SUBJECT-IDENTITY REMOVAL DOES NOT IMPROVE FROZEN EEG FOUNDATION MODEL TRANSFER: A NEGATIVE RESULT

FARS
Analemma
fars@analemma.ai

ABSTRACT

EEG foundation models enable cross-subject transfer learning for brain-computer interfaces, but inter-subject variability remains a challenge. We hypothesize that removing linearly decodable subject-identity information from frozen embeddings could improve transfer. We apply Iterative Nullspace Projection (INLP) to frozen CBraMod embeddings, combined with Euclidean Alignment preprocessing, and evaluate on BNCI2014001 motor imagery classification using leave-one-subject-out cross-validation. Our experiments refute this hypothesis: the Euclidean Alignment baseline achieves 56.27% accuracy, already exceeding reported fine-tuning (53.03%) by +3.24 percentage points, while optimized INLP achieves 56.29%, providing no meaningful improvement (+0.02 pp). Analysis reveals that inner cross-validation consistently selects minimal intervention configurations (1–3 iterations in 77.8% of folds), indicating that the optimization learns to avoid removing information. This negative result suggests that subject identity is entangled with task-discriminative signal in frozen EEG foundation model embeddings, and post-hoc linear debiasing is insufficient for improving cross-subject transfer.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Electroencephalography (EEG)-based brain-computer interfaces (BCIs) enable direct communication between the brain and external devices, with applications ranging from motor rehabilitation to assistive technology. A fundamental challenge in deploying BCIs is cross-subject generalization: models trained on one set of subjects often fail to transfer to new users due to substantial inter-subject variability in EEG signals arising from anatomical, physiological, and cognitive differences. This variability necessitates time-consuming calibration sessions for each new user, limiting the practical adoption of BCI systems.

EEG foundation models have emerged as a promising solution to this challenge. Models such as CBraMod (Wang et al., 2024), BENDR (Kostas et al., 2021), and LaBraM (Jiang et al., 2024) are pre-trained on large-scale EEG corpora and can be fine-tuned or used with linear probing for downstream tasks. However, even with these pre-trained representations, cross-subject transfer remains imperfect. Prior work has shown that subject identity is highly decodable from frozen EEG foundation model embeddings, suggesting that subject-specific confounds may interfere with task classification.

We hypothesize that removing linearly decodable subject-identity information from frozen EEG foundation model embeddings could improve cross-subject transfer. To test this hypothesis, we apply Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020), a method originally developed for debiasing word embeddings, to project embeddings onto the nullspace of subject-identity classifiers. Combined with Euclidean Alignment (EA) (He & Wu, 2018) preprocessing, this approach aims to produce subject-invariant representations while preserving task-relevant information.

[1] https://gitlab.com/fars-a/inlp-subject-nullspace-eeg-linear-probe

Our experiments on BNCI2014001 motor imagery classification refute this hypothesis. The EA baseline with frozen CBraMod embeddings achieves 56.27% accuracy, already exceeding reported fine-tuning (53.03%) by +3.24 pp. Optimized INLP achieves 56.29%, providing no meaningful improvement (+0.02 pp). Analysis reveals that inner cross-validation consistently selects minimal intervention configurations, suggesting that subject identity is entangled with task signal in frozen embeddings.

Our contributions are:

- We provide the first systematic evaluation of post-hoc subject-identity removal for frozen EEG foundation model transfer, testing whether INLP can improve cross-subject generalization.
- We conduct rigorous experiments with PCA-k controls and inner cross-validation optimization, establishing that all dimension removal methods degrade performance.
- We report a negative result: INLP provides no benefit over simple Euclidean Alignment, refuting the hypothesis that linear subject-identity removal improves transfer.
- We provide insight into the structure of frozen EEG-FM embeddings: subject identity appears entangled with task-discriminative signal, as evidenced by the optimization learning to minimize intervention.

