MCD-EEG: CAN A NON-STATIONARY SPATIO-TEMPORAL EEG BENCHMARK EVALUATE BRAINPRINT RECOGNITION ACROSS TASKS, RE-CAPS, AND POPULATIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG), owing to its high temporal resolution, non-invasive acquisition, convenient deployment, and compatibility with wearable devices, has become a fundamental technology for brain–computer interfaces (BCIs) and task-based brainprint recognition. However, existing public EEG datasets are generally limited to a single cognitive task and a single acquisition session, lacking realistic distribution shifts across tasks, populations, and acquisition stages (e.g., EEG cap removal and reattachment, referred to as *re-cap*). Consequently, they are insufficient for systematically evaluating the robustness, generalization, and fairness of EEG representation learning and brainprint recognition models in real-world scenarios. To address these challenges, we introduce **MCD-EEG**, a large-scale EEG benchmark designed for cross-temporal robust and neurodiversity-aware representation learning. MCD-EEG contains recordings from 45 participants, including 22 prelingually deaf individuals and 23 hearing controls, collected using 60-channel high-density EEG together with behavioral measurements. A multi-stage cognitive paradigm consisting of reading, interference, and delayed recall tasks was employed, with repeated recordings acquired across sessions. Notably, the second session involved complete EEG cap removal and reattachment, explicitly introducing realistic cross-temporal session drift. Based on this dataset, we further establish standardized benchmark tasks for within-task, cross-task, and cross-re-cap brainprint recognition, providing a new testbed for developing robust and fair EEG representation learning algorithms. The anonymous code repository is available at `https://mcd-eeg.netlify.app/`

1 INTRODUCTION

In recent years, advances in deep representation learning have substantially promoted research on electroencephalography (EEG)-based biometric decoding and brainprint recognition (Wachinger et al., 2015; Wang et al., 2020; Jouandet et al., 1989). Compared with conventional biometric traits, EEG signals encode individualized neural dynamics and possess an inherent resistance to forgery, making them highly promising for secure authentication and brain–computer interface (BCI) applications. Nevertheless, a significant gap still exists between current EEG benchmarks and real-world deployment scenarios (Barbera et al., 2025; Mihajlović et al., 2014).

Most existing brainprint datasets are collected under stationary cognitive settings (Erdogmus et al., 2005), typically involving resting-state or single-task paradigms, and therefore fail to capture the cross-task (Zhang et al., 2018; Zhou et al., 2022) and cross-temporal (Khorasani & Demazure, 2026; Lu, 2020) distribution shifts commonly encountered in real-world applications. In particular, electrode removal and reattachment (*re-cap*) introduces substantial cross-session physical variations that can severely degrade model performance, yet this factor is largely overlooked in existing datasets. Moreover, current datasets exhibit neurotypicality bias, with limited representation of populations exhibiting neural reorganization, such as prelingually deaf individuals (Neo et al., 2023; Jiang et al., 2026; Fayyazi et al., 2025; Kaplan-Kahn et al., 2021), potentially leading to hidden group-level biases and limiting the development of fair and trustworthy EEG biometrics.

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.

Table 1: Comparison of representative public EEG datasets and MCD-EEG across key evaluation dimensions.

Dataset	Cross-Task	Re-cap EEG Cap	Longitudinal Session	Cross-Population
PhysioNet MI	Limited	×	×	×
DEAP	Limited	×	×	×
DREAMER	Limited	×	×	×
SEED	Limited	Limited	✓	×
BCI IV-2a	Limited	×	✓	×
MAMEM	×	×	×	×
AMIGOS	Limited	×	×	×
MCD-EEG (Ours)	✓	✓	✓	✓

Table 1 compares MCD-EEG with existing datasets.

Representative public EEG datasets exhibit notable limitations. The PhysioNet EEG Motor Movement/Imagery dataset (Schalk et al., 2004) collects all sessions within a single experiment, limiting long-term stability evaluation. DEAP (Koelstra et al., 2011) and DREAMER (Katsigiannis & Ramzan, 2017) contain only single-session or small-scale recordings, while the SEED series (Duan et al., 2013; Jiang et al., 2025; Liu et al., 2022; Schaefer et al., 2010) and SEED-IV (Zheng, 2016) are restricted to emotion recognition without cross-task settings. The BCI Competition IV-2a dataset (Tangermann et al., 2012) focuses solely on motor imagery, and MAMEM (Oikonomou et al., 2016) targets SSVEP paradigms with constrained task diversity. Although AMIGOS (Miranda-Correa et al., 2018) provides more naturalistic recordings, it lacks long-term repeated sessions. Overall, existing public EEG datasets remain limited by insufficient longitudinal recordings, restricted cognitive diversity, small cohorts, and inadequate real-world coverage, hindering comprehensive evaluation of brainprint recognition systems.

To bridge these gaps, we introduce **MCD-EEG**, a large-scale EEG benchmark designed for modeling cross-temporal distribution shifts and neurodiversity. MCD-EEG consists of recordings from 45 participants, including 22 prelingually deaf individuals and 23 normal-hearing controls. EEG data were collected using a multi-stage cognitive paradigm involving reading, interference, and delayed recall tasks, with repeated recordings acquired at different time points. Importantly, the benchmark explicitly incorporates electrode removal and reattachment procedures to simulate realistic cross-temporal session drift.

The main contributions of this work are summarized as follows:

1. **The first public benchmark explicitly modeling cross-temporal re-cap brainprint recognition.** We systematically characterize the realistic session-level distribution shifts induced by electrode reattachment, providing a more deployment-oriented evaluation environment for EEG-based biometric identification.
2. **A standardized EEG benchmark for dynamic cognition and cross-task interference.** By employing a multi-stage cognitive paradigm, MCD-EEG introduces complex brain-state transitions and substantially increases the difficulty of representation learning and model generalization.
3. **A Neurodiversity-Aware Evaluation Benchmark.** By incorporating both prelingually deaf and normal-hearing populations, MCD-EEG enables group-wise comparative analysis and provides a fairness auditing benchmark for trustworthy EEG representation learning.

2 Dataset Construction and Problem Formulation

2.1 Participants and Dataset Composition

The proposed **MCD-EEG** dataset comprises high-density EEG recordings from 45 participants, including 22 prelingually deaf individuals and 23 age-matched healthy controls. All participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological disorders. EEG signals were recorded using a 60-channel high-density acquisition system to ensure

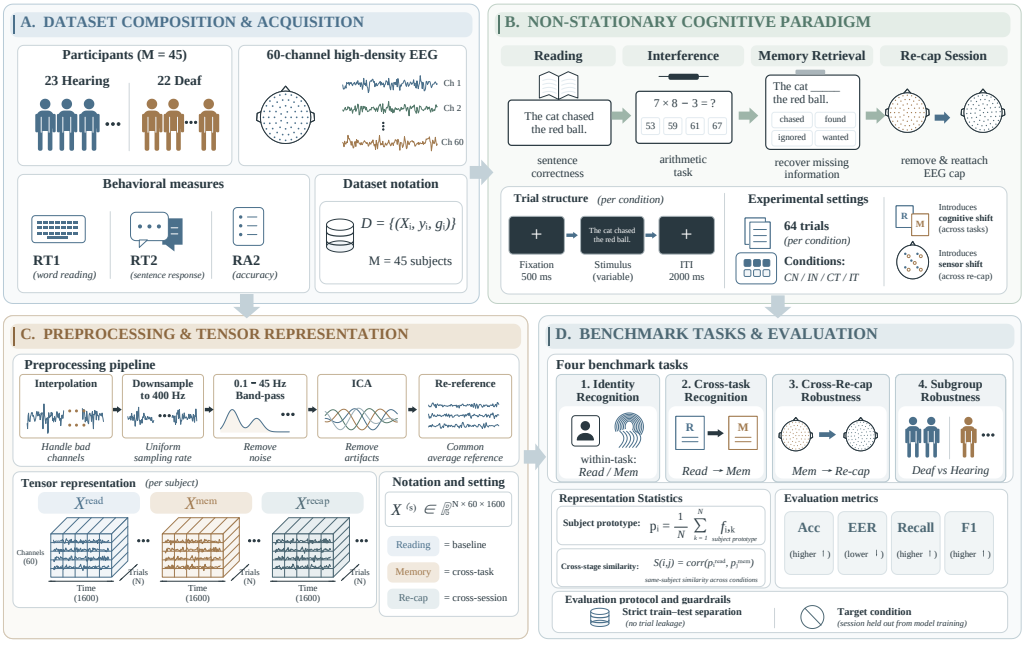


Figure 1: Overview of MCD-EEG. (A) Dataset composition and multimodal acquisition from 45 participants, including 23 hearing and 22 deaf individuals. (B) A multi-stage cognitive paradigm spanning reading, interference, memory retrieval, and re-cap recordings, introducing both cognitive and acquisition shifts. (C) EEG preprocessing and unified tensor construction across tasks and sessions. (D) Benchmark protocols covering identity recognition, cross-task recognition, cross-re-cap robustness, and subgroup robustness, with representation-level analysis and standard evaluation metrics.

sufficient spatial resolution and to comprehensively capture individual-specific neural characteristics. Prior to the experiment, demographic information and standardized cognitive assessments were collected to ensure baseline comparability between the two groups.The construction and benchmark evaluation framework of the MCD-EEG dataset is illustrated in Fig. 1.

