A practical and convenient non-linear classifier. *Biomedical Engineering / Biomedizinische Technik*, 61(1):77–86, 2016. doi: 10.1515/bmt-2014-0117.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=uXl3bZLkr3c.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4540d267eeec4e5dbd9dae9448f0b739-Abstract-Conference.html.

Kun Xia, Lingfei Deng, Wlodzislaw Duch, and Dongrui Wu. Privacy-preserving domain adaptation for motor imagery-based brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(11):3365–3376, 2022. doi: 10.1109/TBME.2022.3168570.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2203.09081.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Liqiang Yuan, Ruilin Li, Jian Cui, and Mohammed Yakoob Siyal. SPARK: A high-efficiency black-box domain adaptation framework for source privacy-preserving drowsiness detection. *IEEE Journal of Biomedical and Health Informatics*, 28(6):3478–3488, 2024. doi: 10.1109/JBHI.2024.3377373.

Wen Zhang and Dongrui Wu. Lightweight source-free transfer for privacy-preserving motor imagery classification. *IEEE Transactions on Cognitive and Developmental Systems*, 15(2):938–949, 2023. doi: 10.1109/TCDS.2022.3193731.

A Extended Related Work

Cross-subject alignment with source access. Cross-subject EEG adaptation commonly uses unsupervised domain adaptation (UDA), with labeled source and unlabeled target trials jointly available during training. Input-level methods reduce subject-specific covariance shift. EA whitens each trial using the subject's unlabeled mean channel covariance and can precede handcrafted or learned feature extraction (He & Wu, 2020). Feature-level UDA instead reduces the encoder's sensitivity to domain identity. DAN matches source and target distributions through multi-kernel maximum mean discrepancy (Long et al., 2015). DANN introduces a gradient-reversal domain discriminator (Ganin et al., 2016), whereas CDAN conditions discrimination on joint feature–prediction representations to account for multimodal class structure (Long et al., 2018). Together, these methods motivate covariance normalization and structure-aware alignment. The adaptation objectives of DAN, DANN, and CDAN require source samples and therefore fall outside our source-free deployment setting.

Source-free EEG adaptation. SFDA transfers source knowledge without accessing source recordings during target adaptation. ASFA combines source-side EEG augmentation with target-side uncertainty reduction and consistency regularization (Xia et al., 2022). It initializes the target model from source parameters, fixes the classifier, and adapts the feature extractor. A separate black-box extension uses source-model queries for initialization. LSFT instead obtains predictions from locally trained source classifiers through model interfaces (Zhang & Wu, 2023). Target samples with high agreement across source predictions form a pseudo-labeled virtual intermediate domain, which supports subsequent feature adaptation without accessing source model parameters. These approaches differ in how source knowledge is exposed and used. GeMoT retains classifier prototypes together with class-specific feature moments to guide target pseudo-labeling and optimization. It shares the source-data-free setting while retaining compact source statistics alongside model parameters.

Test-time adaptation. Test-time adaptation (TTA) updates a pretrained model using unlabeled inputs observed after deployment. TENT minimizes prediction entropy while updating normalization statistics and channel-wise affine parameters batch by batch (Wang et al., 2021). This provides self-supervision from an unlabeled test stream without replaying source examples. Our protocol instead performs offline, transductive adaptation on the complete held-out subject. Building on target-only optimization, GeMoT retains fixed source feature moments and uses neural-collapse-guided parameter replacement to introduce source class prototypes into the classifier. Neighbourhood attraction, progressively weighted compensation-guided supervision, and inter-class dispersal then constrain target predictions.

EEG foundation models. Large-scale self-supervised pretraining provides another route to transferable EEG representations. LaBraM combines channel-patch tokenization with masked prediction across heterogeneous EEG corpora (Jiang et al., 2024), whereas EEGPT aligns spatial and temporal representations in a pretrained Transformer (Wang et al., 2024). Foundation-model-guided source-free adaptation further uses a pretrained EEG model as deployment-time guidance (Gong et al., 2026). Such representations can transfer across datasets and tasks, although subject-specific deployment shift remains. GeMoT offers a complementary interface that encodes task-specific source knowledge in classifier prototypes and dual feature moments. We evaluate this interface with CNN, EEGNet, and SincNet. Integration with EEG foundation models remains future work.

B Experimental Details

This appendix documents preprocessing, baseline access regimes, backbone architectures, and source and target training configurations.

B.1 Datasets and preprocessing

- **P300.** We use the public healthy-participant P300 release described by Ramele et al. (2019), with eight subjects and eight EEG channels (Fz, Cz, P3, Pz, P4, PO7, PO8, and Oz) sampled at 250 Hz. Each subject completes 35 character-level trials containing 120 flash

events each. We apply a fifth-order 1–12 Hz Butterworth filter in second-order-section form and extract one-second epochs from usable flash onsets. Within each character, all usable target epochs are retained, together with an equally sized non-target subset selected deterministically at evenly spaced indices. The resulting balanced binary data contain 1,396–1,400 epochs per subject, each of shape 8×250.

- **BNCI2014002.** This motor-imagery dataset includes 14 subjects performing right-hand versus feet imagery, with 15 EEG channels sampled at 512 Hz (BNCI Horizon 2020, 2026; Steyrl et al., 2016). We merge each subject's training and evaluation sessions and extract the five-second interval from 3 to 8 s after each event marker. Preprocessing applies a fifth-order 8–40 Hz Butterworth band-pass filter, common-average referencing, and channel-wise z-normalization within each trial. The channel-first data contain 160 trials per subject, each of shape 15×2560, with exactly 80 trials per class.
- **SADT.** We use the balanced 11-subject release (Cui, 2021) derived from the sustained-attention driving data of Cao et al. (2019), following the extraction described by Cui et al. (2022). It contains 30-channel, 384-sample epochs at 128 Hz. Our preprocessing groups the three-second epochs by subject, converts samples to 32-bit floating point, and maps the original alert/drowsy labels from $\{0, 1\}$ to $\{1, 2\}$. We apply no additional temporal filtering, referencing, or z-normalization before EA. The processed subjects have balanced class counts and 102–314 trials each, of shape 30×384.

Subject-wise EA. All datasets undergo the same subject-wise alignment. The covariance operator treats channels as variables and temporal samples as observations, after centering each channel over time. For subject u, we average trial covariance matrices to obtain R_u and left-multiply each trial by the symmetric inverse square root $R_u^{-1/2}$. Source subjects are aligned separately before pooling. The complete unlabeled target set determines target alignment. Neither labels nor cross-subject covariance statistics enter EA.

B.2 Compared methods

Following Table 2, we group CSP, EA, and the standalone EEGNet baseline together, followed by source-available UDA and source-free methods. Their objectives and target-stage information access are summarized below.

