SCoRE: LEARNABLE REFLECTION EQUIVARIANCE FOR CROSS-SUBJECT EEG DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) is the principal non-invasive modality for brain-computer interfaces (BCI). Generalizing across pronounced inter-individual variability has long been a crucial challenge for cross-subject decoding. An intuitive strategy for addressing this challenge is to scale up: pretrain high-capacity models on large corpora to obtain subject-invariant representations. Yet scaling up increases model complexity and inference cost, and the shared structure such models capture remains implicit in their weights. We instead propose *SCoRE*, a method that remains compact while making this structure explicit through reflection equivariance, an inductive bias for EEG decoding. Physiologically, its most familiar manifestation is contralateral motor control: each hemisphere controls the opposite hand, and hand movements evoke mirror-image activity across the scalp. We posit that reflection equivariance is not confined to motor control but is intrinsic to brain activity. We operationalize this property by constraining the classifier such that a spatial reflection of the signal induces a corresponding transformation of class evidence. Specifically, *SCoRE* trains a shared network on each input and its reflection, constructing an equivariant classifier through joint training and parameterization. This classifier separately captures class evidence that reverses under reflection and that remains invariant. Notably, *SCoRE* consistently outperforms compact EEG baselines by 10.37 accuracy points and surpasses state-of-the-art pretrained models at a fraction of their size. Our method not only advances cross-subject decoding but also offers neuroscientific insight. The learned input reflection and output transformation are interpretable and agree with established physiology. *SCoRE*, together with EEG experiments, opens a new pathway for discovering reflection-equivariant structure across tasks and brain states.

1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive window into human brain activity and underpins applications such as brain-computer interfaces (BCIs), continuous sleep monitoring, and the assessment of cognitive and affective states (Roy et al., 2019). Although deep representation learning has driven rapid progress in EEG decoding (Wang et al., 2025), cross-subject generalization remains a central challenge. This challenge stems not only from intrinsic signal properties such as low signal-to-noise ratio and pronounced non-stationarity, but, more fundamentally, from substantial inter-individual variability (Wu et al., 2022).

Existing approaches tackle inter-individual variability through two broad strategies. The first trains subject-specific models using individually collected EEG data, as in BCI-assisted stroke rehabilitation (Biasiucci et al., 2018), which requires dedicated acquisition and calibration. To reduce the need for repeated subject-specific training, the second strategy pools data across subjects to learn a shared model, which is subsequently deployed to unseen individuals either zero-shot or after personalization with a few labeled samples. Such models directly learn from task-specific data and differ mainly in module design: EEGNet (Lawhern et al., 2018) uses depthwise and separable convolutions; EEGConformer (Song et al., 2023) combines convolution with self-attention; SPaRCNet (Jing et al., 2023) uses densely connected convolutions. To learn transferable representations, pretraining draws on large-scale EEG corpora spanning diverse tasks and recording configurations. BIOT (Yang et al., 2023a) tokenizes individual channels to accommodate heterogeneous recordings; LaBraM (Jiang et al., 2024) predicts discrete neural codes; CBraMod (Wang et al., 2025)

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.

separates spatial and temporal attention during masked reconstruction. Beyond pretraining objectives, REVE (El Ouahidi et al., 2025) combines masked autoencoding with electrode-coordinate and temporal positional encodings to support different recording configurations.

Nevertheless, existing cross-subject methods face a trade-off between decoding performance and resource requirements. Compact models trained on task-specific data may fail to capture inter-individual variability. Pretraining can improve transferability (El Ouahidi et al., 2025) but is computationally and data intensive. Architectures accommodating heterogeneous montages and recording protocols can increase complexity (Yang et al., 2023a). Rarer tasks may remain underrepresented, and mismatches between pretrained representations and target distributions can lead to negative transfer (Wang et al., 2019). These limitations motivate the explicit modeling of physiological structure shared across individuals as a complementary inductive bias for cross-subject decoding.

Which patterns of brain activity are shared across individuals? A familiar example is sensorimotor lateralization: hand motor imagery can elicit contralateral activity, motivating a left–right correspondence between EEG patterns and hand-related class evidence (Pfurtscheller & Neuper, 1997). Shared regularities also include coordinated regional responses. Higher working-memory load engages frontoparietal regions while suppressing the default mode network (Owen et al., 2005; McKiernan et al., 2003). In passive perception, auditory and visual stimuli can elicit reciprocal activation and deactivation of the corresponding sensory cortices (Laurienti et al., 2002). These observations motivate testing whether a learnable reflection of EEG inputs, coupled with a task-dependent transformation of class evidence, provides a useful structural bias for decoding.

To test this hypothesis, we propose Structured Contrast Reflection Equivariance (*SCoRE*), a compact and efficient network that encodes reflection equivariance as a structural inductive bias (Figure 1). A shared backbone processes each EEG input and its spatial reflection under an involutive constraint. Before fusion, a Householder reflection maps the reflected logits back to the original output space. This parameterization satisfies the output constraints while reducing the transformation to a single learnable direction. After warm-up, this direction is initialized from the principal axis of reflection-induced logit changes and jointly optimized for multiclass tasks. By construction, the classifier guarantees exact logit equivariance under deterministic inference, regardless of the backbone.