The participants were finally divided into two subpopulations:

$$g = \begin{cases} 0, & \text{Hearing group,} \\ 1, & \text{Deaf group.} \end{cases} \tag{1}$$

2.2 Non-Stationary Cognitive Paradigm

To capture realistic cognitive and sensor-induced distribution shifts, we designed a four-stage experimental paradigm with 64 trials per participant, including Reading, Interference, Memory Retrieval, and Re-cap stages. During the Reading stage, participants judged sentence correctness and the response time (RT_1) was recorded. The Interference stage introduced transient cognitive variations through an arithmetic task. In the Memory Retrieval stage, participants recovered missing sentence information, with retrieval response time (RT_2) and accuracy (RA_2) recorded. Stimuli followed a 2×2 factorial design, including four conditions: *CN*, *IN*, *CT*, and *IT*. Each trial contained a 500-ms fixation period and a 2000-ms inter-trial interval to ensure temporal alignment. To further evaluate brainprint robustness against sensor-related distribution shifts, an additional *Re-cap* stage was conducted. After completing the first three stages, participants removed and then re-wore the EEG cap before performing a standardized IQ assessment task. EEG signals were continuously recorded throughout the assessment, and 2000-ms EEG segments were extracted for analysis. This stage was specifically designed to simulate realistic electrode repositioning and physical sensor displacement, enabling the evaluation of brainprint stability under hardware-induced distribution shifts.

2.3 PREPROCESSING AND TENSOR REPRESENTATION

Raw EEG signals underwent standard preprocessing (interpolation, downsampling to 400 Hz, 0.1–45 Hz band-pass filtering, ICA, and re-referencing). The benchmark dataset is formulated as $\mathcal{D} = \{(\mathbf{X}_i, y_i, g_i)\}_{i=1}^{M}$, where $M = 45$ subjects, y_i is the identity label, and $g_i \in \{0, 1\}$ indicates the neurodiversity subgroup.

To simulate real-world wearable deployment, participants underwent a cross-session re-setup protocol involving complete EEG cap removal, reattachment (*re-cap*), and recalibration. Consequently, each participant provides three structured EEG tensors $\mathbf{X}^{\text{read}}, \mathbf{X}^{\text{mem}}, \mathbf{X}^{\text{recap}} \in \mathbb{R}^{N \times C \times T}$ (N trials, $C = 60$ channels, $T = 1600$ time points). $\mathbf{X}^{\text{read}}$ serves as the baseline for representation learning, while $\mathbf{X}^{\text{mem}}$ and $\mathbf{X}^{\text{recap}}$ evaluate cross-task (cognitive shift) and cross-session (electrode repositioning) robustness, respectively.

2.4 PROBLEM FORMULATION AND EVALUATION GUARDRAILS

Given multi-condition observations, the objective is to learn a representation function $f_\theta : \mathbf{X}_i^{(*)} \rightarrow \mathbf{z}_i$ that satisfies three core properties: 1) **Identity consistency** across cognitive states ($\mathbf{z}_i \approx \mathbf{z}_i^{\text{read}} \approx \mathbf{z}_i^{\text{mem}} \approx \mathbf{z}_i^{\text{recap}}$); 2) **Cross-session robustness** against physical shifts ($\mathbf{z}_i^{(1)} \approx \mathbf{z}_i^{(2)}$); 3) **Subgroup invariance** for fairness ($P(\hat{y} \mid g = 0) \approx P(\hat{y} \mid g = 1)$).

Thus, MCD-EEG explicitly formulates a *multi-shift representation learning benchmark*. Furthermore, to ensure realistic Edge AI evaluation, strict train-test separation is enforced, absolutely prohibiting test-set leakage during feature extraction, or model optimization.

2.5 BENCHMARK TASKS

We define four benchmark tasks to evaluate MCD-EEG from complementary perspectives, including population-level neurodiversity characterization, within-task identity discrimination, cross-task generalization, and cross-session robustness.

- **Population Classification.** EEG functional-connectivity representations are used to distinguish hearing and prelingually deaf participants, providing a population-level evaluation of neurodiversity-related neural differences.
- **Within-task Brainprint Recognition.** Models are trained and tested within the same cognitive stage, including *Read-to-Read* and *Mem-to-Mem* identification, to evaluate identity discrimination under stationary task conditions.
- **Cross-task Brainprint Recognition.** Models are trained on the Reading stage and tested on the Memory Retrieval stage (*Read* $\rightarrow$ *Mem*) to evaluate identity generalization across cognitive tasks.
- **Cross-Re-cap Brainprint Recognition.** Models are trained on the Memory Retrieval stage and tested on the Re-cap session (*Mem* $\rightarrow$ *Re-cap*) to evaluate robustness to cross-session sensor shifts induced by EEG cap removal and reattachment.

For the three brainprint-recognition benchmarks, performance is additionally reported separately for the hearing and prelingually deaf groups to assess subgroup robustness and potential performance disparities.

2.6 BEHAVIORAL AND NEUROPHYSIOLOGICAL VALIDATION

Mixed-effects analyses confirm the paradigm successfully induces structured cognitive perturbations. Tongue-twister sentences significantly increased reading latency ($p < 0.001$), while semantic incongruence impaired memory retrieval latency ($p < 0.001$) and accuracy ($p = 0.021$), with distinct group-by-condition interactions. Furthermore, Event-Related Potential (ERP) analyses validated signal quality. While the hearing group exhibited canonical N400 responses (300–700 ms) and the deaf group showed distinct fronto-central positivities (early $\sim$80 ms and late 500–1500 ms), ERP latencies remained highly stable across subjects and sessions. Together, these results demonstrate that MCD-EEG provides high-quality, temporally reliable recordings ideal for evaluating cross-task EEG representation learning. The detailed results are presented in Appendix Table 8.

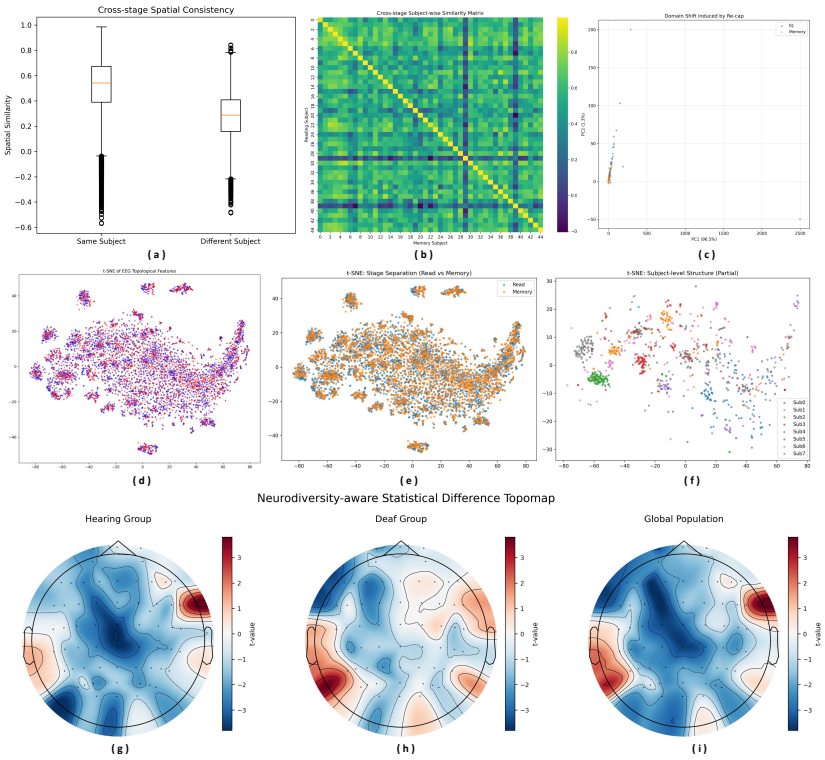


Figure 2: Cross-stage EEG distribution analysis.

2.7 Spatial Stability and Topological Consistency

To evaluate whether MCD-EEG preserves stable individual-specific neural structures under non-stationary cognitive perturbations, we systematically analyzed EEG spatial topology across the Reading and Memory stages.