2 Related Work

2.1 EEG Foundation Models

Recent advances in self-supervised learning have enabled the development of EEG foundation models that learn transferable representations from large-scale heterogeneous recordings. BENDR (Kostas et al., 2021) pioneered this direction by adapting contrastive self-supervised objectives from speech recognition to EEG, demonstrating that a single pre-trained model can generalize across different hardware, subjects, and tasks. LaBraM (Jiang et al., 2024) introduced vector-quantized neural spectrum prediction to train a semantically rich neural tokenizer, enabling cross-dataset learning by segmenting EEG signals into channel patches. BIOT (Yang et al., 2023) proposed a flexible biosignal transformer architecture that handles mismatched channels, variable lengths, and missing values by tokenizing each channel separately into unified sentence structures. CBraMod (Wang et al., 2024) devised a criss-cross transformer that models spatial and temporal dependencies separately through parallel attention mechanisms, achieving state-of-the-art performance across diverse BCI tasks. Recent benchmarking efforts (Liu et al., 2026) have systematically evaluated these models, revealing that linear probing is frequently insufficient and specialist models trained from scratch remain competitive across many tasks.

2.2 Domain Adaptation for EEG

Inter-subject variability poses a fundamental challenge for EEG-based BCIs, as neural signals exhibit substantial individual differences in amplitude, spatial patterns, and temporal dynamics. Euclidean Alignment (EA) (He & Wu, 2018) addresses this by aligning EEG trials from different subjects through a whitening transformation based on per-subject covariance matrices, enabling effective transfer learning with minimal computational cost. Recent work (Wu, 2025) has revisited EA across ten BCI paradigms, demonstrating its consistent effectiveness and efficiency. Riemannian geometry-based approaches (Tibermacine et al., 2024) offer an alternative by operating on the manifold of symmetric positive definite matrices, providing robustness to noise and non-stationarity. In the deep learning domain, CORAL (Sun & Saenko, 2016) aligns second-order statistics of source and target distributions through correlation alignment, while Domain-Adversarial Neural Networks (DANN) (Ganin et al., 2015) learn domain-invariant features through adversarial training with a gradient reversal layer. Domain Separation Networks (Bousmalis et al., 2016) explicitly model private and shared components between domains. Our work investigates whether post-hoc subject-identity removal can complement these alignment strategies for frozen EEG foundation model transfer.

2.3 NULLSPACE PROJECTION METHODS

Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) was introduced to remove protected attributes from neural representations by iteratively training linear classifiers and projecting embeddings onto their nullspace. This approach builds on earlier work in word embedding debiasing (Bolukbasi et al., 2016), which identified that gender bias is captured by a direction in the embedding space and proposed geometric methods to neutralize gender-stereotyped associations. However, subsequent analysis (Gonen & Goldberg, 2019) revealed that such linear debiasing methods may only superficially hide bias rather than truly remove it, as the protected information can often be recovered from the debiased embeddings. We apply INLP to a novel domain—frozen EEG foundation model embeddings—to test whether removing linearly decodable subject identity can improve cross-subject transfer. Our negative results suggest that, similar to the findings in NLP, subject-identity information may be entangled with task-relevant signal in ways that linear projection cannot cleanly separate.

3 METHOD

3.1 PROBLEM SETUP

We consider cross-subject EEG classification under a leave-one-subject-out (LOSO) protocol. Given EEG recordings from S subjects, we train on $S-1$ subjects and evaluate on the held-out subject, repeating for all S folds. Let $\mathcal{D} = \{(\mathbf{X}_i, y_i, s_i)\}_{i=1}^{N}$ denote the dataset where $\mathbf{X}_i \in \mathbb{R}^{C\times T}$ is a C-channel EEG trial of length T, $y_i \in \{1,\dots,K\}$ is the task label, and $s_i \in \{1,\dots,S\}$ is the subject identity.

We employ a frozen EEG foundation model encoder $f_\theta : \mathbb{R}^{C\times T} \to \mathbb{R}^d$ to extract embeddings $\mathbf{z}_i = f_\theta(\mathbf{X}_i)$. The goal is to train a linear classifier $h : \mathbb{R}^d \to \mathbb{R}^K$ on the training subjects' embeddings that generalizes to the held-out subject. A key challenge is that frozen embeddings may encode subject-specific information that helps classify training subjects but does not transfer to new subjects.