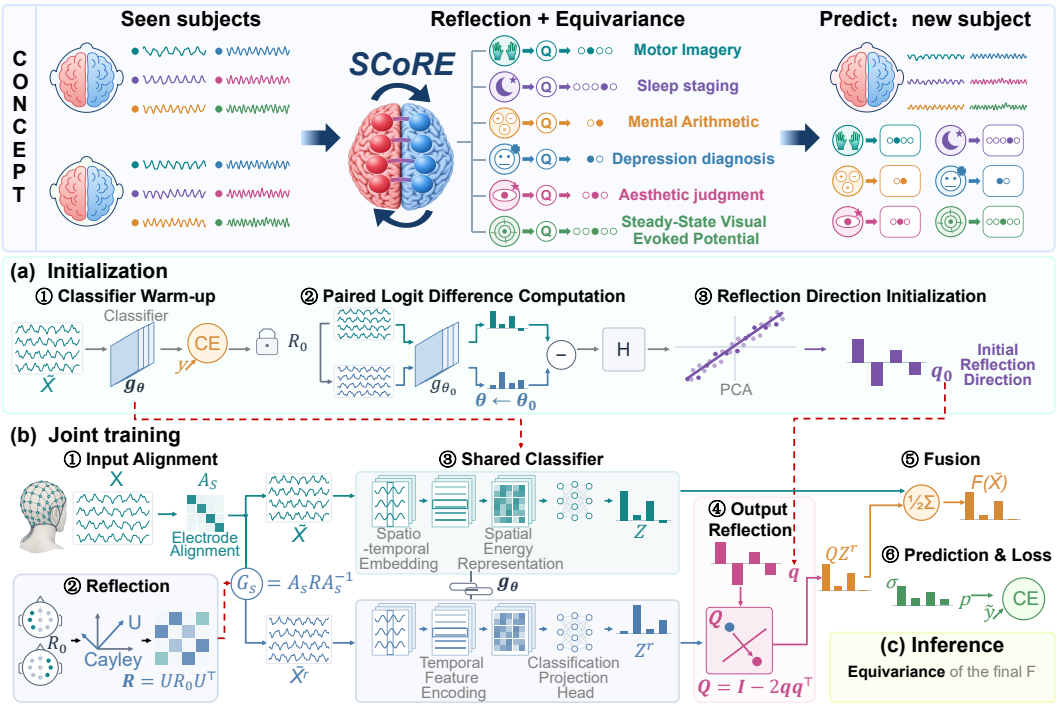


Figure 1: Overview of *SCoRE*. The top panel shows the overall concept across subjects and tasks. (a) Initialization of the output reflection direction. (b) Joint training with learnable input and output reflections. (c) Equivariant inference.

The main contributions of this work are summarized as follows:

• **A novel reflection-equivariant model.** We propose *SCoRE*, a compact network that jointly learns input and output reflections with a shared classifier. The output reflection adapts to each task, and the model is exactly equivariant by construction.

• **Open-access EEG datasets.** We collect and publicly release three EEG datasets covering motor imagery, steady-state visual evoked potentials, and aesthetic judgment, comprising 84 subjects and over 49 hours of recordings.

• **Physiological interpretability.** *SCoRE* enhances model interpretability and provides neuroscientific insight. Its learned reflections align with known neural activity and hold promise as biomarkers.

• **Extensive empirical evaluation.** We evaluate *SCoRE* on nine datasets spanning six tasks across cross-subject, subject-dependent, and personalization settings. In cross-subject decoding, it outperforms non-pretrained baselines and rivals pretrained foundation models with far fewer parameters.

2 PRELIMINARIES

2.1 EUCLIDEAN ALIGNMENT

To mitigate distribution shifts between training and target subjects, we employ Euclidean alignment (He & Wu, 2020). Given EEG trials $X_i^{(s)} \in \mathbb{R}^{C\times T}$ from session s, with C channels and T temporal samples, we compute the reference matrix, the alignment matrix, and the aligned trials as

$$\Sigma_s = \frac{1}{N_s}\sum_{i=1}^{N_s} X_i^{(s)} X_i^{(s)\top}, \qquad A_s = \Sigma_s^{-1/2}, \qquad \widetilde{X}_i^{(s)} = A_s X_i^{(s)}, \tag{1}$$

where N_s is the trial count in session s. Following alignment, each session's reference matrix becomes the identity matrix I_C, aligning spatial second-order statistics across subjects and sessions.

2.2 EEG REFLECTION

Reflection. Reflection transforms EEG signals according to spatial correspondences among channels. For example, left–right reflection swaps signals between bilateral electrode pairs (e.g., C3–C4, FC3–FC4, and CP3–CP4) while preserving signals at midline electrodes (e.g., Cz).

Extending this operation to the full electrode montage yields a reflection matrix $R \in \mathbb{R}^{C\times C}$. In the aligned coordinates, the reflected input $\widetilde{X}^r$ is given by

$$\widetilde{X}^r = G_s\widetilde{X} = A_s R X, \qquad G_s = A_s R A_s^{-1}, \tag{2}$$

where R is learnable and shared across subjects and sessions, whereas G_s depends on the subject–session-specific alignment matrix A_s. The involutive constraint $R^2 = I_C$ ensures that two successive reflections restore the aligned input, i.e., $G_s^2\widetilde{X} = G_s\widetilde{X}^r = A_s R^2 A_s^{-1}\widetilde{X} = \widetilde{X}$.