- **CSP (CSP+LDA).** This supervised pipeline combines common spatial patterns with linear discriminant analysis (Blankertz et al., 2008). CSP learns spatial filters that contrast class-specific projected variances, and LDA classifies the resulting log-variance features. It uses labeled source trials without covariance alignment or target-time parameter updates.
- **EA (CSP+EA).** This baseline applies subject-wise EA before the same CSP+LDA pipeline (He & Wu, 2020). Each subject is whitened by the inverse square root of its mean trial covariance. CSP+EA is therefore shorthand for EA preprocessing followed by CSP+LDA. The baseline isolates input-space alignment without deep representation adaptation.
- **EEGNet.** This standalone classifier combines temporal convolution, depthwise spatial filtering, and separable temporal feature extraction (Lawhern et al., 2018). It is grouped with CSP and EA in Table 2, and its source-trained classifier is evaluated without target adaptation. The EEGNet backbone in the backbone comparisons instead operates within the complete GeMoT framework.
- **DAN.** Deep Adaptation Network aligns source and target feature distributions through multi-kernel maximum mean discrepancy (Long et al., 2015). Source classification preserves discriminability, while the discrepancy term matches kernel mean embeddings at task-specific layers. Adaptation requires simultaneous labeled source and unlabeled target mini-batches.
- **DANN.** Domain-Adversarial Neural Network connects a binary domain classifier to the encoder through gradient reversal (Ganin et al., 2016). The encoder minimizes source label loss while maximizing domain classification error, encouraging class-discriminative

features that obscure domain identity. Each adversarial update requires labeled source and unlabeled target trials.

- **CDAN.** Conditional Domain Adversarial Network feeds multilinear feature–prediction representations to a domain discriminator (Long et al., 2018). Prediction conditioning accounts for multimodal class structure, while entropy conditioning downweights uncertain examples. Like DAN and DANN, CDAN requires source trials during target adaptation.
- **SHOT.** Source Hypothesis Transfer freezes a trained source classifier and adapts its encoder (Liang et al., 2020). Conditional-entropy minimization promotes confidence, a diversity term discourages single-class collapse, and centroid-based pseudo-label refinement supplies self-supervision. Adaptation accesses model parameters and unlabeled target trials, without source samples.
- **ASFA.** Augmentation-based Source-Free Adaptation separates source training from target adaptation (Xia et al., 2022). Source-side EEG augmentation promotes cross-subject generalization. Target-side uncertainty reduction emphasizes reliable predictions, and consistency regularization penalizes disagreement under perturbations. Adaptation uses the source model and unlabeled target trials without raw source EEG. The table entry denotes ASFA, rather than the original work's separate black-box extension.
- **LSFT.** Lightweight Source-Free Transfer encapsulates locally trained source models as application programming interfaces, avoiding transmission of source recordings or complete models (Zhang & Wu, 2023). Interface predictions construct a virtual intermediate domain, whose knowledge is transferred to the unlabeled target subject through feature adaptation. This limits exposed source information and the parameter and communication costs of conventional source-data transfer.

Comparison boundary. The baselines retain method-specific architectures, training procedures, and source-access regimes. The main comparison evaluates these configurations. The backbone study and ablations examine choices within GeMoT. The GeMoT row uses EEGNet for all P300 subjects, CNN for all BNCI2014002 subjects, and SincNet for all SADT subjects. The paired tests in Supplementary Section D.2 quantify differences under these settings, without isolating architecture, tuning, or source-access effects.

B.3 Backbone architectures

Common interface. Aligned trials enter each network as $X \in \mathbb{R}^{B\times 1\times E\times T}$, with batch size B, electrode count E, and temporal length T. The singleton axis is the input feature channel. Each encoder returns a structured map $M_\theta(X)$ for prototype construction and an embedding $z_\theta(X)$ for classification. The corresponding feature-map shapes and embedding dimensions are listed in Table S1, with the batch axis omitted. L denotes pooled temporal length and d the learned feature-channel count. The flattened structural size is $r = T$ for SincNet and $r = 2L$ for CNN and EEGNet.

Adaptive global average pooling (AGAP) averages structural coordinates while retaining learned channels. It maps $d \times T$ to d for SincNet and $d \times 2 \times L$ to d for CNN and EEGNet. Applied to a stored feature-moment map, pooling returns one class-specific scalar. Pooling is one-dimensional for SincNet and two-dimensional for CNN and EEGNet.

Table S1: Backbone feature maps and embedding dimensions, with the batch axis omitted.

Backbone	Structured map $M_\theta(X)$	Embedding $z_\theta(X)$	Dataset-dependent length
CNN	$40 \times 2 \times L$	40	$L = \{12, 165, 19\}$
EEGNet	$16 \times 2 \times L$	16	$L = \{7, 80, 12\}$
SincNet	$32 \times T$	32	$T = \{250, 2560, 384\}$

The length entries follow the fixed dataset order P300, BNCI2014002, and SADT.

CNN backbone. The CNN first applies a $1 \to 40$ convolution with kernel $(1, E)$, followed by a $40 \to 40$ convolution with kernel $(E, 1)$, batch normalization over 40 channels, and ELU activation. Two parallel operators summarize the resulting map, both with window 75 and stride

15. The first uses average pooling and dropout with probability 0.5. The second computes $\log(\mathrm{clip}(\mathrm{Var}(\cdot), 10^{-6}, 10^{6}))$ in each window. Here, $\mathrm{Var}(\cdot)$ uses the sample-variance denominator over temporal samples, and $\mathrm{clip}(\cdot, a, b)$ clamps values to $[a, b]$. A shared 1×1 convolution maps both summaries from 40 to 40 channels. After rearrangement and concatenation, they form $M_\theta(X) \in \mathbb{R}^{B\times 40\times 2\times L}$. The average-pooled branch is flattened and projected by a $40L \to 40$ linear bottleneck to produce $z_\theta(X)$. A separate 40-dimensional batch-normalization layer normalizes class-conditioned features outside the ordinary embedding path.

EEGNet backbone. The EEGNet-based encoder begins with a bias-free $1 \to 8$ temporal convolution of kernel $(1, 64)$ and padding $(0, 32)$, followed by batch normalization. A grouped spatial convolution with kernel $(E, 1)$, eight groups, and 16 output maps then collapses the electrode dimension. This block uses batch normalization, ELU, 1×4 average pooling, and dropout with probability 0.25. A second bias-free temporal convolution keeps 16 channels, uses kernel $(1, 16)$ with padding $(0, 8)$, and is followed by batch normalization and ELU. From this final convolutional map, the average branch applies 1×8 pooling and dropout 0.25, while the variance branch applies log-variance pooling with window and stride 8 and the same $[10^{-6}, 10^{6}]$ clamp. Concatenating the two branches yields $M_\theta(X) \in \mathbb{R}^{B\times 16\times 2\times L}$. The flattened average branch is mapped by a $16L \to 16$ linear layer to the embedding. A separate 16-dimensional batch-normalization layer normalizes compensated target features, as in the CNN.

SincNet backbone. The SincNet-inspired encoder first applies a bias-free spatial convolution $1 \to 16$ with kernel $(E, 1)$. After permutation, 24 learnable sinc-parameterized band-pass filters operate along time. Cutoff parameters are initialized on a 0.5–50 Hz lower-frequency grid with a 4 Hz bandwidth. The forward pass uses absolute-value reparameterization with minimum-frequency and minimum-bandwidth offsets and caps the upper cutoff at the normalized bound $100/f_s$. Each band-pass kernel is the difference of two low-pass sinc filters, followed by peak-amplitude normalization and a Hamming window. Here, f_s denotes the EEG sampling frequency in Hz. Even filter lengths are increased to the next odd number. Effective lengths are 51 samples for P300, 27 for SADT, and 103 for BNCI2014002 at 512 Hz.