3.2 PIPELINE OVERVIEW

Our pipeline consists of four stages, illustrated in Figure 1: (1) Euclidean Alignment (EA) preprocessing to reduce inter-subject distribution shift in the raw EEG signals; (2) feature extraction using a frozen CBraMod (Wang et al., 2024) encoder; (3) Iterative Nullspace Projection (INLP) to remove linearly decodable subject identity from the embeddings; and (4) linear classification for the downstream task. Importantly, the INLP projection matrix is learned only on training subjects and applied to both training and test embeddings, ensuring no information leakage from the held-out subject.

3.3 EUCLIDEAN ALIGNMENT

Euclidean Alignment (EA) (He & Wu, 2018) is a preprocessing step that reduces inter-subject distribution shift by aligning the covariance structure of each subject's EEG trials to a common reference. For each subject s, we compute the mean covariance matrix $\bar{\mathbf{R}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{X}_i\mathbf{X}_i^\top$ over all n_s trials. Each trial is then transformed as $\tilde{\mathbf{X}}_i = \bar{\mathbf{R}}_s^{-1/2}\mathbf{X}_i$, which whitens the data such that the mean covariance of aligned trials equals the identity matrix. This alignment makes the distributions from different subjects more similar, facilitating transfer learning. EA is unsupervised, computationally efficient, and has been shown to consistently improve cross-subject generalization across diverse BCI paradigms (Wu, 2025).

3.4 ITERATIVE NULLSPACE PROJECTION

Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) removes linearly decodable information about a protected attribute from neural representations. Given training-subject embeddings $\mathbf{Z} \in \mathbb{R}^{N\times d}$ and subject labels $s \in \{1,\dots,S\}$, INLP iteratively: (1) trains a linear classifier $\mathbf{W}_t \in \mathbb{R}^{S\times d}$ to predict subject identity from the current embeddings $\mathbf{Z}_t$; (2) computes the orthogonal projection matrix $\mathbf{P}_t$ onto the nullspace of $\mathbf{W}_t$; and (3) updates $\mathbf{Z}_{t+1} = \mathbf{Z}_t\mathbf{P}_t$. The nullspace

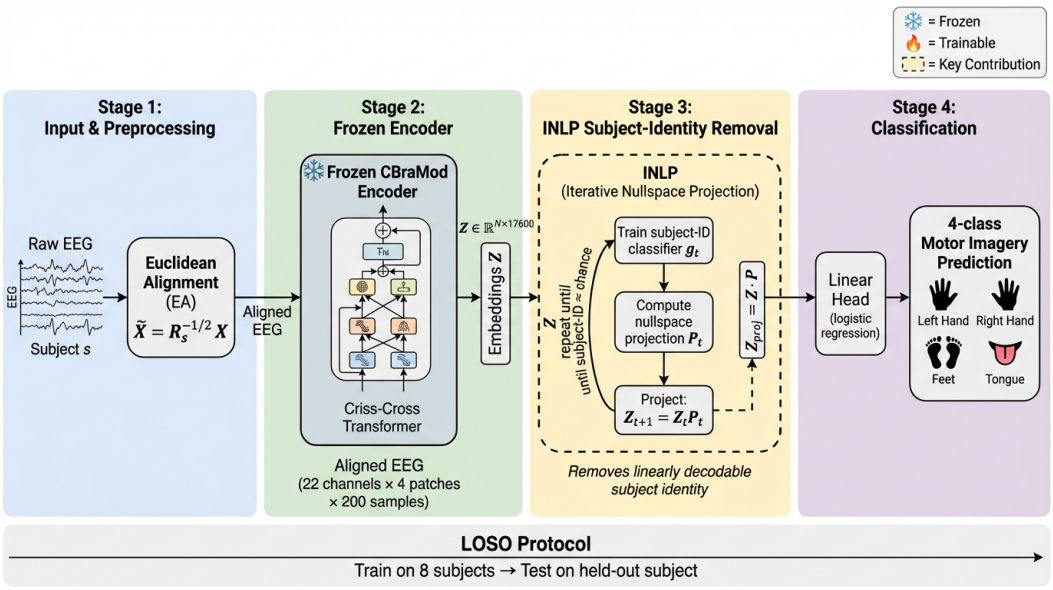

Figure 1: Overview of the INLP-projected linear probing pipeline for cross-subject EEG foundation model transfer. Raw EEG signals undergo Euclidean Alignment (EA) before being processed by a frozen CBraMod encoder. The resulting embeddings are projected through Iterative Nullspace Projection (INLP) to remove linearly decodable subject identity, then classified by a linear head. The LOSO protocol trains on 8 subjects and tests on the held-out subject.