2.7.1 Cross-stage Spatial Similarity

Differential entropy (DE) was adopted to characterize channel-wise spatial topology:

$$DE_c = \frac{1}{2}\log(2\pi e\sigma_c^2), \tag{2}$$

where σ_c^2 denotes the signal variance of channel c. Each trial was represented as a spatial topological vector $\mathbf{f} = [DE_1, \ldots, DE_C]$. Pearson correlation analysis revealed that the within-subject cross-stage similarity (0.521 ± 0.207) was significantly higher than the between-subject similarity (0.280 ± 0.181; $t = 251.31$, $p < 0.001$). As illustrated in Fig. 2(a), individual spatial topologies remain highly consistent across cognitive stages while exhibiting substantial inter-subject variability, indicating that MCD-EEG effectively preserves identity-specific neural signatures.

2.7.2 Subject-wise Similarity Matrix Analysis

To further examine identity consistency, subject prototypes were constructed by averaging all trials:

$$\mathbf{p}_i = \frac{1}{N}\sum_{k=1}^{N}\mathbf{f}_{i,k}. \tag{3}$$

The cross-stage similarity matrix,

$$S(i,j) = corr(\mathbf{p}_i^{read}, \mathbf{p}_j^{memory}), \tag{4}$$

exhibited strong diagonal dominance (Fig. 2(b)), demonstrating that the highest similarity consistently occurred between recordings from the same subject. This result further confirms the stability of individual spatial representations across cognitive stages.

2.7.3 Analysis of Re-cap-induced Distribution Shifts

As shown in Fig. 2(c), EEG signals acquired after electrode removal and reattachment (Re-cap) exhibit substantial distribution shifts relative to the Memory stage. Quantitative analysis revealed an average channel-wise Wasserstein distance of 0.095 (maximum: 0.162), indicating moderate marginal signal perturbations. However, the average cross-session identity similarity dropped to -0.085, suggesting an almost complete loss of identity consistency after re-cap. Moreover, PCA analysis showed that the first principal component explained 96.5% of the total variance, implying that session drift dominates the feature space. These findings indicate that re-cap introduces a structural distribution shift that fundamentally disrupts brainprint representations and constitutes the major challenge for real-world deployment.

2.7.4 Topological Separability Analysis

As illustrated in Fig. 2(e)–(f), Reading and Memory samples exhibit substantial overlap in the global feature space, suggesting that task effects are relatively weak. In contrast, EEG features from different subjects form compact and well-separated clusters, demonstrating strong spatial separability and highly stable individual-specific structures. These observations indicate that individual effects substantially outweigh task effects, highlighting the suitability of MCD-EEG for brainprint recognition research.

2.7.5 Difference Topomap Analysis

Channel-wise paired t-tests were performed to identify task-related spatial differences:

$$t_c = ttest(DE_c^{read}, DE_c^{memory}). \tag{5}$$

As shown in Fig. 2(g)–(i), significant spatial reorganization was observed between Reading and Memory stages. Specifically, the Hearing group exhibited broader Memory-related activations than the Deaf group, suggesting that memory retrieval recruits more extensive neural resources in hearing participants. Furthermore, distinct topographic patterns were observed between the two populations, indicating different neural representation strategies during cognitive processing. These findings demonstrate the effectiveness of MCD-EEG for both cognitive-state modeling and neurodiversity analysis.

3 Benchmark Protocol

We evaluate four benchmark tasks: (1) population classification between hearing and prelingually deaf participants using EEG functional-connectivity representations; (2) within-task brainprint recognition in the Reading and Memory Retrieval stages; (3) cross-task brainprint recognition by training on Reading and testing on Memory Retrieval; and (4) cross-Re-cap brainprint recognition by training on Memory Retrieval and testing on recordings acquired after EEG cap removal and reattachment.

Within-task experiments use five-fold cross-validation, whereas the cross-condition benchmarks evaluate generalization to a different cognitive task or recording session. For the brainprint-recognition tasks, results are reported separately for the hearing and prelingually deaf groups to support subgroup robustness analysis. The full task definitions and guardrails are specified in Appendix B and Appendix D.

We compare eight representative approaches: LRMD, FBCNet, EEGNet, EEGViT, EEGConformer, EEGPT, BrainprintNet, and ATCNet (Kong et al., 2018; Mane et al., 2020; Lawhern et al., 2018;

Table 2: Reading Stage performance under five-fold cross-validation. Results are reported as mean ± standard deviation over five folds. The inconsistency between Acc and Recall is due to rounding in the *train_test_split* division.

Methods	Hearing				Deaf			
	Acc	EER	Recall	F1	Acc	EER	Recall	F1
LRMD	0.4955±0.01	0.1730±0.01	0.4961±0.01	0.4892±0.01	0.5271±0.01	0.1833±0.01	0.5284±0.01	0.5193±0.01
FBCNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGViT	0.8167±0.02	0.0532±0.01	0.8172±0.02	0.8127±0.02	0.9078±0.02	0.0326±0.01	0.9075±0.02	0.9070±0.02
EEGConformer	0.9457±0.01	0.0183±0.00	0.9465±0.01	0.9444±0.01	0.9976±0.00	0.0024±0.00	0.9976±0.00	0.9976±0.00
EEGPT	0.9480±0.01	0.0113±0.00	0.9482±0.01	0.9466±0.01	0.9527±0.01	0.0165±0.00	0.9532±0.01	0.9518±0.01
BrainprintNet	0.9593±0.01	0.0204±0.00	0.9592±0.01	0.9597±0.01	0.9683±0.01	0.0158±0.00	0.9689±0.01	0.9691±0.01
ATCNet	0.7330±0.02	0.1086±0.01	0.7339±0.02	0.7305±0.02	0.7683±0.01	0.0757±0.01	0.7671±0.01	0.7666±0.01

Table 3: Memory Stage performance under five-fold cross-validation. Results are reported as mean ± standard deviation over five folds. The inconsistency between Acc and Recall is due to rounding in the *train_test_split* division.

Methods	Hearing				Deaf			
	Acc	EER	Recall	F1	Acc	EER	Recall	F1
LRMD	0.5113±0.01	0.1855±0.01	0.5125±0.01	0.4989±0.01	0.4819±0.01	0.1954±0.01	0.4828±0.01	0.4733±0.01
FBCNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGViT	0.8077±0.02	0.0566±0.01	0.8082±0.02	0.8031±0.02	0.8582±0.02	0.0492±0.01	0.8577±0.02	0.8579±0.02
EEGConformer	0.9525±0.01	0.0220±0.00	0.9532±0.01	0.9513±0.01	0.9929±0.00	0.0024±0.00	0.9928±0.00	0.9928±0.00
EEGPT	0.9389±0.01	0.0126±0.00	0.9396±0.01	0.9383±0.01	0.9504±0.01	0.0095±0.00	0.9506±0.01	0.9501±0.01
BrainprintNet	0.9545±0.01	0.0201±0.00	0.9564±0.01	0.9528±0.01	0.9632±0.01	0.0147±0.00	0.9670±0.01	0.9611±0.01
ATCNet	0.6493±0.02	0.1244±0.01	0.6495±0.02	0.6400±0.02	0.8463±0.01	0.0617±0.01	0.8461±0.01	0.8416±0.01

Altaheri et al., 2023; Song et al., 2022; Arjun et al., 2021; Wang et al., 2024; Tu et al., 2026). Accuracy, equal error rate (EER), recall, and F1 are reported separately for the two participant groups. Full mean and standard deviation results appear in Tables 2–5.

4 RESULTS

This section systematically evaluates and decouples the performance of current mainstream Electroencephalogram (EEG) deep learning architectures and traditional machine learning methods on the MCD-EEG benchmark suite. All experiments strictly follow a training and testing set separation protocol and adopt a unified five-fold cross-validation for technical verification.

The representative backbones participating in the evaluation include:

- **LRMD (Kong et al., 2018):** A method combining traditional time-frequency statistical features with linear regression, representing the classic shallow machine learning paradigm.
- **FBCNet (Mane et al., 2020):** A deep network based on filter bank convolutional structures, which enhances EEG representation capabilities through multi-band feature learning.
- **EEGNet (Lawhern et al., 2018):** A typical spatio-temporal Convolutional Neural Network (CNN) capable of jointly learning the temporal and spatial features of EEG signals.
- **ATCNet (Altaheri et al., 2023):** A deep network integrating attention mechanisms with temporal convolutional structures, designed to capture long-range temporal dependencies.
- **EEGConformer (Song et al., 2022):** A spatio-temporal joint framework combining CNN-based spatial feature extraction and Transformer-based temporal context modeling.
- **EEGViT (Arjun et al., 2021):** An EEG representation learning model based on the Vision Transformer (ViT), utilizing a global self-attention mechanism to model cross-channel dependencies.