The sinc output has 24×16 temporal–spatial channels. At each time point, LayerNorm normalizes the reshaped output over $24 \cdot 16 = 384$ features. A bias-free $24 \to 32$ convolution with kernel $(16, 1)$ collapses the spatial-response dimension, followed by ELU. This produces $M_\theta(X) \in \mathbb{R}^{B\times 32\times T}$. After transposition, LayerNorm is applied over the 32 channels and temporal averaging produces the 32-dimensional embedding. The same 32-dimensional LayerNorm module normalizes class-conditioned target features.

Initialization and classifier. For CNN and EEGNet, convolutional weights use Kaiming-uniform initialization, and linear weights use Xavier-normal initialization. Batch-normalization scales follow $\mathcal{N}(1, 0.02^2)$, and available biases are zeroed. Here, $\mathcal{N}(\cdot, \cdot)$ denotes a Gaussian distribution parameterized by its mean and variance. Both ordinary SincNet convolutional layers use Xavier-uniform weight initialization. Every backbone feeds a bias-free linear classifier $W \in \mathbb{R}^{C\times d}$ initialized with Xavier-normal weights, where $d \in \{40, 16, 32\}$ for CNN, EEGNet, and SincNet, respectively.

B.4 Training Configuration and Source Learning

For each LOSO fold, subject-wise EA precedes pooling of all non-target trials. A fixed run seed controls shuffling, followed by a trial-level 70% training and 30% validation split. Adam jointly optimizes the encoder and classifier at learning rates of 10^{-3} and 10^{-2}, respectively. Weight decay is 10^{-4}, batch size is 32, and label smoothing uses $\epsilon = 0.1$. A residual mini-batch of size one is skipped. Validation and prototype construction use batch sizes of 64 and 96, respectively.

Source training runs for the prescribed epoch limit E_s, with validation used to trigger prototype replacement rather than terminate optimization. After epoch e, validation accuracy a_e^{val} triggers prototype replacement whenever

$$a_e^{\mathrm{val}} \geq \max\big(\{0\} \cup \{a_\ell^{\mathrm{val}} : \ell < e\}\big) .$$

The zero initializes the running maximum, and ties trigger replacement. The complete source-training partition is then forwarded without gradients, and within-class map averages define A_c.

Cross-channel mean and sample-variance vectors form $\mathcal{P}_s$. Variance uses denominator $d-1$. Each classifier row is replaced by $\mathrm{AGAP}(A_c)$, and source optimization resumes. These prototypes summarize structured feature maps, rather than empirical class means of classifier embeddings. Subsequent gradients may change the classifier prototypes after replacement.

Target training settings. Target adaptation uses the same optimizer settings as source training, with a maximum of $E_t = 25$ target adaptation epochs. We set the neighbourhood size to $K = 4$, loss weights to $\lambda_{\mathrm{sim}} = 1.6$ and $\lambda_{\mathrm{dis}} = 0.6$, and supervision growth rate to $\gamma = 5$. Pseudo-labels and h-norm are evaluated every five updates on P300 and every update on BNCI2014002 and SADT. The stopping grace period is $g = 5$ evaluation intervals. For numerical evaluation of Equation 20, $\log \rho_c$ is approximated by $\log(\rho_c + \delta)$ with $\delta = 10^{-8}$.

B.5 Target Adaptation and Model Selection

Target adaptation jointly updates the source-initialized encoder and classifier while keeping the source feature-moment memory fixed. Let E_t denote the target adaptation epoch limit and J its corresponding update budget. The target representations and predictions define a neighbourhood memory, initialized from all unlabeled target trials and refreshed as adaptation proceeds. Complete-set pseudo-labels and h-norm are recomputed at prescribed intervals.

Feature compensation and normalization. At each pseudo-label refresh, AGAP reduces the target structured map to m_i. Pooling the stored source moments gives the class-specific scale

$$a_c = \frac{1}{r} \sum_{q=1}^{r} \big[\lambda \mu_{c,q} + (1-\lambda)\nu_{c,q}\big].$$

Here, $\mu_{c,q}$ and $\nu_{c,q}$ are the cross-channel mean and variance of source class prototype A_c at structural coordinate q, and $\lambda \in [0,1]$ balances their contributions. The normalized, class-conditioned target feature is

$$h_{ic} = \mathcal{N}_b(a_c m_i),$$

where $\mathcal{N}_b$ is the compensation-normalization operator and b identifies the backbone. SincNet applies LayerNorm over the d learned features using the same transformation as its embedding path. CNN and EEGNet use separate compensation BatchNorm layers with stored statistics. Each normalized feature is compared with its same-index classifier prototype, and the maximum cosine score determines the pseudo-label.

This scalar modulation does not perform coordinate-wise moment matching. Its effect depends on coefficient signs and the normalization operator's sensitivity to scale. Positive scaling cancels in cosine similarity and, in the zero-stabilizer limit for nonconstant inputs, in LayerNorm. Stored-statistics BatchNorm can retain scale dependence through its effective affine offset. When this offset is zero, positive scaling again cancels in cosine similarity. Changes in moment magnitude therefore need not change pseudo-label assignments.

Target objective. The joint objective in Equation 18 combines neighbourhood attraction, compensation-guided supervision, and inter-class dispersal. Neighbours are selected by cosine similarity between normalized target embeddings, excluding the query itself and without pseudo-class filtering. Their stored predictions remain constant during the current update. For query i, let

$$q_i = \sum_{k \in \mathcal{N}_K(i)} \bar{p}_k.$$

Here, $\bar{p}_k$ is the stored class-probability vector for neighbour k. Each batch refreshes its own representations and predictions before neighbour selection. Summing K probability vectors gives $\sum_{c=1}^{C} q_{ic} = K$. The sum is not divided by K.

Neighbourhood attraction also admits the form

$$\widetilde{\mathcal{L}}_{\mathrm{sim}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{C} q_{ic}(\log q_{ic} - p_{ic}),$$

where B is the batch size, p_{ic} is the current class-c prediction for trial i, and $0 \log 0 = 0$. Relative to Equation 14, this expression adds

$$\kappa(q) = \frac{1}{B} \sum_{i=1}^{B} \sum_{c=1}^{C} q_{ic} \log q_{ic}.$$

Because the neighbour predictions remain fixed within an update, $\kappa(q)$ is constant with respect to the current model parameters. The two forms therefore yield identical parameter gradients, although $\kappa(q)$ may change when the neighbourhood memory is refreshed.

For target update $j \in \{0, \ldots, J-1\}$, the compensation-guided supervision weight is

$$\alpha(j) = 1 - \exp(-\gamma j / J),$$

where $\gamma > 0$ controls its growth rate. The weights λ_{sim} and λ_{dis} respectively scale local agreement and inter-class separation.

Model selection. At each evaluation, pseudo-label class proportions determine h-norm as defined in Equation 20, with $0 \log 0 = 0$. The model with the largest observed h-norm is retained. The first evaluation satisfying $\mathcal{H}(j) \le \tau_h^{(D,b)}$ triggers termination after a grace period of g further evaluations, or at the update budget if reached earlier. Here, D and b identify the dataset and backbone, $\tau_h^{(D,b)}$ is their associated threshold, and g is the fixed number of grace evaluations. If the threshold is never reached, adaptation continues to the budget with the same model-selection criterion. This statistic measures class balance rather than prediction correctness and thus imposes a balanced-class prior. Imbalanced or missing-class target populations require a different prior. Target accuracy is logged during adaptation, but labels do not determine updates, checkpoint selection, or stopping.