Equivariance. Equivariance links input transformations to corresponding output transformations. Sensorimotor lateralization (Pfurtscheller & Neuper, 1997) suggests that left–right reflection may transform EEG evidence for right-hand motor imagery into evidence for left-hand motor imagery.

Beyond left–right hand imagery, we extend this formulation to arbitrary classes and states. We assume that reflection-equivariant structure is intrinsic to the brain activity recorded by EEG. We model this structure using a K-class classifier F satisfying

$$F(G_s\widetilde{X}) = QF(\widetilde{X}), \tag{3}$$

where $Q \in \mathbb{R}^{K\times K}$ denotes an output transformation shared across subjects and sessions. The problem thus becomes learning Q and constructing an equivariant classifier F. The model learns from the training data which class responses and output patterns are affected by reflection.

Constraints. Since input reflection is involutive, applying Q twice must restore the logits. Moreover, softmax probabilities are unchanged by adding the same scalar to all logits. To preserve this invariance, we constrain Q to leave such common shifts unchanged. We therefore impose

$$Q^2 = I_K, \qquad Q\mathbf{1} = \mathbf{1}, \tag{4}$$

where I_K denotes the $K \times K$ identity matrix and $\mathbf{1} \in \mathbb{R}^K$ is the all-ones vector.

2.3 HOUSEHOLDER REFLECTION

A Householder reflection reverses the component along a unit vector while preserving all orthogonal components. To satisfy the constraints in equation 4, we parameterize Q as

$$Q = I_K - 2qq^\top, \qquad q^\top q = 1, \qquad q^\top \mathbf{1} = 0, \tag{5}$$

where $q \in \mathbb{R}^K$ denotes a learnable direction in the logit space. The unit-norm condition ensures $Q^2 = I_K$, while orthogonality to $\mathbf{1}$ gives $Q\mathbf{1} = \mathbf{1}$. This parameterization reduces learning Q to learning a class-contrast direction q whose component changes sign under reflection.

3 METHODOLOGY

SCoRE is trained in two stages (Fig. 1). (1) **Warm-up and Contrast Initialization** establishes discriminative class evidence to guide the initialization of a task-relevant reflection contrast. (2) **Joint Equivariant Training** learns discriminative EEG representations and the input–output reflection correspondence, unifying classification and reflection equivariance within a unified model. Related approaches and their connections to *SCoRE* are discussed in Appendix E.

3.1 WARM-UP AND CONTRAST INITIALIZATION

Classifier Warm-up. We warm up an EEG classifier g_θ by supervising its responses to the aligned, unreflected inputs. This warm-up establishes a class-discriminative logit space in which reflection-induced changes can be assessed. We denote the resulting classifier parameters by θ_0.

Paired Logit Differences. For each training trial, we compute the logit difference between the aligned input and its reflected counterpart:

$$\Delta_n = g_{\theta_0}\left(\widetilde{X}_n\right) - g_{\theta_0}\left(\widetilde{X}_n^r\right), \qquad d_n = H\Delta_n, \tag{6}$$

where $n = 1, \ldots, N_{\mathrm{tr}}$ indexes all training trials; equation 1 and equation 2 define $\widetilde{X}_n$ and $\widetilde{X}_n^r$, respectively. At this stage, the input reflection is fixed at the initial electrode-pairing matrix R_0. The softmax-gauge projection $H = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ removes common logit shifts by projecting Δ_n onto the zero-sum logit space, yielding d_n, which captures reflection-induced changes in class contrasts.

Contrast Initialization. We apply principal component analysis (PCA) to the projected logit differences to initialize the reflection-sensitive class contrast. After centering across training trials, we extract the leading principal direction:

$$q_0 = \mathrm{PCA}_1\left(\{d_n\}_{n=1}^{N_{\mathrm{tr}}}\right), \tag{7}$$

where PCA_1 returns the unit vector along the first principal axis in the zero-sum logit space. The resulting q_0 initializes the sign-changing direction q of the Householder reflection in equation 5. For subsequent joint training, we set $\theta \leftarrow \theta_0$ and $q_a \leftarrow q_0$.

3.2 JOINT EQUIVARIANT TRAINING

Learnable Input Reflection. We parameterize R by conjugating R_0 with a learnable orthogonal matrix constructed via the Cayley transform. This construction preserves the involutive constraint $R^2 = I_C$ throughout optimization.

Shared Classifier Network. A shallow convolutional classifier g_θ processes the aligned and reflected inputs. Temporal and spatial convolutions learn spatiotemporal responses, followed by batch normalization, response squaring, temporal average pooling, and logarithmic compression. Adaptive temporal average pooling yields a fixed-dimensional log-energy representation while retaining coarse temporal structure. A nonlinear embedding and a linear classification head then map these representations to K-class logits. Both branches share all parameters and are trained end-to-end.