Table 4: Read-Memory Stage performance under five-fold cross-validation. Results are reported as mean ± standard deviation over five folds. Acc = Recall because the number of samples in each class is equal.

Methods	Hearing				Deaf			
	Acc	EER	Recall	F1	Acc	EER	Recall	F1
LRMD	0.3859±0.01	0.2349±0.01	0.3859±0.01	0.3712±0.01	0.4624±0.01	0.1882±0.01	0.4624±0.01	0.4634±0.01
FBCNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGNet	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00
EEGViT	0.8410±0.02	0.0462±0.01	0.8410±0.02	0.8392±0.02	0.8523±0.02	0.0419±0.01	0.8523±0.02	0.8488±0.02
EEGConformer	1.0000±0.00	0.0000±0.00	1.0000±0.00	1.0000±0.00	0.9986±0.00	0.0007±0.00	0.9986±0.00	0.9986±0.00
EEGPT	0.9524±0.01	0.0149±0.00	0.9524±0.01	0.9522±0.01	0.9773±0.01	0.0065±0.00	0.9773±0.01	0.9770±0.01
BrainprintNet	0.9631±0.01	0.0192±0.00	0.9692±0.01	0.9688±0.01	0.9707±0.01	0.0123±0.00	0.9716±0.01	0.9756±0.01
ATCNet	0.7255±0.02	0.0968±0.01	0.7255±0.02	0.7239±0.02	0.7386±0.01	0.0881±0.01	0.7386±0.01	0.7361±0.01

Table 5: Re-cap Stage performance under five-fold cross-validation. Results are reported as mean ± standard deviation over five folds. Acc = Recall because the number of samples in each class is equal.

Methods	Hearing				Deaf			
	Acc	EER	Recall	F1	Acc	EER	Recall	F1
LRMD	0.0822±0.01	0.4574±0.01	0.0822±0.01	0.0807±0.01	0.0959±0.01	0.4418±0.01	0.0959±0.01	0.0965±0.01
FBCNet	0.4823±0.02	0.2088±0.01	0.4823±0.02	0.3986±0.02	0.8239±0.01	0.0746±0.01	0.8239±0.01	0.8032±0.01
EEGNet	0.5190±0.02	0.2894±0.01	0.5190±0.02	0.4163±0.02	0.7358±0.02	0.0909±0.01	0.7358±0.02	0.6913±0.02
EEGViT	0.4083±0.02	0.2656±0.01	0.4083±0.02	0.3844±0.02	0.5895±0.02	0.1413±0.01	0.5895±0.02	0.5785±0.02
EEGConformer	0.3899±0.02	0.2738±0.01	0.3899±0.02	0.3267±0.02	0.5760±0.02	0.1903±0.01	0.5760±0.02	0.5248±0.02
EEGPT	0.3533±0.02	0.2947±0.01	0.3533±0.02	0.3163±0.02	0.4510±0.02	0.2091±0.01	0.4510±0.02	0.4398±0.02
BrainprintNet	0.3743±0.02	0.2303±0.01	0.3743±0.02	0.3598±0.02	0.4162±0.02	0.2152±0.01	0.4162±0.02	0.3836±0.02
ATCNet	0.1624±0.01	0.3879±0.01	0.1624±0.01	0.1604±0.01	0.2159±0.02	0.3098±0.01	0.2159±0.02	0.2007±0.02

- **EEGPT (Wang et al., 2024):** A Transformer architecture built entirely on self-attention mechanisms and a generative pre-training strategy, aimed at learning universal EEG representations.
- **BrainprintNet (Tu et al., 2026):** A deep brainprint network specifically designed for individual identity recognition tasks, optimized for extracting individual discriminative features.

4.1 Multi-Metric Comparative Analysis for Brainprint Recognition

Tables 2, 3, and 4 systematically present the identity recognition performance of various mainstream backbones under single-task (Read and Mem stages) and cross-task (Read → Mem stage) conditions.

(1) Single-Task Brainprint Recognition Performance Analysis (Reading and Memory Stages)

As shown in Table 2 and Table 3, under the single-task setting, all models exhibit consistent performance trends across the reading and memory stages. The traditional LRMD method achieves relatively low performance (Acc $\approx$ 0.48–0.53) with higher EER, indicating limited identity discrimination capability. Deep learning-based approaches substantially improve recognition accuracy, where FBCNet and EEGNet achieve near-perfect performance (Acc $\approx$ 1.00) under controlled single-task conditions. Transformer-based and contrastive learning models further enhance robustness, with EEGViT achieving Acc of approximately 0.81–0.90, while EEGConformer, EEGPT, and BrainprintNet consistently exceed 0.94 Acc with lower EER. Among them, BrainprintNet achieves competitive performance (Acc $\approx$ 0.95–0.97), demonstrating strong subject-specific representation learning ability, whereas ATCNet shows relatively weaker discrimination capability.

(2) CROSS-TASK BRAINPRINT RECOGNITION PERFORMANCE ANALYSIS (READ → MEM)

As shown in Table 4, under the cross-task setting, all models exhibit varying degrees of performance degradation, reflecting the significant challenges posed by task distribution shifts. The traditional LRMD method suffers the most obvious decline, with its Acc dropping to 0.38–0.46, demonstrating high sensitivity to task variations. Among the deep learning methods, EEGViT remains relatively stable (Acc $\approx$ 0.84), though it degrades compared to the single-task scenario.

Notably, EEGConformer approaches 1.0 in some settings, demonstrating strong cross-task generalization. EEGPT and BrainprintNet maintain high accuracy (Acc $\approx$ 0.95–0.97) with low EER, indicating strong robustness and transferability. ATCNet shows larger fluctuations but remains relatively stable. Overall, pre-training and architectural enhancements effectively alleviate cross-task distribution shifts.

4.2 PERFORMANCE CLIFF AND DECOUPLING ANALYSIS IN CROSS-TIME EQUIPMENT RECONSTRUCTION (RE-CAP)

In Task 4, which introduces hardware-level electrode re-wearing (Re-cap), all models experience a "cliff-like" performance decline due to spatial topological shifts. As shown in Table 5, the traditional LRMD method fails completely (Acc dropping to 0.08–0.09). Deep learning architectures also degrade significantly: CNNs like FBCNet and EEGNet drop to 0.48–0.82, while Transformers and contrastive models (EEGViT, EEGConformer, EEGPT) fall to the 0.35–0.58 range. ATCNet drops near random levels (0.16–0.21).

Compared with cross-task cognitive shifts (Table 4), where models achieve near 100% accuracy, Re-cap induced physical sensor drift causes severe degradation, reducing CNN accuracy to $\sim$60% and alignment-based models to 30%–40%. These results indicate that sensor-level shifts, rather than cognitive variation, are the primary barrier to real-world deployment, highlighting the need for domain adaptation and geometrically invariant representations.

4.3 SUBGROUP ROBUSTNESS EVALUATION AND FAIRNESS AUDITING

To address potential "neurotypical bias," we audit model fairness between hearing ($g = 0$) and prelingually deaf ($g = 1$) subgroups using the Subgroup Accuracy Gap (ΔAcc) and Equal Error Rate Gap (ΔEER).

As shown in Appendix Figure 3 and Tables 2–5, the deaf group unexpectedly outperforms the hearing group under intra-session and cross-task settings. For example, EEGViT achieves a +9.11% ΔAcc improvement, while EEGPT maintains a low ΔEER of 0.0084, suggesting that early auditory deprivation may induce cortical reorganization and more discriminable neural topologies under stable sensor conditions. However, this advantage diminishes under the Task 4 Re-cap stress test. CNN-based models exhibit stronger physical robustness (e.g., FBCNet achieves 0.8239 Acc in the deaf group), whereas Transformers and linear models degrade substantially. These results expose architecture-dependent biases under physical non-stationarity and emphasize the necessity of geometry-invariant optimization and fairness-aware evaluation for neurodiverse EEG recognition.

4.4 POPULATION CLASSIFICATION

In the model construction using population groups as classification labels, good separability was demonstrated in both the reading and memory stages. The performance of each model is shown in Appendix Table 9 . The XGBoost model achieved the best performance in the reading stage, while the SVM model performed best in the memory stage.

5 CONCLUSION

MCD-EEG combines cognitive shifts, EEG cap reattachment, and two participant groups in one brainprint benchmark. Results show that strong cross-task identity recognition does not imply robustness to re-cap drift. The benchmark therefore supports work on representations that retain identity under sensor changes and can be evaluated separately across participant groups.

AI USE STATEMENT

Generative AI tools were used to assist with English translation and language editing, mathematical typesetting, and manuscript organization. They were not used to formulate research hypotheses or mathematical claims, develop the proposed methodology, design or conduct experiments, generate or process experimental data, implement the proposed method, interpret experimental results, or determine the scientific conclusions of this work. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

H. Altaheri, G. Muhammad, and M. Alsulaiman. Physics-informed attention temporal convolutional network for eeg-based motor imagery classification. *IEEE Trans. Ind. Informat.*, 19(2):2249–2258, Feb. 2023.