C Training and Adaptation Algorithm

Algorithm S1 summarizes source training, moment-memory construction, and target adaptation. Source trials and labels are discarded before target optimization.

Here, $\mathcal{D}_u$ denotes subject u's dataset and E_s the source epoch limit. Stage-local X_i, y_i, and $z_i = z_\theta(X_i)$ omit domain superscripts, and X_i is already aligned when used for feature extraction. The sets $\mathcal{I}_s$ and $\mathcal{I}_c$ index the source-training partition and its class-c trials. Subscripted mean and var denote arithmetic mean and sample variance. $A_{c,k}$ is the vector of structural coordinates in prototype channel k. The memories F and P contain normalized target embeddings and class-probability vectors. Cosine similarity follows Equation 12. The scalar $H^\star$ records the running maximum of h-norm, and $(\theta^\star, W^\star)$ denotes the retained model. The threshold τ_h abbreviates $\tau_h^{(D,b)}$.

Algorithm S1: GeMoT source training and target adaptation

Input: $\{\mathcal{D}_u\}_{u \in \mathcal{U}}$, t, f_θ, W, E_s, E_t, J, B, ϵ, λ, K, λ_{sim}, λ_{dis}, γ, evaluation interval Δ, τ_h, g.
Output: $(\theta^\star, W^\star)$, $\{\arg\max_c p_{\theta^\star, W^\star, c}(X_i^t)\}_{i=1}^{n_t}$.

Source stage

for $u \in \mathcal{U}$ **do**
  $R_u \leftarrow n_u^{-1} \sum_{i=1}^{n_u} \mathrm{Cov}(X_i^u)$; $\widetilde{X}_i^u \leftarrow R_u^{-1/2} X_i^u$, $i = 1, \ldots, n_u$.
end for
$\mathcal{D}_t \leftarrow \mathcal{D}_{u=t}$; $\mathcal{D}_s \leftarrow \bigcup_{u \neq t} \mathcal{D}_u$; $(\mathcal{I}_s, \mathcal{I}_v) \leftarrow \mathrm{Split}(\mathcal{D}_s)$; $(\theta, W) \leftarrow \mathrm{Init}()$.
for $e = 1, \ldots, E_s$ **do**
  for $\mathcal{B}_s \subset \mathcal{I}_s$, $|\mathcal{B}_s| \le B$ **do**; **if** $|\mathcal{B}_s| = 1$ **then continue**.
    $\mathcal{L}_s \leftarrow -|\mathcal{B}_s|^{-1} \sum_{i \in \mathcal{B}_s} \sum_{c=1}^{C} \widetilde{y}_{ic}^s \log p_{ic}^s$; $(\theta, W) \leftarrow \mathrm{Adam}(\theta, W; \nabla_{\theta,W} \mathcal{L}_s)$.
  end for; $a_e^{\text{val}} \leftarrow \mathrm{Acc}(\mathcal{I}_v; \theta, W)$.
  if $a_e^{\text{val}} \ge \max(\{0\} \cup \{a_\ell^{\text{val}} : \ell < e\})$ **then**
    $A_c \leftarrow |\mathcal{I}_c|^{-1} \sum_{i \in \mathcal{I}_c} M_\theta(\widetilde{X}_i^s)$; $\pi_c \leftarrow \mathrm{AGAP}(A_c)$; $w_c \leftarrow \pi_c$, $c = 1, \ldots, C$.
    $\mu_{c,q} \leftarrow d^{-1} \sum_{k=1}^{d} A_{c,kq}$; $\nu_{c,q} \leftarrow (d-1)^{-1} \sum_{k=1}^{d} (A_{c,kq} - \mu_{c,q})^2$.
    $\mathcal{P}_s \leftarrow \{(\mu_c, \nu_c)\}_{c=1}^{C}$; $(\bar{\theta}, \bar{W}, \bar{\mathcal{P}}_s) \leftarrow (\theta, W, \mathcal{P}_s)$.
  end if
end for; $(\theta, W, \mathcal{P}_s) \leftarrow (\bar{\theta}, \bar{W}, \bar{\mathcal{P}}_s)$; $(\mathcal{D}_s, \boldsymbol{y}_s) \leftarrow \varnothing$.

Algorithm S1 (continued)

Target initialization

$F_i \leftarrow z_i/\|z_i\|_2$; $P_i \leftarrow \mathrm{softmax}(Wz_i)$, $i = 1, \ldots, n_t$.
$H^\star \leftarrow 0$; $j_{\mathrm{stop}} \leftarrow \infty$; $(\theta^\star, W^\star) \leftarrow \varnothing$.

Source-free target stage

for $j = 0, \ldots, J-1$ **do**
  if $j \bmod \Delta = 0$ **then**
    $m_i \leftarrow \mathrm{AGAP}(M_\theta(\widetilde{X}_i^t))$; $a_c \leftarrow r^{-1}\sum_{q=1}^{r}[\lambda\mu_{c,q} + (1-\lambda)\nu_{c,q}]$.
    $h_{ic} \leftarrow \mathcal{N}_b(a_c m_i)$; $s_{ic} \leftarrow h_{ic}^\top w_c/(\|h_{ic}\|_2\|w_c\|_2)$; $\widehat{y}_i \leftarrow \arg\max_{c\in\{1,\ldots,C\}} s_{ic}$.
  end if
  $\mathcal{B}_j \leftarrow \mathrm{Sample}(\mathcal{D}_t, B)$; $z_i \leftarrow z_\theta(\widetilde{X}_i^t)$; $p_i \leftarrow \mathrm{softmax}(Wz_i)$, $i \in \mathcal{B}_j$.
  $F_i \leftarrow \mathrm{detach}(z_i/\|z_i\|_2)$; $P_i \leftarrow \mathrm{detach}(p_i)$, $i \in \mathcal{B}_j$.
  $\mathcal{N}_K(i) \leftarrow \mathrm{TopK}_{k\neq i}(F_i^\top F_k)$; $q_i \leftarrow \sum_{k\in\mathcal{N}_K(i)} P_k$; $\nabla_{\theta,W} q_i \leftarrow 0$.
  $\mathcal{L}_{\mathrm{sim}} \leftarrow -B^{-1}\sum_{i\in\mathcal{B}_j} p_i^\top q_i$; $\mathcal{L}_{\mathrm{cgs}} \leftarrow -B^{-1}\sum_{i\in\mathcal{B}_j}\log p_{i,\widehat{y}_i}$.
  $\mathcal{L}_{\mathrm{dis}} \leftarrow B^{-1}\sum_{i,\ell\in\mathcal{B}_j}\mathbb{I}[\widehat{y}_i \neq \widehat{y}_\ell]p_i^\top p_\ell$.
  $\alpha(j) \leftarrow 1 - \exp(-\gamma j/J)$; $\mathcal{L}_t \leftarrow \lambda_{\mathrm{sim}}\mathcal{L}_{\mathrm{sim}} + \alpha(j)\mathcal{L}_{\mathrm{cgs}} + \lambda_{\mathrm{dis}}\mathcal{L}_{\mathrm{dis}}$.
  $(\theta, W) \leftarrow \mathrm{Adam}(\theta, W; \nabla_{\theta,W}\mathcal{L}_t)$; $\mathcal{P}_s \leftarrow \widetilde{\mathcal{P}}_s$.
  if $(j+1) \bmod \Delta = 0$ **or** $j+1 = J$ **then**
    $m_i \leftarrow \mathrm{AGAP}(M_\theta(\widetilde{X}_i^t))$; $a_c \leftarrow r^{-1}\sum_{q=1}^{r}[\lambda\mu_{c,q} + (1-\lambda)\nu_{c,q}]$.
    $h_{ic} \leftarrow \mathcal{N}_b(a_c m_i)$; $s_{ic} \leftarrow h_{ic}^\top w_c/(\|h_{ic}\|_2\|w_c\|_2)$; $\widehat{y}_i \leftarrow \arg\max_{c\in\{1,\ldots,C\}} s_{ic}$.
    $\rho_c(j+1) \leftarrow n_t^{-1}\sum_{i=1}^{n_t}\mathbb{I}[\widehat{y}_i = c]$; $\mathcal{H}(j+1) \leftarrow -(\log C)^{-1}\sum_{c=1}^{C}\rho_c(j+1)\log\rho_c(j+1)$.
    if $\mathcal{H}(j+1) \geq H^\star$ **then** $(H^\star, \theta^\star, W^\star) \leftarrow (\mathcal{H}(j+1), \theta, W)$.
    if $\mathcal{H}(j+1) \leq \tau_h$ **and** $j_{\mathrm{stop}} = \infty$ **then** $j_{\mathrm{stop}} \leftarrow \min\{j+1+g\Delta, J\}$.
    if $j+1 \geq j_{\mathrm{stop}}$ **then break**; **end if**
end for; $(\theta, W) \leftarrow (\theta^\star, W^\star)$; $\widehat{y}_i^\star \leftarrow \arg\max_c p_{\theta^\star,W^\star,c}(\widetilde{X}_i^t)$.