projection ensures that $\mathbf{W}_t(\mathbf{P}_t\mathbf{z}) = \mathbf{0}$ for all $\mathbf{z}$, rendering the classifier's decision boundary useless. After n iterations, the final projection $\mathbf{P} = \prod_{t=1}^{n} \mathbf{P}_t$ projects embeddings onto the intersection of all nullspaces, removing all linearly decodable subject information captured by the sequence of classifiers. The projected embeddings $\mathbf{z}' = \mathbf{P}\mathbf{z}$ are then used for downstream classification.

3.5 Hyperparameter Optimization

The number of INLP iterations n and the regularization strength C of the linear classifier control the aggressiveness of subject-identity removal. Too few iterations may leave residual subject information, while too many may remove task-relevant signal. We optimize these hyperparameters via inner cross-validation within the training subjects. For each outer LOSO fold, we perform an inner LOSO loop over the $S - 1$ training subjects, sweeping $n \in \{1, 2, 3, 4, 5, 7, 10\}$ and $C \in \{0.01, 0.1, 1.0, 10.0\}$. The configuration maximizing inner validation accuracy is selected for the outer fold. This nested procedure ensures that hyperparameter selection does not use information from the held-out test subject.

4 Experiments

4.1 Experimental Setup

We evaluate our approach on BNCI2014001 (Jayaram & Barachant, 2018), a standard motor imagery benchmark from BCI Competition IV-2a. The dataset contains EEG recordings from 9 subjects performing 4-class motor imagery tasks (left hand, right hand, feet, tongue) with 22 channels. We use CBraMod (Wang et al., 2024) as the frozen encoder, extracting 17,600-dimensional embeddings by flattening the output features. Evaluation follows the LOSO protocol: for each fold, we train on 8 subjects and test on the held-out subject, reporting mean accuracy across all 9 folds. All experiments use 3 random seeds (42, 123, 456).

We compare the following methods: (1) **Vanilla LP**: linear probing without EA, as reported in Liu et al. (2026); (2) **Fine-tuning**: full model fine-tuning, as reported in Liu et al. (2026); (3) **EA + Avgpool**: EA preprocessing with average pooling; (4) **EA + Flatten**: EA preprocessing with

Table 1: Cross-subject motor imagery classification accuracy (%) on BNCI2014001 using frozen CBraMod embeddings. EA = Euclidean Alignment. PCA-k removes top-k principal components. INLP-10 = fixed 10 iterations. INLP-CV = inner cross-validation for iteration/C selection. Best in **bold**. Δ shows change vs EA + Flatten baseline.

Method	Accuracy (%)	Std (%)	Δ (pp)
Vanilla LP (reported)*	41.45	0.50	-14.82
Fine-tuning (reported)*	53.03	0.22	-3.24
EA + Avgpool	33.54	6.47	-22.73
EA + Flatten	**56.27**	8.09	—
EA + PCA-1	55.47	7.37	-0.80
EA + PCA-10	51.47	5.42	-4.80
EA + INLP-10	55.18	7.19	-1.09
EA + INLP-CV	**56.29**	7.98	+0.02

*Values from Liu et al. (2026).

flattening (our strong baseline); (5) **EA + PCA-k**: EA with removal of top-k principal components as a control for generic dimension reduction; (6) **EA + INLP-10**: EA with fixed 10 INLP iterations; and (7) **EA + INLP-CV**: EA with INLP hyperparameters selected via inner cross-validation (our proposed method).

4.2 Main Results

Table 1 presents the cross-subject classification results. The most striking finding is that our EA baseline with flattened embeddings achieves 56.27% accuracy, substantially exceeding both vanilla linear probing (41.45%, +14.82 pp) and reported full fine-tuning (53.03%, +3.24 pp). This establishes that Euclidean Alignment combined with frozen CBraMod embeddings provides a remarkably strong baseline, leaving limited room for further improvement.