Learnable Output Reflection. We learn the reflection direction q by optimizing q_a through the parameterization $q = Hq_a/\|Hq_a\|_2$, with $Hq_a \neq 0$. The softmax-gauge projection H and normalization ensure $q^\top \mathbf{1} = 0$ and $q^\top q = 1$ throughout optimization. The resulting q defines the output reflection Q through the Householder parameterization in equation 5.

Joint Optimization. For each training trial, the shared network produces class logits for the aligned input and its reflected counterpart:

$$z_n = g_\theta(\widetilde{X}_n), \qquad z_n^r = g_\theta(\widetilde{X}_n^r). \tag{8}$$

Since $Q^{-1} = Q$, we map the reflected logits back to the original output coordinates before averaging. The resulting classifier and class probabilities are

$$F(\widetilde{X}_n) = \frac{1}{2}\left(z_n + Qz_n^r\right), \qquad p_n = \mathrm{softmax}\Big(F(\widetilde{X}_n)\Big). \tag{9}$$

We jointly optimize θ, q_a, and the learnable parameters of R by minimizing label-smoothed cross-entropy on the aggregated predictions, with smoothing coefficient ε.

Equivariance Guarantee. For deterministic g_θ, input reflection exchanges z_n and z_n^r because $G_s^2 = I_C$. Since $Q^2 = I_K$, we can substitute $z_n^r = Q^2 z_n^r$ and factor out Q. Together with equation 8 and equation 9, these identities yield

$$F(G_s\widetilde{X}_n) = \tfrac{1}{2}(z_n^r + Qz_n) = \tfrac{1}{2}(Qz_n + Q^2 z_n^r) = \tfrac{1}{2}Q(z_n + Qz_n^r) = QF(\widetilde{X}_n), \tag{10}$$

which establishes equation 3. Empirical verification is provided in Appendix D.

Remark 1. For $K = 2$, the one-dimensional zero-sum logit space reduces PCA initialization to the closed-form direction $q = \pm(1, -1)^\top/\sqrt{2}$, yielding a fixed class-swap reflection Q. With Q fixed, joint training still learns task-adaptive input reflections and discriminative representations through R and g_θ, while preserving exact logit equivariance under deterministic inference.

4 Experiments

Public benchmarks. We evaluate *SCoRE* on nine datasets spanning six EEG decoding tasks. The public benchmarks include BCI-IV-2a and PhysioNet-MI for motor imagery (Tangermann et al., 2012; Schalk, 2009), ISRUC and HMC for sleep staging (Khalighi et al., 2016; Alvarez-Estevez & Rijsman, 2021), MentalArithmetic (MAT) for stress detection (Zyma et al., 2019), and Mumtaz for depression detection (Mumtaz, 2016). For these benchmarks, we follow the train/validation/test splits used in prior studies to maintain comparability with existing results.

New open EEG datasets. We additionally collect and release HandMI for left- and right-hand motor imagery, SSVEP-EEG for four-frequency visual stimulation and rest, and AesthEEG for aesthetic judgments of faces and landscapes. Together, these datasets comprise 84 participants and over 49 hours of EEG recordings. We use subject-disjoint train/validation/test splits, assigning all recordings from each participant to a single partition. Across these datasets, *SCoRE* and all baselines use identical data partitions and a common evaluation protocol to ensure fair comparisons. Table 3 summarizes the datasets, with detailed descriptions and preprocessing in Appendices A and B.

Baselines. We compare *SCoRE* against two groups of baselines. Non-pretrained models include EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2023), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2023b), CNN-Transformer (Peh et al., 2022), FFCL (Li et al., 2022), and ST-Transformer (Song et al., 2021). Pretrained models include BIOT (Yang et al., 2023a), LaBraM-Base (Jiang et al., 2024), CBraMod (Wang et al., 2025), and REVE (El Ouahidi et al., 2025).

Implementation. Unless otherwise specified, we set the number of temporal filters to 40, the embedding dimension to 96, and the dropout rate to 0.5. We optimize *SCoRE* using AdamW with a batch size of 64 and a cosine learning-rate schedule preceded by linear warm-up. The base learning rates are 3×10^{-4} for the convolutional encoder and the input reflection, and 10^{-3} for the embedding layer, the classification head, and the output contrast. Code and the three new datasets are available on Anonymous GitHub and Figshare.

Metrics. We primarily use balanced accuracy for all tasks. Secondary metrics include Cohen's κ and weighted F1 for multiclass classification, and AUROC and AUC-PR for binary classification. Result provenance and evaluation protocols are detailed in Appendix B.8.

5 RESULTS

5.1 TASK PERFORMANCE AND ABLATION STUDY

Results. *SCoRE* achieves the highest balanced accuracy on eight of nine datasets and outperforms all non-pretrained baselines on every dataset (Table 1). Its macro-average balanced accuracy reaches 76.30%, exceeding EEGConformer (65.93%), the strongest non-pretrained baseline, by 10.37 percentage points. It also surpasses the strongest pretrained baseline, REVE (72.58%), by 3.72 percentage points. These results establish reflection-equivariant modeling as an effective route to cross-subject generalization through task-specific training, without large-scale EEG pretraining. These gains extend across distinct forms of brain activity. *SCoRE* improves over REVE by 4.53 and 5.78 percentage points on BCI-IV-2a and HMC respectively, and also leads on ISRUC, MAT, and Mumtaz, extending its advantage from motor imagery to sleep staging, stress detection, and depression detection. This breadth supports the reflection-based structural prior beyond motor imagery. The newly collected datasets cover paradigms underrepresented in EEG pretraining corpora, notably aesthetic judgment. Pretrained baselines barely outperform non-pretrained models, whereas *SCoRE* leads on all three datasets, supporting reflection-based priors under limited pretraining coverage. Full results across evaluation metrics and detailed statistical comparisons are provided in Appendices B and B.9, respectively.