D Supplementary Results

These tables provide the subject-level results underlying the main-text summaries, ordered as P300, BNCI2014002, and SADT. Each GeMoT entry averages repeated runs after h-norm model selection. Dataset summaries give the mean and sample standard deviation across these subject-level averages.

D.1 Subject-Level Method Comparisons

Subject-level accuracies and dataset summaries are provided in Tables S2–S4, allowing cross-subject performance differences to be examined. GeMoT uses EEGNet for every P300 subject, CNN for every BNCI2014002 subject, and SincNet for every SADT subject. These accuracies also supply the matched observations for the paired analysis below.

P300. As shown in Table S2, GeMoT achieves 69.36% cross-subject mean accuracy for eight held-out subjects, exceeding CDAN, the highest-mean baseline, by 1.95 points. It attains the highest accuracy for six subjects, although the mean advantage does not imply improvement for every subject.

Table S2: Subject-level LOSO accuracy (%) on P300. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima. GeMoT uses one fixed backbone for every subject.

Method	S1	S2	S3	S4	S5	S6	S7	S8	Avg.±std
CSP	50.93	49.93	56.14	51.29	53.72	50.79	50.36	50.36	$51.69_{\pm 2.14}$
EA	52.93	52.07	56.21	51.07	55.16	53.93	51.86	52.36	$53.20_{\pm 1.77}$
EEGNet	62.10	51.80	61.40	67.70	61.80	61.80	65.40	62.10	$61.76_{\pm 4.60}$
DAN	63.64	63.71	67.50	64.02	68.48	69.71	76.36	63.73	$67.15_{\pm 4.45}$
DANN	**64.00**	50.00	61.00	67.30	57.70	65.80	67.30	62.50	$61.95_{\pm 5.84}$
CDAN	63.57	62.36	67.50	63.95	69.91	70.71	75.21	66.09	$67.41_{\pm 4.34}$
SHOT	63.60	49.60	64.70	**69.90**	62.50	64.00	66.50	63.20	$63.00_{\pm 5.90}$
ASFA	54.74	52.42	54.09	48.07	54.73	53.24	51.64	54.59	$52.94_{\pm 2.28}$
LSFT	54.14	52.43	58.71	52.00	54.58	53.86	51.86	53.58	$53.90_{\pm 2.20}$
GeMoT	63.07	**65.00**	**69.29**	66.38	**71.85**	**73.93**	**78.22**	**67.17**	$\mathbf{69.36}_{\pm 5.04}$

BNCI2014002. Table S3 presents cross-subject method comparisons for 14 held-out subjects. GeMoT obtains 72.50% mean accuracy, 3.39 points above SHOT, the highest-mean baseline. It ties for the highest accuracy on one subject, and the mean advantage does not imply improvement for every subject.

Table S3: Subject-level LOSO accuracy (%) on BNCI2014002. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima. GeMoT uses one fixed backbone for every subject.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Avg.±std
CSP	55.63	60.00	91.88	81.25	56.88	57.50	83.12	56.88	91.25	65.00	50.63	57.50	50.00	48.13	64.69±15.38
EA	56.25	61.88	93.13	82.50	57.50	66.25	69.38	61.88	90.00	61.88	50.00	56.25	50.00	49.38	64.73±14.34
EEGNet	50.00	50.70	53.70	50.00	51.60	50.00	55.00	50.60	51.80	51.40	50.80	55.10	50.30	86.74	54.12±9.55
DAN	60.63	81.25	88.75	73.75	**71.25**	**74.38**	79.38	66.88	77.50	59.38	63.75	65.00	56.25	47.50	68.97±11.14
DANN	61.88	81.88	86.88	78.13	66.25	65.00	78.13	60.63	80.63	58.13	66.25	67.50	**61.25**	48.13	68.62±10.93
CDAN	63.60	61.00	66.10	56.00	52.20	45.40	75.40	**89.30**	69.00	**70.50**	56.20	**78.80**	59.20	**91.40**	66.72±13.45
SHOT	**67.50**	**83.75**	**99.38**	**85.00**	65.00	61.25	**86.88**	50.63	**93.75**	60.00	51.88	61.88	50.63	50.00	69.11±17.25
ASFA	55.00	82.50	93.75	68.75	62.50	64.38	67.50	50.00	89.38	63.12	54.37	50.62	54.37	43.12	64.24±15.15
LSFT	55.62	82.50	93.75	69.38	65.00	61.25	86.25	54.37	**93.75**	58.13	**72.50**	61.88	51.88	49.38	68.26±15.26
GeMoT	66.88	**83.75**	95.62	79.38	64.38	68.75	81.88	64.38	92.50	65.00	71.25	67.50	60.62	53.12	**72.50**±12.32

SADT. For 11 held-out subjects, GeMoT achieves 82.45% cross-subject mean accuracy, 4.23 points above DAN, the highest-mean baseline, as shown in Table S4. It attains the highest accuracy for three subjects, although the mean advantage does not imply improvement for every subject.