A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker. Introducing attention mechanism for eeg signals: Emotion recognition with vision transformers. In *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, pp. 5723–5726, Nov. 2021.

T. Barbera, J. Burger, A. D'Amelio, et al. On using ai for eeg-based bci applications: problems, current challenges and future trends. *International Journal of Human–Computer Interaction*, pp. 1–20, 2025.

R. N. Duan, J. Y. Zhu, and B. L. Lu. Differential entropy feature for eeg-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, 2013.

D. Erdogmus, A. Adami, M. Pavel, et al. Cognitive state estimation based on eeg for augmented cognition. In *Conference Proceedings. 2nd International IEEE EMBS Conference on Neural Engineering, 2005.*, pp. 566–569. IEEE, 2005.

A. Fayyazi, S. Safari, and S. Farashi. Brain connectome differences between attention deficit hyperactivity disorder (adhd) and neurotypical children during visual attention: A study using a minimum spanning tree graph, multichannel eeg recording and machine learning. *Applied Neuropsychology: Child*, pp. 1–10, 2025.

Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, 16(2):969–985, 2025. doi: 10.1109/TAFFC.2024.3485057.

X. Jiang, S. Shang, A. Wilkinson, et al. Interpretable fuzzy modeling reveals population-level representation differences in p300 brain computer interfaces across neurodivergent and neurotypical cohorts. *arXiv preprint arXiv:2604.24765*, 2026.

M. L. Jouandet, M. J. Tramo, D. M. Herron, et al. Brainprints: computer-generated two-dimensional maps of the human cerebral cortex in vivo. *Journal of Cognitive Neuroscience*, 1(1):88–117, 1989.

E. A. Kaplan-Kahn, A. Park, and N. Russo. Pathways of perceptual primacy: Erp evidence for relationships between autism traits and enhanced perceptual functioning. *Neuropsychologia*, 163: 108065, 2021.

S. Katsigiannis and N. Ramzan. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2017.

A. Khorasani and T. Demazure. Cross-temporal attention fusion (ctaf) for multimodal physiological signals in self-supervised learning. *arXiv preprint arXiv:2602.02784*, 2026.

S. Koelstra, C. Muhl, M. Soleymani, et al. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2011.

X. Kong, W. Kong, Q. Fan, Q. Zhao, and A. Cichocki. Taskindependent eeg identification via low-rank matrix decomposition. In *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, pp. 412–419, Dec. 2018.

V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance. Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *J. Neural Eng.*, 15(5):056013, Oct. 2018.

Wei Liu, Wei-Long Zheng, Ziyi Li, Si-Yuan Wu, Lu Gan, and Bao-Liang Lu. Identifying similarities and differences in emotion recognition with eeg and eye movements among chinese, german, and french people. *Journal of Neural Engineering*, 19(2):026012, 2022.

Z. Lu. Pyctrsa: A python package for cross-temporal representational similarity analysis-based e/meg decoding, 2020.

R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan. A multiview cnn with novel variance layer for motor imagery brain computer interface. In *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, pp. 2950–2953, Jul. 2020.

V. Mihajlović, B. Grundlehner, R. Vullers, et al. Wearable, wireless eeg solutions in daily life applications: what are we missing? *IEEE Journal of Biomedical and Health Informatics*, 19(1): 6–21, 2014.

J. A. Miranda-Correa, M. K. Abadi, N. Sebe, et al. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2): 479–493, 2018.

W. S. Neo, D. Foti, B. Keehn, et al. Resting-state eeg power differences in autism spectrum disorder: a systematic review and meta-analysis. *Translational Psychiatry*, 13(1):389, 2023.

V. P. Oikonomou, G. Liaros, K. Georgiadis, et al. Comparative evaluation of state-of-the-art algorithms for ssvep-based bcis. *arXiv preprint arXiv:1602.00904*, 2016.

A. Schaefer, F. Nils, X. Sanchez, and P. Philippot. Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. *Cognition and Emotion*, 24(7): 1153–1172, 2010.

G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. Bci2000: A general-purpose brain-computer interface (bci) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Y. Song, Q. Zheng, B. Liu, and X. Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 31:710–719, 2022.

M. Tangermann, K. R. Müller, A. Aertsen, et al. Review of the bci competition iv. *Frontiers in Neuroscience*, 6:55, 2012.

Y. Tu, S. Li, X. Chen, and D. Wu. Brainprintnet: A multiscale cross-band fusion network for eeg-based brainprint recognition. *IEEE Transactions on Information Forensics and Security*, 21: 2757–2768, 2026. doi: 10.1109/TIFS.2026.3672000.

C. Wachinger, P. Golland, W. Kremen, et al. Brainprint: A discriminative characterization of brain morphology. *NeuroImage*, 109:232–248, 2015.

G. Wang, W. Liu, Y. He, C. Xu, L. Ma, and H. Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 39249–39280, Dec. 2024.

M. Wang, J. Hu, and H. A. Abbass. Brainprint: Eeg biometric identification based on analyzing brain connectivity graphs. *Pattern Recognition*, 105:107381, 2020.

P. Zhang, X. Wang, W. Zhang, et al. Learning spatial–spectral–temporal eeg features with recurrent 3d convolutional neural networks for cross-task mental workload assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(1):31–42, 2018.

W. Zheng. Multichannel eeg-based emotion recognition via group sparse canonical correlation analysis. *IEEE Transactions on Cognitive and Developmental Systems*, 9(3):281–290, 2016.

Y. Zhou, Z. Xu, Y. Niu, et al. Cross-task cognitive workload recognition based on eeg and domain adaptation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:50–60, 2022.

A Additional Subgroup Fairness Analysis

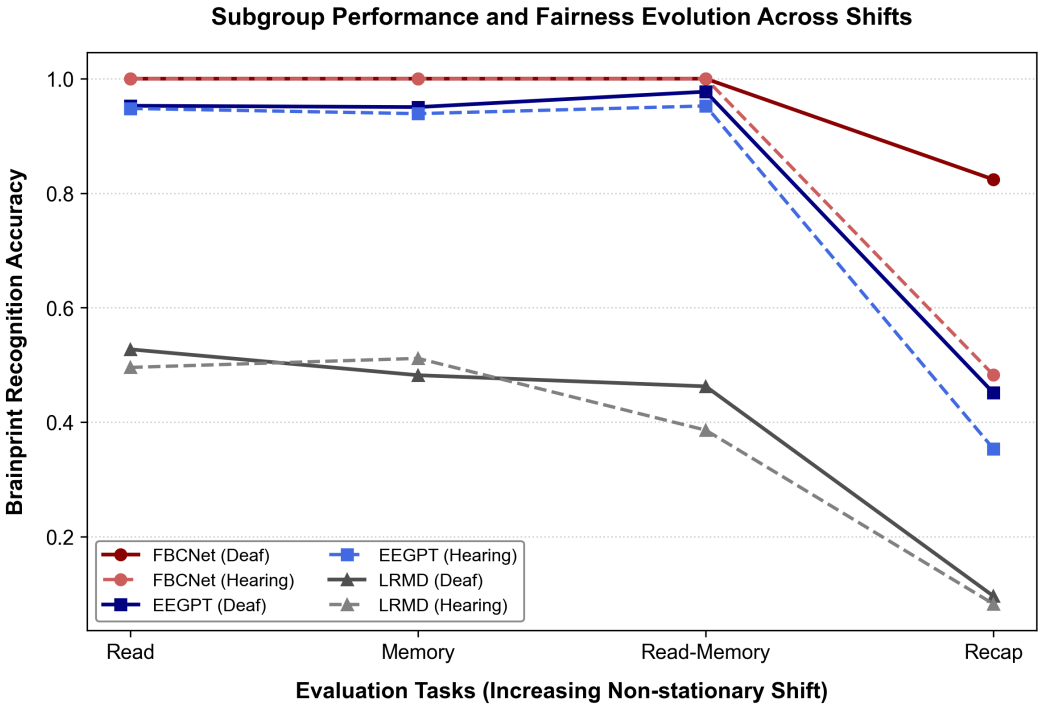


Figure 3: Evolution analysis of subgroup performance and fairness during distribution shifts.

B Detailed Dataset Construction and Analysis

B.1 Participants and Dataset Composition

The proposed **MCD-EEG** dataset comprises high-density EEG recordings from 45 participants, including 22 prelingually deaf individuals and 23 age-matched healthy controls. All participants were right-handed, had normal or corrected-to-normal vision, and reported no history of neurological disorders. EEG signals were recorded using a 60-channel high-density acquisition system to ensure sufficient spatial resolution and to comprehensively capture individual-specific neural characteristics. Prior to the experiment, demographic information and standardized cognitive assessments were collected to ensure baseline comparability between the two groups.The construction and benchmark evaluation framework of the MCD-EEG dataset is illustrated in Fig. 1.