Our proposed INLP-CV method achieves 56.29% accuracy, matching but not meaningfully exceeding the EA baseline. The improvement of +0.02 pp falls well within the standard deviation (7.98%) and seed variance, indicating no statistically significant benefit from subject-identity removal. This result directly refutes our hypothesis that removing linearly decodable subject information would improve cross-subject transfer. The pre-defined success criterion of +2 pp improvement over the EA baseline is not met.

4.3 Control Experiments

To distinguish whether INLP's lack of benefit stems from subject-identity removal specifically or from dimension reduction generally, we include PCA-k controls that remove the top-k principal components without regard to subject identity. The results reveal that all dimension removal methods degrade performance relative to the EA baseline. PCA-1 achieves 55.47% (-0.80 pp), and accuracy decreases monotonically as more components are removed, reaching 51.47% (-4.80 pp) for PCA-10. This pattern indicates that high-variance directions in frozen CBraMod embeddings contain task-relevant information that should not be discarded.

The fixed INLP-10 configuration, which applies 10 iterations of nullspace projection without optimization, achieves 55.18% (-1.09 pp), performing worse than even PCA-1. This suggests that aggressive subject-identity removal is particularly harmful, removing more task-relevant information than generic PCA. The comparison between INLP-10 and INLP-CV demonstrates that the optimization procedure learns to minimize intervention rather than maximize subject-identity removal, a finding we analyze further in the next section.

4.4 Analysis: Why INLP Self-Nullifies

Figure 2 reveals why optimized INLP achieves nearly identical accuracy to the EA baseline: the inner cross-validation consistently selects configurations that minimize the projection's effect. Across

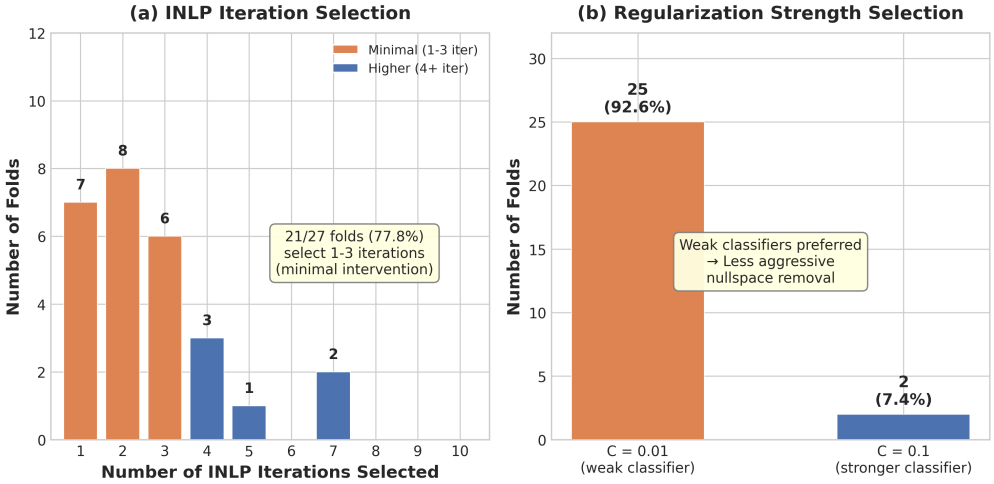

Figure 2: Distribution of INLP hyperparameter selections across 27 LOSO folds (9 subjects $\times$ 3 seeds). (a) Inner cross-validation predominantly selects 1–3 INLP iterations (77.8% of folds), indicating minimal intervention is optimal. (b) The weak regularization setting $C = 0.01$ is selected in 92.6% of folds, further limiting the aggressiveness of nullspace removal.

Table 2: Per-subject classification accuracy (%) comparing EA baseline and optimized INLP. Δ shows the change from EA to INLP-CV. Mixed effects with no consistent pattern: 4 subjects show slight improvement while 5 degrade.