Table S4: Subject-level LOSO accuracy (%) on SADT. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima. GeMoT uses one fixed backbone for every subject.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Avg.±std
CSP	54.79	36.36	75.33	67.57	77.23	98.80	49.02	53.79	72.93	77.78	72.12	66.88±17.17
EA	**83.51**	75.76	72.00	77.70	76.79	82.53	58.82	67.05	83.12	74.07	68.58	74.54±7.61
EEGNet	72.87	62.12	58.00	72.29	62.94	77.10	**72.54**	68.93	70.89	84.55	66.81	69.91±7.42
DAN	78.72	76.52	73.33	70.27	**92.86**	84.94	63.73	75.38	83.44	84.26	76.99	78.22±7.96
DANN	57.97	56.81	62.00	64.86	64.73	66.26	67.64	58.71	58.28	70.37	63.27	62.81±4.45
CDAN	75.29	81.86	73.50	69.82	87.02	81.64	60.21	75.49	82.44	89.94	83.11	78.21±8.46
SHOT	76.60	73.48	74.67	71.62	90.18	84.34	67.65	64.02	**87.26**	84.26	65.93	76.36±8.98
ASFA	67.98	**82.73**	76.93	75.20	80.89	96.02	49.61	69.47	83.34	69.26	58.85	73.66±12.68
LSFT	82.98	80.30	**83.33**	**80.41**	78.57	**100.00**	63.73	84.09	83.12	65.74	56.19	78.04±12.01
GeMoT	76.06	78.03	74.67	72.97	89.73	99.40	66.67	**84.85**	83.76	**95.37**	**85.40**	**82.45**±9.92

D.2 Subject-Paired Statistical Comparisons

For each baseline, the paired difference is GeMoT accuracy minus baseline accuracy for the same held-out subject. A two-sided paired Student's t-test evaluates whether the population mean difference is zero. P300, BNCI2014002, and SADT provide $n = 8$, 14, and 11 pairs, with 7, 13, and 10 degrees of freedom, respectively. Repeated runs are averaged within subjects and are not additional independent observations.

Holm's step-down procedure adjusts the nine comparisons separately within each dataset. Adjusted $p < 0.05$ defines significance. Tables S5–S7 present the mean difference Δ in percentage points, its marginal 95% confidence interval (CI), and the t-statistic. They also list the raw and adjusted p-values, denoted p_{raw} and p_{Holm}. Positive Δ favours GeMoT. The confidence intervals are not multiplicity-adjusted. Here, df denotes degrees of freedom.

All 27 comparisons have positive mean paired differences, but only 9 of 9 on P300, 6 of 9 on BNCI2014002, and 3 of 9 on SADT meet the adjusted threshold. Higher means therefore do not establish significant superiority over every baseline. The marginal confidence intervals and multiplicity-adjusted significance decisions answer different inferential questions.

Paired-t inference assumes approximately normal subject-level differences. Small sample sizes and overlapping source-training pools across LOSO folds limit strict inferential independence. Furthermore, the comparisons retain method-specific architectures, training, and source-access regimes. The tests quantify differences between these configurations and do not isolate the contribution of each factor.

P300. Table S5 summarizes paired comparisons between GeMoT and the nine baselines using eight matched subjects.

Table S5: Two-sided paired t-tests of GeMoT against nine baselines on P300 ($n = 8$, $\mathrm{df} = 7$). Δ is mean GeMoT accuracy minus baseline (percentage points). CI denotes the unadjusted 95% paired-t confidence interval. Both raw and within-dataset Holm-adjusted p-values are shown. Boldface marks adjusted $p < 0.05$.

Baseline	Δ	95% CI	t	p_{raw}	p_{Holm}
CSP	17.67	[13.21, 22.13]	9.371	3.28×10^{-5}	$\mathbf{2.95 \times 10^{-4}}$
EA	16.16	[11.95, 20.38]	9.073	4.05×10^{-5}	$\mathbf{3.24 \times 10^{-4}}$
EEGNet	7.60	[2.97, 12.24]	3.878	0.00607	**0.01820**
DAN	2.22	[0.96, 3.47]	4.177	0.00415	**0.01662**
DANN	7.41	[2.29, 12.54]	3.420	0.01113	**0.02226**
CDAN	1.95	[0.94, 2.96]	4.562	0.00260	**0.01300**
SHOT	6.36	[1.02, 11.71]	2.816	0.02594	**0.02594**
ASFA	16.42	[11.72, 21.13]	8.251	7.48×10^{-5}	$\mathbf{5.23 \times 10^{-4}}$
LSFT	15.47	[10.75, 20.18]	7.754	1.11×10^{-4}	$\mathbf{6.67 \times 10^{-4}}$

BNCI2014002. The BNCI2014002 comparisons use 14 matched subjects, with paired differences and test results shown in Table S6.

Table S6: Two-sided paired t-tests of GeMoT against nine baselines on BNCI2014002 ($n = 14$, $\mathrm{df} = 13$). Δ is mean GeMoT accuracy minus baseline (percentage points). CI denotes the unadjusted 95% paired-t confidence interval. Both raw and within-dataset Holm-adjusted p-values are shown. Boldface marks adjusted $p < 0.05$.

Baseline	Δ	95% CI	t	p_{raw}	p_{Holm}
CSP	7.81	[3.40, 12.22]	3.825	0.00210	**0.01302**
EA	7.77	[3.53, 12.00]	3.961	0.00163	**0.01302**
EEGNet	18.38	[7.82, 28.94]	3.760	0.00238	**0.01302**
DAN	3.53	[0.28, 6.77]	2.345	0.03554	0.10662
DANN	3.88	[1.74, 6.03]	3.907	0.00180	**0.01302**
CDAN	5.78	[−5.81, 17.37]	1.077	0.30100	0.30100
SHOT	3.39	[−0.84, 7.62]	1.731	0.10711	0.21422
ASFA	8.26	[4.82, 11.70]	5.184	1.76×10^{-4}	**0.00158**
LSFT	4.24	[1.34, 7.14]	3.158	0.00755	**0.03020**

SADT. As shown in Table S7, the SADT analysis compares GeMoT with each baseline using 11 matched subjects.

Table S7: Two-sided paired t-tests of GeMoT against nine baselines on SADT ($n = 11$, $\mathrm{df} = 10$). Δ is mean GeMoT accuracy minus baseline (percentage points). CI denotes the unadjusted 95% paired-t confidence interval. Both raw and within-dataset Holm-adjusted p-values are shown. Boldface marks adjusted $p < 0.05$.

Baseline	Δ	95% CI	t	p_{raw}	p_{Holm}
CSP	15.56	[7.09, 24.03]	4.093	0.00217	**0.01517**
EA	7.91	[1.29, 14.52]	2.663	0.02379	0.14272
EEGNet	12.53	[6.00, 19.06]	4.277	0.00162	**0.01296**
DAN	4.23	[0.34, 8.11]	2.426	0.03568	0.14272
DANN	19.64	[13.15, 26.13]	6.741	5.10×10^{-5}	$\mathbf{4.59 \times 10^{-4}}$
CDAN	4.23	[0.45, 8.02]	2.491	0.03191	0.14272
SHOT	6.08	[0.10, 12.06]	2.266	0.04688	0.14272
ASFA	8.78	[1.26, 16.30]	2.603	0.02637	0.14272
LSFT	4.40	[−4.70, 13.51]	1.078	0.30627	0.30627

D.3 Backbone Comparisons

As shown in Tables S8–S10, cross-subject backbone performance varies within each dataset. The adaptation framework remains the same for CNN, EEGNet, and SincNet. The highest mean changes with the dataset, and no encoder is uniformly best for individual subjects.