The participants were finally divided into two subpopulations:

$$g = \begin{cases} 0, & \text{Hearing group,} \\ 1, & \text{Deaf group.} \end{cases} \tag{6}$$

B.2 Non-Stationary Cognitive Paradigm

To capture realistic cognitive and sensor-induced distribution shifts, we designed a four-stage experimental paradigm with 64 trials per participant, including Reading, Interference, Memory Retrieval, and Re-cap stages.During the Reading stage, participants judged sentence correctness and the response time (RT_1) was recorded. The Interference stage introduced transient cognitive variations through an arithmetic task. In the Memory Retrieval stage, participants recovered missing sentence information, with retrieval response time (RT_2) and accuracy (RA_2) recorded. Stimuli followed a 2×2 factorial design, including four conditions: *CN*, *IN*, *CT*, and *IT*. Each trial contained a 500-ms

fixation period and a 2000-ms inter-trial interval to ensure temporal alignment. To further evaluate brainprint robustness against sensor-related distribution shifts, an additional *Re-cap* stage was conducted. After completing the first three stages, participants removed and then re-wore the EEG cap before performing a standardized IQ assessment task. EEG signals were continuously recorded throughout the assessment, and 2000-ms EEG segments were extracted for analysis. This stage was specifically designed to simulate realistic electrode repositioning and physical sensor displacement, enabling the evaluation of brainprint stability under hardware-induced distribution shifts.

B.3 Preprocessing and Tensor Representation

Raw EEG signals underwent standard preprocessing (interpolation, downsampling to 400 Hz, 0.1–45 Hz band-pass filtering, ICA, and re-referencing). The benchmark dataset is formulated as $\mathcal{D} = \{(\mathbf{X}_i, y_i, g_i)\}_{i=1}^{M}$, where $M = 45$ subjects, y_i is the identity label, and $g_i \in \{0, 1\}$ indicates the neurodiversity subgroup.

To simulate real-world wearable deployment, participants underwent a cross-session re-setup protocol involving complete EEG cap removal, reattachment (*re-cap*), and recalibration. Consequently, each participant provides three structured EEG tensors $\mathbf{X}^{\text{read}}, \mathbf{X}^{\text{mem}}, \mathbf{X}^{\text{recap}} \in \mathbb{R}^{N \times C \times T}$ (N trials, $C = 60$ channels, $T = 1600$ time points). $\mathbf{X}^{\text{read}}$ serves as the baseline for representation learning, while $\mathbf{X}^{\text{mem}}$ and $\mathbf{X}^{\text{recap}}$ evaluate cross-task (cognitive shift) and cross-session (electrode repositioning) robustness, respectively.

B.4 Problem Formulation and Evaluation Guardrails

Given multi-condition observations, the objective is to learn a representation function $f_\theta : \mathbf{X}_i^{(*)} \to \mathbf{z}_i$ that satisfies three core properties: 1) **Identity consistency** across cognitive states ($\mathbf{z}_i \approx \mathbf{z}_i^{\text{read}} \approx \mathbf{z}_i^{\text{mem}} \approx \mathbf{z}_i^{\text{recap}}$); 2) **Cross-session robustness** against physical shifts ($\mathbf{z}_i^{(1)} \approx \mathbf{z}_i^{(2)}$); 3) **Subgroup invariance** for fairness ($P(\hat{y} \mid g = 0) \approx P(\hat{y} \mid g = 1)$).

Thus, MCD-EEG explicitly formulates a *multi-shift representation learning benchmark*. Furthermore, to ensure realistic Edge AI evaluation, strict train-test separation is enforced, absolutely prohibiting test-set leakage during feature extraction, or model optimization.

C Benchmark Reliability and Dataset Validation

C.1 Behavioral and Neurophysiological Validation

Mixed-effects analyses confirm the paradigm successfully induces structured cognitive perturbations. Tongue-twister sentences significantly increased reading latency ($p < 0.001$), while semantic incongruence impaired memory retrieval latency ($p < 0.001$) and accuracy ($p = 0.021$), with distinct group-by-condition interactions. Furthermore, Event-Related Potential (ERP) analyses validated signal quality. While the hearing group exhibited canonical N400 responses (300–700 ms) and the deaf group showed distinct fronto-central positivities (early ∼80 ms and late 500–1500 ms), ERP latencies remained highly stable across subjects and sessions. Together, these results demonstrate that MCD-EEG provides high-quality, temporally reliable recordings ideal for evaluating cross-task EEG representation learning. The detailed results are presented in Appendix Table 8.

D Full Benchmark Task Definitions

Task 1: Neurodiversity-Aware Macro-Network Classification

This task evaluates the model's capacity to capture cortical reorganization from auditory deprivation. Using connectivity matrices

$$\mathbf{A} \in \mathbb{R}^{C \times C}$$

(where C is the channel count), the model learns the mapping

$$f : \mathbf{A} \to y, \quad y \in \{0, 1\}$$

to classify hearing ($y = 0$) and prelingually deaf ($y = 1$) populations. The detailed results are presented in Appendix Table 9.

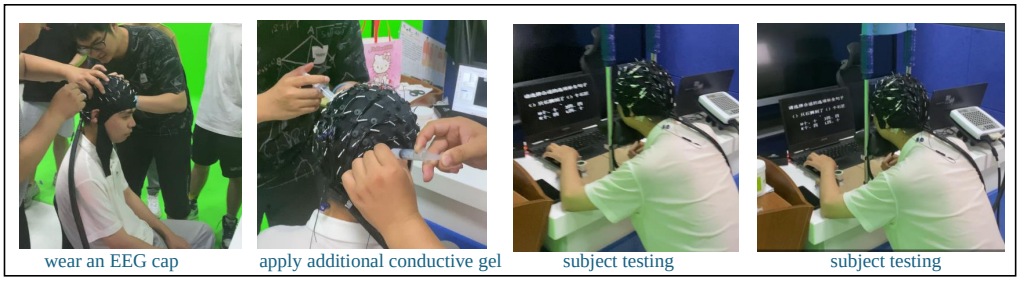


Figure 4: **Real-world EEG data acquisition environment.** (a) Wearing an EEG cap. (b) Applying additional conductive gel to ensure signal quality. (c–d) Participants performing a linguistic text similarity judgment task during EEG recording.

TASK 2: WITHIN-TASK BRAINPRINT RECOGNITION

This baseline establishes the upper limit of identity extraction under stationary cognitive states using a 5-fold cross-validation on 45 subjects. It comprises two sub-tasks:

- **Read-to-Read Identification:** Training and testing within the Reading Stage of Session 1.
- **Mem-to-Mem Identification:** Training and testing within the Memory Retrieval Stage of Session 1 to evaluate performance under a single, fixed cognitive load.

TASK 3: CROSS-TASK NON-STATIONARY BRAINPRINT RECOGNITION

This zero-shot task evaluates identity robustness against severe cognitive distribution shifts. Models are trained on Reading stage EEG tensors and tested on Memory Retrieval tensors:

$$\text{Train: } f(\mathbf{X}_{\text{read}}) \to y_{\text{id}}, \quad \text{Test: } \mathbf{X}_{\text{mem}} \mapsto \hat{y}_{\text{id}}$$

TASK 4: CROSS-TIME RE-CAP BRAINPRINT RECOGNITION

The core benchmark stress-tests robustness against physical sensor non-stationarity (e.g., electrode displacement, impedance changes) during cross-session deployments. Models train on Session 1 Memory data and verify identity on Session 2 Re-cap data (collected after re-wearing the EEG cap):

$$\text{Train: } f(\mathbf{X}_{\text{mem}}) \to y_{\text{id}}, \quad \text{Test: } \mathbf{X}_{\text{re-cap}} \mapsto \hat{y}_{\text{id}}$$

This systematically evaluates potential performance cliffs caused by real-world physical and environmental distribution shifts.

E SUPPLEMENTARY EXPERIMENTAL DETAILS

E.1 SUBJECTS

This experiment recruited two groups of university student participants, including 23 deaf individuals and 25 hearing individuals. Due to incomplete experiments and poor EEG data quality, 1 deaf participant and 2 hearing participants were excluded. The final sample consisted of 22 deaf participants in the experimental group (12 males, 10 females) and 23 hearing participants in the control group (11 males, 12 females). All deaf participants had profound hearing loss (bilateral hearing loss > 85 dB) and were prelingually deaf (onset of deafness before age 3 or congenitally deaf). Among them, 6 had cochlear implants, 14 wore hearing aids, and 2 did not use any hearing assistive devices.