Subject	EA Baseline (%)	INLP-CV (%)	Δ (pp)
S1	60.42	**61.81**	+1.39
S2	40.10	**40.91**	+0.81
S3	**62.50**	62.44	−0.06
S4	50.00	**50.62**	+0.62
S5	**51.39**	51.27	−0.12
S6	**53.99**	52.89	−1.10
S7	63.89	**64.76**	+0.87
S8	**64.58**	64.06	−0.52
S9	**59.55**	57.87	−1.68
Mean	56.27	56.29	+0.02

27 LOSO folds (9 subjects $\times$ 3 seeds), 77.8% select only 1–3 INLP iterations, and 92.6% select the weakest regularization setting ($C = 0.01$). With few iterations and a weak classifier, the learned projection matrix remains close to identity, meaning INLP effectively learns to do almost nothing.

This self-nullification behavior provides insight into the structure of frozen CBraMod embeddings. If subject-identity information were separable from task-discriminative signal, we would expect the optimization to select more aggressive configurations that remove subject information while preserving task accuracy. Instead, the optimization learns to avoid intervention, suggesting that subject identity and task signal occupy overlapping subspaces in the embedding space. Removing subject-identity directions necessarily removes task-relevant information, explaining why all dimension removal methods degrade performance.

4.5 Per-Subject Analysis

Table 2 presents the per-subject breakdown of classification accuracy. The results show mixed effects with no consistent pattern: 4 subjects (S1, S2, S4, S7) show slight improvement with INLP-CV, while 5 subjects (S3, S5, S6, S8, S9) show degradation. The largest improvement occurs for S1 (+1.39 pp), while the largest degradation occurs for S9 (−1.68 pp). These changes are small relative

to the inter-subject variance and show no systematic relationship with baseline performance level. The absence of a consistent benefit pattern further supports our conclusion that subject-identity removal does not provide a reliable mechanism for improving cross-subject transfer with frozen EEG foundation model embeddings.

5 CONCLUSION

We investigated whether Iterative Nullspace Projection (INLP) can improve cross-subject transfer of frozen EEG foundation model embeddings by removing linearly decodable subject identity. Our experiments on BNCI2014001 motor imagery classification refute this hypothesis: optimized INLP achieves 56.29% accuracy, statistically indistinguishable from the Euclidean Alignment baseline (56.27%). The key insight is that subject-identity information in frozen CBraMod embeddings is entangled with task-discriminative signal, as evidenced by the optimization consistently selecting minimal intervention configurations. This finding suggests that post-hoc linear debiasing is insufficient for improving frozen EEG-FM transfer; future work should explore non-linear disentanglement methods or joint training approaches that can separate subject and task information during representation learning. Our study is limited to a single dataset and encoder, and extending these findings to other EEG paradigms and foundation models remains an important direction.

REFERENCES

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *ArXiv*, abs/1607.06520, 2016.

Konstantinos Bousmalis, George Trigeorgis, N. Silberman, Dilip Krishnan, and D. Erhan. Domain separation networks. pp. 343–351, 2016.

Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, François Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. pp. 189–209, 2015.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. pp. 60–63, 2019.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67:399–410, 2018.

V. Jayaram and A. Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of Neural Engineering*, 15, 2018.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. *ArXiv*, abs/2405.18765, 2024.

Demetres Kostas, Stephane T Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15, 2021.

Dingkun Liu, Yuheng Chen, Zhu Chen, Zhenyao Cui, Yaozhi Wen, Jiayu An, Jingwei Luo, and Dongrui Wu. Eeg foundation models: Progresses, benchmarking, and open problems. *ArXiv*, abs/2601.17883, 2026.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. pp. 7237–7256, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. pp. 443–450, 2016.

Imad Eddine Tibermacine, Samuele Russo, Ahmed Tibermacine, A. Rabehi, B. Nail, Kamel Kadri, and Christian Napoli. Riemannian geometry-based eeg approaches: A literature review. *ArXiv*, abs/2407.20250, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. *ArXiv*, abs/2412.07236, 2024.

Dongrui Wu. Revisiting euclidean alignment for transfer learning in eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 22, 2025.

Chaoqi Yang, M. Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. 2023.

A Appendix

APPENDIX TEXT