P300.

Table S8: Subject-level backbone accuracy (%) of GeMoT on P300. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

Backbone	S1	S2	S3	S4	S5	S6	S7	S8	Avg.±std
CNN	**64.57**	57.57	61.71	64.09	61.25	67.29	74.14	63.88	$64.31_{\pm 4.89}$
EEGNet	63.07	**65.00**	**69.29**	**66.38**	**71.85**	**73.93**	**78.22**	**67.17**	$\mathbf{69.36}_{\pm 5.04}$
SincNet	50.43	53.79	50.57	56.01	51.07	52.71	53.30	52.72	$52.58_{\pm 1.88}$

BNCI2014002.

Table S9: Subject-level backbone accuracy (%) of GeMoT on BNCI2014002. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

Backbone	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Avg.±std
CNN	**66.88**	**83.75**	**95.62**	**79.38**	**64.38**	68.75	81.88	64.38	92.50	**65.00**	**71.25**	67.50	60.62	**53.12**	$\mathbf{72.50}_{\pm 12.32}$
EEGNet	58.13	75.00	86.25	68.12	63.12	68.12	73.75	51.88	93.12	**65.00**	59.38	**69.38**	53.12	45.62	$66.43_{\pm 13.01}$
SincNet	50.00	82.50	86.88	**79.38**	58.75	**74.38**	**88.75**	**66.25**	**94.38**	50.00	53.12	63.75	**66.88**	46.25	$68.66_{\pm 15.95}$

SADT.

Table S10: Subject-level backbone accuracy (%) of GeMoT on SADT. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

Backbone	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Avg.±std
CNN	85.64	**84.09**	**76.67**	**80.41**	87.95	92.77	64.71	82.58	86.62	83.33	77.43	$82.02_{\pm 7.37}$
EEGNet	**86.17**	79.55	70.67	77.70	86.61	91.57	**66.67**	71.21	**86.94**	83.33	76.99	$79.76_{\pm 7.94}$
SincNet	76.06	78.03	74.67	72.97	**89.73**	**99.40**	**66.67**	**84.85**	83.76	**95.37**	**85.40**	$\mathbf{82.45}_{\pm 9.92}$

D.4 Target-Loss Ablations

We evaluate neighbourhood attraction ($\mathcal{L}_{\text{sim}}$), compensation-guided supervision ($\mathcal{L}_{\text{cgs}}$), and inter-class dispersal ($\mathcal{L}_{\text{dis}}$) individually and in every possible combination. The complete objective has the highest mean on each dataset, but some subjects favour reduced objectives, as shown in Tables S11–S13. The strongest reduced configuration differs across datasets, supporting complementary roles for the losses without implying universal subject-level dominance. These ablations are descriptive. Their marginal differences are not separately tested.

P300. As shown in Table S11, the complete objective achieves the highest mean accuracy among the seven loss combinations, at 69.36%. This exceeds the strongest reduced configuration (68.59%) by 0.77 points, although some subjects achieve higher accuracy with reduced objectives.

Table S11: Subject-level accuracy (%) for target-loss ablations on P300. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cgs}}$	$\mathcal{L}_{\text{dis}}$	S1	S2	S3	S4	S5	S6	S7	S8	Avg.±std
✓	×	×	**63.79**	63.93	65.93	58.51	**71.98**	74.14	76.43	66.87	$67.70_{\pm 6.02}$
×	✓	×	58.15	64.86	62.79	57.80	68.90	72.28	74.71	61.87	$65.17_{\pm 6.28}$
×	×	✓	61.43	63.36	68.07	61.73	71.12	73.42	75.14	64.37	$67.33_{\pm 5.39}$
✓	✓	×	60.15	64.15	68.79	65.09	71.70	**75.14**	77.01	66.73	$68.59_{\pm 5.73}$
✓	×	✓	61.43	64.86	67.93	59.94	71.41	74.35	76.65	66.80	$67.92_{\pm 5.93}$
×	✓	✓	60.08	**65.72**	63.21	60.37	70.05	74.21	75.36	65.30	$66.79_{\pm 5.87}$
✓	✓	✓	63.07	65.00	**69.29**	**66.38**	71.85	73.93	**78.22**	**67.17**	$\mathbf{69.36}_{\pm 5.04}$

BNCI2014002. Table S12 lists subject-level accuracies for all seven loss combinations. The complete objective yields the highest mean of 72.50%, exceeding the strongest reduced configuration (72.03%) by 0.47 points. Individual subjects do not uniformly favour the complete objective.

Table S12: Subject-level accuracy (%) for target-loss ablations on BNCI2014002. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cgs}}$	$\mathcal{L}_{\text{dis}}$	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	Avg.±std
✓	×	×	47.88	81.62	93.50	78.50	55.38	63.50	83.50	73.50	91.00	61.62	57.88	51.00	47.25	46.00	66.58±16.75
×	✓	×	63.50	79.12	87.25	76.62	67.25	70.38	84.75	73.50	88.50	60.38	58.50	**68.50**	56.00	49.75	70.29±12.02
×	×	✓	59.75	82.88	88.50	75.38	77.88	**72.25**	88.50	77.88	91.62	60.38	61.00	67.25	57.88	47.25	72.03±13.50
✓	✓	×	48.50	82.25	71.62	**79.75**	69.12	69.12	88.50	68.50	91.00	61.62	56.00	47.88	47.25	46.00	66.22±15.56
✓	×	✓	61.62	81.62	94.12	64.12	**79.12**	67.88	**89.12**	75.38	91.00	64.75	54.75	67.25	59.75	49.75	71.45±13.86
×	✓	✓	61.00	82.25	89.75	64.12	77.88	**72.25**	85.38	**78.50**	91.00	60.38	58.50	**68.50**	59.75	49.75	71.36±12.98
✓	✓	✓	**66.88**	**83.75**	**95.62**	79.38	64.38	68.75	81.88	64.38	**92.50**	**65.00**	**71.25**	67.50	**60.62**	**53.12**	**72.50**±12.32

SADT. Among the seven loss combinations, the complete objective reaches the highest mean accuracy of 82.45%, as shown in Table S13. Its mean exceeds that of the strongest reduced configuration (82.29%) by 0.16 points, but this advantage does not extend to every subject.

Table S13: Subject-level accuracy (%) for target-loss ablations on SADT. The final column gives the cross-subject mean and sample standard deviation. Boldface marks column maxima.