Before the experiment, the Language Experience and Proficiency Questionnaire (LEAP-Q) developed by Marian et al. was used to measure the spoken and sign language experiences of the deaf university students. Participants were required to rate the percentage of time they used each language, and all deaf participants reported a sign language usage rate of $\geq 50\%$. The native language

Table 6: Demographic Information of Participants: M (SD)

Demographic Info	**Deaf Group** ($N = 22$)	**Hearing Group** ($N = 23$)	**ANOVA Results**
Age	21.64 (1.89)	22.17 (1.69)	$F(1, 43) = 1.01$, $p = .321$
Reading Comprehension	13.45 (0.74)	13.73 (0.75)	$F(1, 43) = 1.64$, $p = .207$
Reading Fluency	387.12 (120.18)	433.16 (99.66)	$F(1, 43) = 1.96$, $p = .168$
IQ	63.91 (4.10)	65.52 (2.66)	$F(1, 43) = 2.47$, $p = .124$

of all hearing participants was Mandarin. All participants had normal or corrected-to-normal vision, were right-handed, and had no history of psychiatric or neurological disorders.

Before participating in the experiment, both groups completed the following three tests:

1. **Reading Comprehension Test.** Referring to the experiment by Chen et al., this study used a cloze test to evaluate reading comprehension, utilizing a previous HSK-5 examination cloze section. The test consisted of 4 short passages, with 3 or 4 blanks in each passage. For each blank, participants needed to select the most appropriate option from four choices. One point was awarded for each correct answer, with a maximum score of 15.
2. **Reading Fluency Test.** A three-minute rapid reading test was employed, comprising 100 sentences with gradually increasing lengths. Participants were required to silently read the sentences as quickly as possible within three minutes and judge the logical correctness of each sentence, such as "Swallows can fly (✓)" or "There are eight days in a week (×)". The final reading fluency score was calculated as the total number of correctly judged characters minus the total number of incorrectly judged characters, divided by the time.
3. **Raven's Non-verbal Intelligence Test.** Using Raven's non-verbal intelligence test and referencing previous studies, the Combined version of Raven's Test was selected, utilizing units A, B, and AB from the Colored version alongside units C, D, and E from the Standard version, totaling 72 items. One point was awarded for each correct answer, with a maximum score of 72. As all participants were in the same age group, scores were directly used to represent IQ levels without further IQ conversion.

One-way Analysis of Variance (ANOVA) revealed no significant differences between the two groups across any of the tests, indicating that the participating deaf individuals were fundamentally matched in reading proficiency with the hearing individuals. The means and standard deviations of the demographic variables are presented in Table 6.

This study obtained ethical approval from the Ethics Committee of Tianjin University. All participants carefully read and signed an informed consent form before the experiment and received monetary compensation upon completion.

E.2 Materials and Experiments

The experimental materials consisted of 32 sets of short sentences. Each set contained 4 sentences: a Correct Normal sentence (CN), an Incorrect Normal sentence (IN), a Correct Tongue twister (CT), and an Incorrect Tongue twister (IT), totaling 128 short sentences. Sentences within the same set were identical in character count and syntactic structure. The correct tongue twisters were selected from the *Dictionary of Tongue Twisters*, and the corresponding three other variations were generated by modifying phonetically similar words and swapping the positions of subjects and objects or double objects. Examples are shown in Table 7.

Before the experiment, 17 healthy university students were invited to rate the logicality of the experimental materials using a 5-point Likert scale, where 1 indicated "completely logical" and 5 indicated "completely illogical". All logically correct sentences scored below 2 ($M = 1.21$, $SD = 0.20$), while all logically incorrect sentences scored above 4 ($M = 4.76$, $SD = 0.15$). A one-way ANOVA showed a significant difference in logicality scores between the two ($F(1, 126) = 12965.75$, $p < .001$). There was no significant difference in logicality scores between tongue twisters and normal sentences ($F(1, 126) = 0$, $p = .988$). The short sentences were divided into two sets, en-

Table 7: Examples of Experimental Materials

Type	Sentence (English)	Selected Words	Distractor Words
CN	Merchants buy and sell goods	merchants, goods	merchants, goods
IN	Goods buy and sell merchants	goods, merchants	goods, merchants
CT	Herdsmen buy and sell horses	herdsmen, horses	shepherd boy, trojan horse
IT	Horses buy and sell herdsmen	horses, herdsmen	trojan horse, shepherd boy

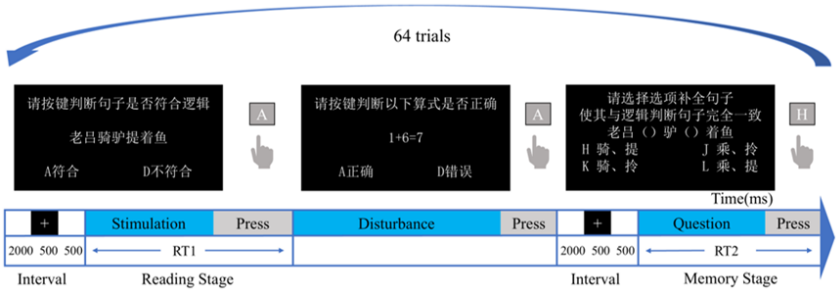


Figure 5: Flowchart of the experimental paradigm.

suring that a correct sentence and its corresponding incorrect version did not appear in the same set. Participants were randomly assigned one of the two sets for the experiment.

The experimental procedure was divided into two stages: reading judgment and cloze test. Participants first pressed a key to judge whether the sentence was logical. Following this, they evaluated an addition or subtraction equation within ten, a step designed to introduce memory interference. Finally, they completed a cloze test based on the sentence they had read earlier. To allow the EEG responses to relatively return to baseline before each stimulus presentation, a 2-second inter-trial interval was employed, followed by a 500 ms fixation cross warning stimulus, and a 500 ms blank screen to eliminate event-related potentials (ERPs) elicited by the cross. The flowchart of the experimental paradigm is shown in Figure 5.

Experimental stimuli were presented using E-Prime software on a monitor with a resolution of 1920×1080 and a refresh rate of 144 Hz. All text was displayed in Song typeface (SimSun) with a font size of 60, presented as white text on a black background, ensuring normal readability for all participants. Details of all experimental materials are provided in Supplementary Material 1.

E.3 EEG Signal Acquisition

EEG signals were acquired using the NeuroScan system coupled with SynAmps amplifiers and Curry 9 acquisition software. The electrode configuration adhered to the international 10-20 system standards. The left and right mastoid electrodes (M1, M2) were selected as reference electrodes, and 62 channels of EEG signals were recorded. The sampling rate was set to 1 kHz, and the signal bandwidth ranged from 0 to 400 Hz. A built-in 50 Hz notch filter was enabled during acquisition to eliminate power line interference. The experiment commenced only when the impedance of all electrodes fell below 15 kΩ.

E.4 Behavioral Data Analysis

The reaction accuracy and reaction time for the reading judgment and cloze test stages were recorded using the built-in keylogging function of E-Prime. For convenience of discussion, the reaction accuracy and reaction time for the reading judgment section are denoted as RA1 and RT1, while those for the cloze test section are denoted as RA2 and RT2.

A Generalized Linear Mixed Model (GLMM) was used to analyze reaction accuracy. For reaction time, a logarithmic transformation was applied first, followed by analysis using a Linear Mixed Model (LMM). Only trials with correct responses were included in the reaction time calculations. Simple effect analyses were conducted when significant interaction effects emerged. Specific analyses were performed in R 4.2.2 using the `lme4` and `lmerTest` packages, while simple effect analyses were implemented via the `emmeans` package. Group (deaf vs. hearing), sentence semantics (correct vs. incorrect), sentence type (tongue twister vs. normal), and their interactions were entered as fixed effects, whereas subjects, materials, and material presentation order were treated as random effects. The models were fitted using Restricted Maximum Likelihood (REML). The model fitting started from the most complex full model and was sequentially simplified; the `anova()` function was used to compare models until the best fit was achieved.

E.5 EEG Data Analysis

The EEG data first underwent preprocessing. Subsequently, feature extraction and further analyses were conducted from two perspectives: the time domain and functional connectivity.

E.5.1 EEG Data Preprocessing

Raw EEG data were processed using EEGLAB, an open-source toolbox for MATLAB (R2017a). Two electrodes with minimal cognitive relevance, CB1 and CB2, were removed. The data were down-sampled to 400 Hz and digitally filtered using a 0.1-45 Hz 3rd-order Butterworth band-pass filter. For instances where specific electrodes exhibited obvious abnormal signal waveforms during certain time segments, spherical interpolation in EEGLAB was utilized to replace these segments with the average values of surrounding electrodes. Artifacts, such as electrooculograms (EOG) and electromyograms (EMG), were removed via Independent Component Analysis (ICA).