Table 1: Balanced accuracy (%, mean ± SD; N as specified in Table 13) across nine EEG datasets. Best and second-best means per column are bold and underlined, respectively. Significance versus second-best (two-sided Welch's t-test): $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. MAT: MentalArithmetic; Macro: unweighted mean across datasets.

Method	BCI-IV-2a	PhysioNetMI	ISRUC	HMC	MAT
EEGNet	44.82±0.94	58.14±1.25	71.54±1.21	68.56±0.64	67.70±1.16
EEGConformer	46.96±1.06	60.49±1.04	74.00±1.33	72.10±1.02	68.05±1.23
SPaRCNet	46.35±1.17	59.32±1.52	74.87±0.75	47.56±11.09	68.79±1.07
ContraWR	46.78±1.25	58.92±1.33	74.02±1.26	42.42±5.41	66.31±0.97
CNN-Transformer	46.00±1.08	60.53±1.18	73.63±0.87	65.73±1.41	67.79±2.68
FFCL	44.70±1.43	57.26±0.92	72.77±1.82	44.27±7.02	67.98±1.42
ST-Transformer	45.75±1.45	60.35±0.81	73.81±2.05	25.59±1.41	66.31±1.73
BIOT	47.48±0.93	61.53±1.54	75.27±1.21	68.62±0.41	68.75±1.86
LaBraM-Base	48.69±0.85	61.73±1.22	76.33±1.02	72.86±1.01	69.09±1.25
CBraMod	51.38±0.66	<u>64.17</u>±0.91	<u>78.65</u>±1.10	73.35±0.47	72.56±1.32
REVE	<u>63.96</u>±0.95	**64.80**±1.40	78.19±0.78	<u>74.01</u>±0.75	<u>76.60</u>±3.55
SCoRE	**68.49*****±1.56	63.95±0.77	**79.25**±0.37	**79.79*****±0.30	**78.47**±2.62
	Mumtaz	AesthEEG	HandMI	SSVEP-EEG	Macro
EEGNet	92.32±1.04	65.54±0.78	51.25±3.02	72.85±2.84	65.86
EEGConformer	93.08±1.17	66.13±1.36	65.00±1.77	47.58±7.70	65.93
SPaRCNet	93.16±0.95	48.50±5.92	64.58±3.12	39.34±4.13	60.27
ContraWR	91.95±1.15	65.22±1.05	59.17±4.58	67.01±1.28	63.53
CNN-Transformer	93.05±0.68	35.80±1.93	67.04±2.62	54.53±5.50	62.68
FFCL	93.14±0.38	44.47±0.44	<u>67.46</u>±2.49	49.10±5.76	60.13
ST-Transformer	91.35±1.03	51.95±1.00	51.13±3.41	45.07±4.88	56.81
BIOT	93.58±0.52	<u>66.33</u>±0.79	64.17±2.45	72.43±2.72	68.68
LaBraM-Base	94.09±0.79	65.61±0.75	65.96±0.52	71.92±0.30	69.59
CBraMod	95.60±0.56	55.47±0.34	67.38±2.39	<u>73.97</u>±0.21	70.28
REVE	<u>96.44</u>±0.97	63.79±0.43	62.38±2.77	73.08±1.71	<u>72.58</u>
SCoRE	**97.44***±0.56	**68.24*****±0.18	**70.13***±2.77	**80.91*****±1.29	**76.30**

Ablation. Table 2 shows that the full model achieves the highest balanced accuracy on all five datasets. Fixing the input reflection reduces macro-average balanced accuracy by 3.96 percentage points, supporting the benefit of learning task-adaptive spatial structure. Removing Euclidean alignment or alignment conjugation reduces performance by 5.68 and 2.56 percentage points, respec-

tively. These results highlight the importance of learning reflections in conjunction with subject-specific alignment. The shared branches and constrained output reflection both contribute to the gains. Using either branch alone reduces macro-average balanced accuracy by 3.60–7.43 percentage points, while untying branch parameters incurs a 4.62-point decrease despite introducing a second backbone. Replacing the output reflection with invariant averaging or an unconstrained mapping reduces performance by 3.52 and 3.49 percentage points, respectively. Freezing the output contrast also degrades performance on all three multiclass datasets. Together, these comparisons support learning task-dependent transformations of class evidence within the reflection constraints. The initialization strategy further improves performance. Removing supervised warm-up reduces macro-average balanced accuracy by 3.39 percentage points. Random contrast initialization and joint training from scratch also underperform on all three multiclass datasets, supporting the initialization of the reflection contrast from discriminative representations before joint optimization.