$\mathcal{L}_{\text{sim}}$	$\mathcal{L}_{\text{cgs}}$	$\mathcal{L}_{\text{dis}}$	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	Avg.±std
✓	×	×	70.21	**90.15**	50.67	68.92	69.65	98.19	61.77	54.55	53.18	64.82	60.17	67.48±14.90
×	✓	×	72.87	70.45	60.00	75.00	78.13	94.58	60.79	65.15	75.79	75.93	70.79	72.68±9.50
×	×	✓	77.12	78.78	**75.33**	**79.73**	87.06	98.19	62.75	73.87	84.07	88.89	84.07	80.90±9.22
✓	✓	×	70.74	72.72	48.67	68.92	66.08	98.19	60.79	57.20	60.82	62.96	59.73	66.07±12.62
✓	×	✓	**85.82**	88.66	72.33	74.70	88.52	95.19	66.61	75.03	**84.89**	91.45	81.95	82.29±8.97
×	✓	✓	78.72	75.75	70.67	78.38	83.49	96.38	62.75	66.67	84.07	79.63	73.00	77.23±9.20
✓	✓	✓	76.06	78.03	74.67	72.97	**89.73**	**99.40**	**66.67**	**84.85**	83.76	**95.37**	**85.40**	**82.45**±9.92

E SUPPLEMENTARY HYPERPARAMETER SENSITIVITY

The dataset-specific curves in Figures S1–S3 provide detailed views of the one-parameter analyses summarized in Figure 2. Unless varied, parameters remain at $K = 4$, $\lambda_{\text{sim}} = 1.6$, and $\lambda_{\text{dis}} = 0.6$. Each point represents subject-mean LOSO accuracy after h-norm selection. The GeMoT backbones are EEGNet, CNN, and SincNet for P300, BNCI2014002, and SADT, respectively. These results characterize the evaluated parameter ranges rather than joint hyperparameter robustness.

Local agreement and neighbourhood size. P300 varies by at most 1.29 points as any one parameter varies. BNCI2014002 stays within 70.39–72.50% for $K \in [1, 7]$ and 70.25–72.50% for $\lambda_{\text{sim}} \in [0.4, 2.8]$. Its declines at larger values are consistent with overly broad or strongly weighted neighbourhood constraints, although these analyses do not identify the mechanism. SADT retains 80.40–82.57% for $K \in [2, 10]$ and 80.97–82.73% for $\lambda_{\text{sim}} \in [0.8, 4.0]$. The moderate regions thus support local consistency without requiring a single shared optimum across datasets.

Inter-class separation. For $\lambda_{\text{dis}} \in [0.3, 1.0]$, accuracy spans 70.43–72.50% on BNCI2014002 and 81.51–82.61% on SADT. At $\lambda_{\text{dis}} = 0.1$, it falls to 65.34% and 68.79%, respectively. The contrast suggests that adequate separation strength helps balance the attraction and pseudo-label supervision objectives. BNCI2014002 reaches the highest accuracy for each parameter at the configured value. P300 and SADT remain within 0.71 and 0.29 points of their respective maxima for individual parameters, indicating useful performance around the selected settings.

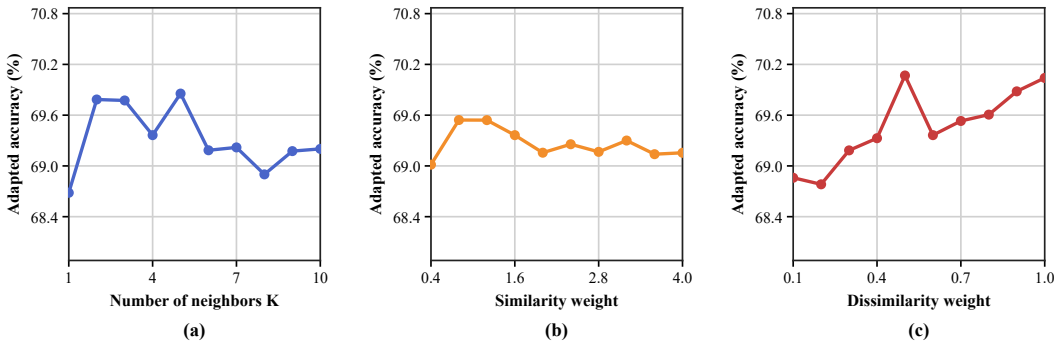


Figure S1: Sensitivity of GeMoT with the EEGNet backbone on P300. Panels vary (a) K, (b) λ_{sim}, and (c) λ_{dis}. Other parameters remain fixed. Mean accuracy ranges are 68.68–69.86%, 69.02–69.54%, and 68.78–70.07%, respectively.

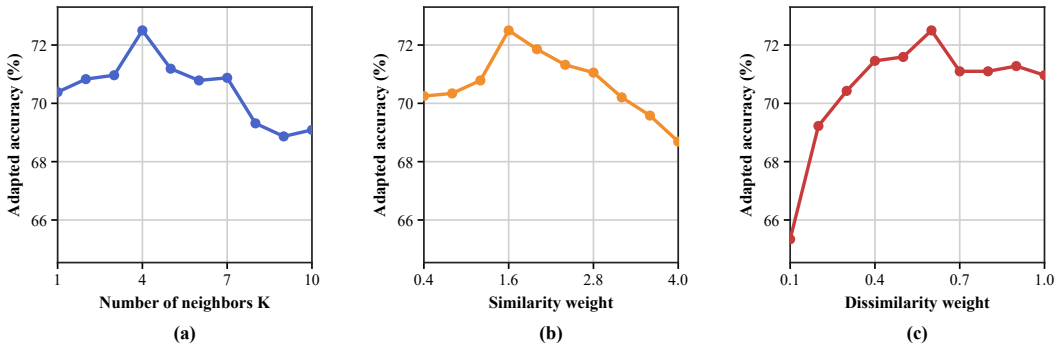


Figure S2: Sensitivity of GeMoT with the CNN backbone on BNCI2014002. Panels vary (a) K, (b) λ_{sim}, and (c) λ_{dis}. Other parameters remain fixed. Mean accuracy ranges are 68.87–72.50%, 68.69–72.50%, and 65.34–72.50%, respectively. Each configured value yields the highest accuracy in its panel.

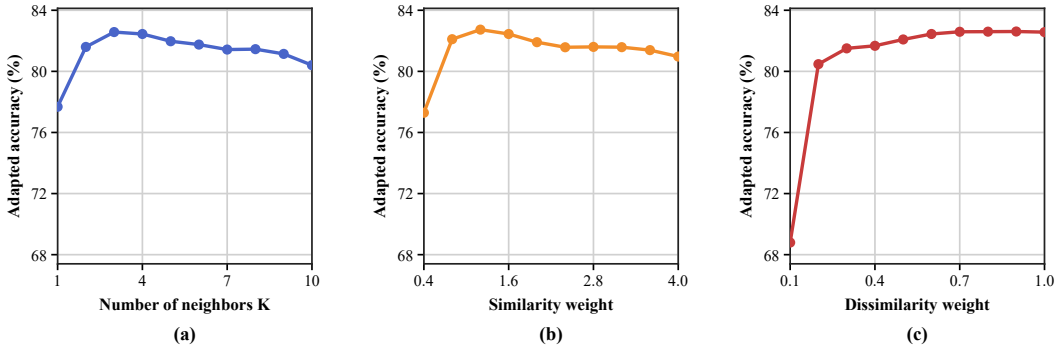


Figure S3: Sensitivity of GeMoT with the SincNet backbone on SADT. Panels vary (a) K, (b) λ_{sim}, and (c) λ_{dis}. Other parameters remain fixed. Mean accuracy ranges are 77.68–82.57%, 77.29–82.73%, and 68.79–82.61%, respectively.