E.5.2 Time-domain Analysis

The preprocessed EEG data were segmented based on event markers, extracting 1500 ms epochs post-stimulus onset, and baseline corrected using the 200 ms pre-stimulus interval. For each participant, an averaging method was applied to process the EEG signals of the two stages, categorized by the different sentence types. Consequently, each participant yielded 8 sets of time-domain signals across the 4 sentence types and 2 stages.

Using phonetic interference and semantic interference as independent variables, a two-way ANOVA was conducted on the data at each sampling point within the deaf group and the hearing group separately. To facilitate subsequent statistical analysis, we divided the data into 15 windows, denoted as W1 to W15, using a 100 ms time window to display the data. The selection of the 100 ms time window represents a balance between temporal resolution and the reduction of multiple comparisons, a practice widely adopted in similar time-series analyses.

E.5.3 Functional Connectivity Analysis

The preprocessed EEG data were epoched using event markers, extracting the segments from the onset of each stimulus to the participant's response. A functional connectivity matrix was calculated for every extracted trial of each participant. Regarding the functional connectivity metric, considering that the reading process is a dynamic process involving information transfer, the directed functional connectivity measure, Transfer Entropy (TE), was selected. TE was computed using the Java Information Dynamics Toolkit (JIDT) in MATLAB.

Compared to traditional undirected connectivity indices (e.g., Phase Locking Value (PLV), coherence), TE can reveal the direction of information flow between brain regions, making it more suitable for exploring causal interactive relationships among brain networks during cognitive processes.

For this analysis, 5632 functional connectivity matrices were obtained from 45 participants across 8 stimulus conditions and 16 trials. Following methods from previous research, the diagonal self-connections of the TE matrices were removed, and the matrices were flattened into one-dimensional feature vectors to serve as inputs for machine learning classification. First, classification was performed on the entire dataset using population groups as labels. Subsequently, classification was

conducted within each of the two participant groups using phonetic and semantic attributes as labels. Six machine learning models were employed: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), XGBoost, and LightGBM. A five-fold cross-validation scheme was adopted, and model performance was evaluated using AUC, accuracy, precision, sensitivity, specificity, and F1-score. The best-performing model was selected for SHapley Additive exPlanations (SHAP) analysis to interpret the differences in functional connectivity under various conditions.

E.6 Behavioral Validation of Cognitive Perturbation

To evaluate whether the proposed experimental paradigm induced measurable distribution shifts in behavioral responses, we conducted a mixed-effects model analysis under different experimental conditions. RT1/RT2 and RA1/RA2 denote the reaction times and accuracies for the reading and memory stages, respectively, which are used to quantify the subjects' processing speed and task execution accuracy under distinct cognitive task conditions.

As shown in Table 8, sentence type significantly affected the reaction time in the reading stage, exhibiting a significant main effect on $\log(\text{RT1})$ ($p < 0.001$). Specifically, compared to normal sentences, tongue twisters led to an increase in cognitive processing costs. The significant interaction effect between group and sentence type ($p = 0.001$) further indicates that different subject groups exhibited heterogeneous response patterns when confronting phonetic complexity.

In contrast, semantic inconsistency did not significantly affect the reaction time in the reading stage ($p = 0.866$), but it significantly impacted performance in the memory stage: it not only reduced accuracy (RA2, $p = 0.021$) but also increased the reaction time ($\log(\text{RT2})$, $p < 0.001$). This suggests that different task factors selectively affect distinct cognitive processing stages, thereby leading to multi-stage behavioral shifts.

Furthermore, the significant "Group $\times$ Sentence Semantics" interaction effect in RA2 performance ($p = 0.034$) indicates systematic differences between groups in semantic processing. No significant higher-order interaction effects were found, suggesting that phonetic and semantic factors primarily operate in a relatively independent manner.

Overall, these results confirm that the experimental design successfully induced structured and multi-stage behavioral variations, thereby providing a reliable controlled environment for evaluating the robustness of EEG-based recognition models under cross-task conditions.

E.7 Full Population Classification Results

Table 8: Results of the mixed-effects linear models.

Measure	Fixed Effect	b	SE	t/z	p
RA1	Group (Deaf vs. Hearing)	-1.434	0.277	-5.176	$< .001$
	Sentence Type (Tongue twist vs. Normal)	-0.359	0.309	-1.160	.246
	Sentence Semantics (Correct vs. Incorrect)	0.354	0.309	1.143	.253
	Group × Sentence Type	0.094	0.401	0.234	.815
	Group × Sentence Semantics	-0.278	0.402	-0.690	.490
	Sentence Type × Sentence Semantics	-0.251	0.619	-0.406	.685
	Group × Sentence Type × Sentence Semantics	1.567	0.918	1.707	.089
log(**RT1**)	Group (Deaf vs. Hearing)	0.297	0.102	2.903	.005
	Sentence Type (Tongue twist vs. Normal)	0.154	0.030	5.217	$< .001$
	Sentence Semantics (Correct vs. Incorrect)	0.005	0.030	0.169	.866
	Group × Sentence Type	-0.084	0.026	-3.218	.001
	Group × Sentence Semantics	0.052	0.026	1.993	.046
	Sentence Type × Sentence Semantics	-0.008	0.054	-0.153	.878
	Group × Sentence Type × Sentence Semantics	-0.034	0.053	-0.648	.517
RA2	Group (Deaf vs. Hearing)	-1.837	0.498	-3.686	$< .001$
	Sentence Type (Tongue twist vs. Normal)	0.666	0.197	3.380	$< .001$
	Sentence Semantics (Correct vs. Incorrect)	-0.613	0.265	-2.310	.021
	Group × Sentence Type	-0.369	0.285	-1.293	.196
	Group × Sentence Semantics	0.584	0.277	2.110	.034
	Sentence Type × Sentence Semantics	-0.236	0.397	-0.596	.551
	Group × Sentence Type × Sentence Semantics	-0.222	0.580	-0.382	.702
log(**RT2**)	Group (Deaf vs. Hearing)	0.285	0.068	4.161	$< .001$
	Sentence Type (Tongue twist vs. Normal)	-0.044	0.024	-1.861	.065
	Sentence Semantics (Correct vs. Incorrect)	0.143	0.024	6.037	$< .001$
	Group × Sentence Type	0.003	0.026	0.100	.921
	Group × Sentence Semantics	0.047	0.026	1.840	.066
	Sentence Type × Sentence Semantics	0.015	0.047	0.310	.757
	Group × Sentence Type × Sentence Semantics	-0.000	0.052	-0.001	.999

Table 9: Performance comparison of machine learning models

Stage	Model	AUC	Accuracy	Precision	Sensitivity	Specificity	F1
Read Stage	DT	0.632 ± 0.017	0.597 ± 0.019	0.558 ± 0.022	0.689 ± 0.055	0.517 ± 0.075	0.615 ± 0.015
	RF	0.755 ± 0.030	0.684 ± 0.021	0.711 ± 0.020	0.544 ± 0.044	0.806 ± 0.015	0.646 ± 0.034
	LR	0.769 ± 0.031	0.695 ± 0.027	0.669 ± 0.022	0.685 ± 0.055	0.704 ± 0.014	0.677 ± 0.037
	SVM	0.787 ± 0.023	0.709 ± 0.024	0.703 ± 0.022	0.655 ± 0.039	0.758 ± 0.015	0.678 ± 0.031
	XGB	0.803 ± 0.020	0.728 ± 0.021	0.721 ± 0.023	0.683 ± 0.036	0.767 ± 0.023	0.701 ± 0.026
	LGBM	0.795 ± 0.022	0.714 ± 0.019	0.711 ± 0.029	0.656 ± 0.018	0.764 ± 0.033	0.682 ± 0.017
Memory Stage	DT	0.625 ± 0.009	0.615 ± 0.012	0.583 ± 0.013	0.753 ± 0.043	0.483 ± 0.049	0.656 ± 0.014
	RF	0.745 ± 0.023	0.674 ± 0.030	0.667 ± 0.034	0.668 ± 0.043	0.679 ± 0.043	0.667 ± 0.031
	LR	0.772 ± 0.012	0.711 ± 0.010	0.691 ± 0.014	0.741 ± 0.034	0.682 ± 0.034	0.715 ± 0.014
	SVM	0.795 ± 0.017	0.726 ± 0.028	0.726 ± 0.036	0.712 ± 0.024	0.740 ± 0.024	0.718 ± 0.023
	XGB	0.787 ± 0.016	0.705 ± 0.019	0.686 ± 0.016	0.731 ± 0.026	0.680 ± 0.026	0.708 ± 0.020
	LGBM	0.787 ± 0.009	0.718 ± 0.007	0.701 ± 0.013	0.741 ± 0.030	0.696 ± 0.030	0.720 ± 0.011