Table 2: Ablation study of *SCoRE*. Balanced accuracy (%) is reported as mean ± SD ($N = 10$). Macro denotes the unweighted average of the five dataset means. Bold and underlining indicate the best and second-best means in each column, respectively.

Variant	BCI-IV-2a	PhysioNetMI	HMC	MAT	Mumtaz	Macro
SCoRE	**68.49**±1.56	**63.95**±0.77	**79.79**±0.30	**78.47**±2.62	**97.44**±0.56	**77.63**
w/o Alignment	60.90±2.73	59.90±0.62	76.14±0.90	69.55±5.54	93.27±0.73	71.95
Original Branch Only	64.84±2.01	59.96±1.49	76.82±0.74	73.16±3.68	95.38±0.38	74.03
Reflected Branch Only	57.55±4.53	51.49±0.28	78.42±1.47	70.49±2.43	93.06±3.63	70.20
Untied Branches	64.84±1.81	58.78±0.81	78.64±0.83	67.26±1.78	95.55±2.34	73.01
Fixed Input Reflection	<u>65.33</u>±0.64	58.62±0.28	76.60±0.50	<u>74.97</u>±2.37	92.82±1.45	73.67
w/o Alignment Conjugation	64.82±1.51	<u>63.34</u>±0.19	79.11±0.79	72.74±1.23	95.33±0.74	<u>75.07</u>
Fixed Output Contrast	64.92±1.40	59.26±0.46	78.76±0.83	—	—	—
Invariant Averaging	62.21±1.18	58.98±1.19	79.05±0.82	73.72±3.21	<u>96.57</u>±2.16	74.11
Unconstrained Output Mapping	64.44±2.27	59.30±0.87	<u>79.49</u>±0.34	70.94±8.40	96.54±0.35	74.14
w/o Warm-up	64.86±1.36	57.10±0.94	79.27±1.38	73.85±2.81	96.10±1.82	74.24
Random Contrast Initialization	64.04±2.78	58.87±1.36	78.44±0.28	—	—	—
Joint Training from Scratch	63.17±1.71	56.14±2.16	78.74±0.36	—	—	—

5.2 Subject-Dependent Decoding and Personalization

SCoRE's advantage extends beyond cross-subject transfer. Under subject-dependent training (Figure 2, left), reflection-constrained learning also supports effective within-subject discrimination. The shared structural prior is thus compatible with exploiting individual-specific training data.

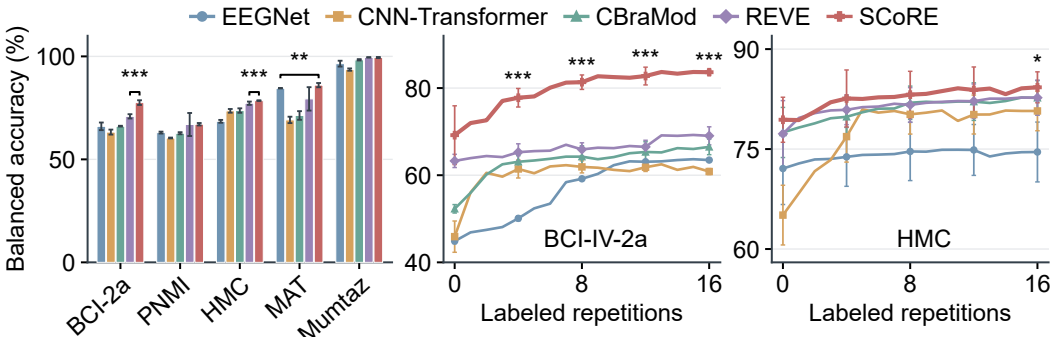


Figure 2: Balanced accuracy for subject-dependent decoding across five datasets (left) and personalization on BCI-IV-2a (middle) and HMC (right).

The personalization results demonstrate a reduced need for individual calibration. On BCI-IV-2a (middle), *SCoRE* reaches the performance of the strongest baseline at the largest evaluated annotation budget without individual labels. A small annotation budget then yields substantial additional gains, providing a clear advantage when calibration data are scarce. On HMC (right), further improvements from an already strong initial model show that population-level generalization and individual adaptation can be complementary. Zero-shot deployment and personalization with few la-

beled samples reduce calibration demands, supporting practical applications such as brain-computer control and sleep monitoring. Detailed protocols and results are provided in Appendices C.1 and C.2.

5.3 Mechanistic Interpretation

Figure 3(a) shows bilateral organization in the learned R and pronounced left/right-hand coupling emerging in Q during training, consistent with contralateral sensorimotor organization (Pfurtscheller & Neuper, 1997). The Q-weighted low-β suppression maps (b) connect the learned output action to an established motor-imagery marker. Contralateral hand and midline foot patterns accord with motor somatotopy (Ehrsson et al., 2003), alongside a spatially distinct tongue pattern. Decomposing route evidence as $z_{\pm} = \frac{1}{4}(I \pm Q)H(z \pm z^r)$ reveals that hand discrimination is carried predominantly by the odd component, whereas feet and tongue are distinguished by the even component (c). This separation preserves both reflection-sensitive and reflection-stable class evidence. Detailed analysis protocols are provided in Appendix C.5.

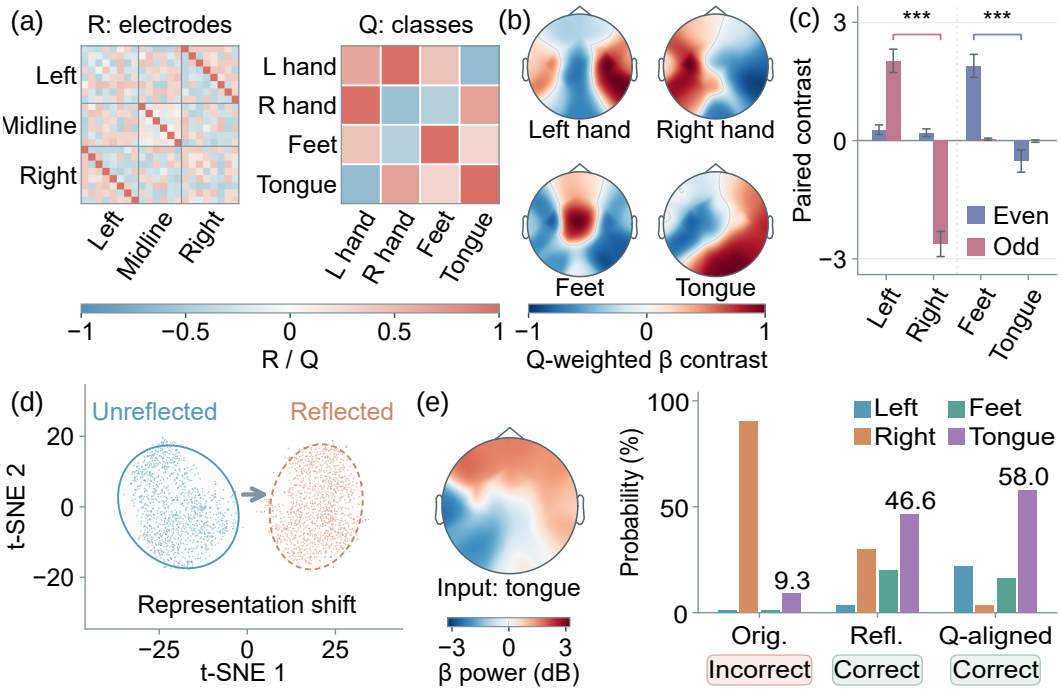


Figure 3: Mechanistic analysis of *SCoRE* on BCI-IV-2a. (a) Learned input and output reflection matrices R and Q. (b) Normalized Q-weighted low-β suppression contrasts (13–20 Hz). (c) Class-wise reflection-even and reflection-odd logit contrasts. (d) Joint t-SNE of unreflected and reflected route embeddings. (e) Input topography and class probabilities from z, z^r, and Qz^r for a tongue-imagery trial. Bars in (c) denote mean $\pm$ SD across ten runs, $^{***}p < 0.001$ (two-sided t-tests).

Joint t-SNE (d) reveals distinct representations of the original and reflected views in a shared embedding space. Their predictive complementarity is illustrated by a tongue trial (e): the reflected route corrects the original misclassification, and Q alignment further strengthens the correct-class evidence. These findings link the equivariant structure to physiological organization and complementary predictive evidence, offering an interpretable account of how reflection supports decoding.

5.4 Quantitative Analysis of Established Markers

On HMC, masking established EEG biomarkers produces higher prediction flip rates than matched random masking (Figure 4, left, all $p < 0.001$), demonstrating that *SCoRE* uses physiological information for sleep staging. The magnitude of R-induced changes also correlates strongly with spectral biomarkers ($|\rho| \geq 0.60$, right), providing a quantitative physiological interpretation of the learned input reflection. Experimental protocols and supporting biomarker references are provided in Appendix C.6.

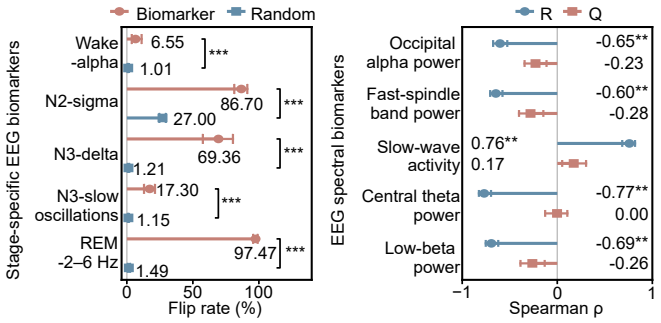


Figure 4: Quantitative assessment of *SCoRE* against established EEG biomarkers. Left, stage-specific prediction flip rates under biomarker and matched random masking. Right, Spearman correlations of spectral biomarkers with R and Q action magnitudes. Error bars indicate 95% bootstrap confidence intervals. $^{**}p < 0.01$, $^{***}p < 0.001$.

5.5 Warm-up Sensitivity and Computational Efficiency

A brief supervised warm-up effectively initializes joint equivariant learning. On BCI-IV-2a (Figure 5, left), five warm-up epochs achieve near-peak accuracy, with little variation across 5–20 epochs. Extending warm-up to 40 epochs reduces performance. This broad plateau provides flexibility in the initialization budget, favoring a timely transition to joint learning over prolonged standalone classifier training.

SCoRE also achieves a favorable balance between decoding performance and computational cost (Figure 5, right). With 0.295M parameters, it exceeds REVE in macro-average balanced accuracy by 3.72 percentage points while using approximately $1/235$ as many parameters. CPU inference is also 7.8 times faster under the same benchmarking conditions. These gains show that improved decoding can accompany substantial reductions in model storage and inference overhead, supporting continuous EEG applications with limited computational resources. Detailed results and profiling settings are provided in Appendices C.3 and C.4.

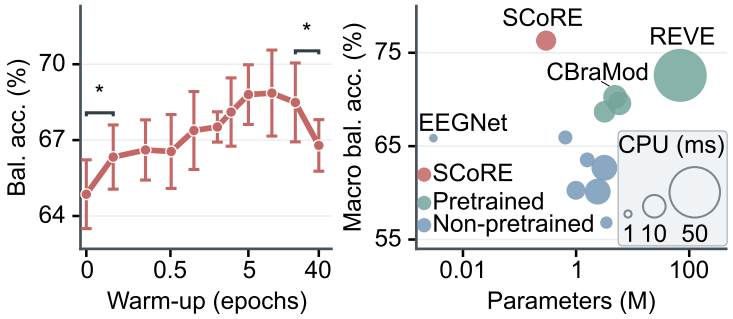


Figure 5: Warm-up sensitivity (left) and macro-average balanced accuracy versus parameter count (right). Bubble size indicates CPU inference latency.

6 Conclusion

We presented *SCoRE*, a lightweight EEG model that learns input and output reflections to capture intrinsic interregional coupling. With 0.295M parameters and no pretraining, it achieves the highest balanced accuracy on eight of nine datasets, including three new benchmarks, highlighting the value of physiological structural priors for efficient, interpretable cross-subject EEG decoding.

Broader implications. *SCoRE* may offer a way to study shared and task-dependent relationships in brain activity. Testing reflection stability across populations and recording configurations could help assess their physiological relevance and generalizability.

Limitations and future work. The gains depend on task-relevant interregional coupling. When this structure is weak, the reflection prior offers limited benefit, as illustrated by TUAB, where fusion attenuates shared class evidence and performance is comparable to strong spatiotemporal models (Appendix C.7). Beyond the current single-pair formulation, we will investigate aggregating multiple learnable (R, Q) pairs, exploiting the lightweight architecture to capture the diversity of reflection-equivariant relations in brain activity.

AI USE STATEMENT

Generative AI tools were used for language polishing to improve the clarity and readability of the manuscript. The authors reviewed the revised text and take full responsibility for the final content.

ETHICS STATEMENT

Human participant data collection for this study was conducted under institutional ethics approval. To preserve anonymity during peer review, the ethics approval documentation will be made publicly available upon publication. Participant information will be released in de-identified form to protect privacy.

REPRODUCIBILITY STATEMENT

The main text and Appendices B and C describe the model, preprocessing procedures, dataset splits, training configurations, and evaluation protocols. We provide the source code, including preprocessing and evaluation scripts and experimental configurations (Anonymous GitHub), along with the three newly collected datasets, HandMI, SSVEP-EEG, and AesthEEG (Figshare), to facilitate reproduction and independent verification.

REFERENCES

Diego Alvarez-Estevez and Roselyne M. Rijsman. Inter-database validation of a deep learning approach for automatic sleep scoring. *PLOS ONE*, 16(8):e0256111, 08 2021. doi: 10.1371/journal.pone.0256111. URL https://doi.org/10.1371/journal.pone.0256111.

Benjamin Baird, Anna Castelnovo, Brady A. Riedner, Antoine Lutz, Fabio Ferrarelli, Melanie Boly, Richard J. Davidson, and Giulio Tononi. Human rapid eye movement sleep shows local increases in low-frequency oscillations and global decreases in high-frequency oscillations compared to resting wakefulness. *eNeuro*, 5(4), 2018. doi: 10.1523/ENEURO.0293-18.2018. URL https://www.eneuro.org/content/5/4/ENEURO.0293-18.2018.

A. Biasiucci, R. Leeb, I. Iturrate, S. Perdikis, A. Al-Khodairy, T. Corbet, A. Schnider, T. Schmidlin, H. Zhang, M. Bassolino, D. Viceic, P. Vuadens, A. G. Guggisberg, and J. d. R. Millán. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nature Communications*, 9(1):2421, Jun 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04673-z. URL https://doi.org/10.1038/s41467-018-04673-z.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Luigi De Gennaro, Michele Ferrara, Giuseppe Curcio, and Riccardo Cristiani. Antero-posterior eeg changes during the wakefulness–sleep transition. *Clinical Neurophysiology*, 112(10):1901–1911, 2001. ISSN 1388-2457. doi: https://doi.org/10.1016/S1388-2457(01)00649-6. URL https://www.sciencedirect.com/science/article/pii/S1388245701006496.

Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2238–2250, 2023. doi: 10.1109/TAFFC.2022.3169001.

H. Henrik Ehrsson, Stefan Geyer, and Eiichi Naito. Imagery of voluntary movement of fingers, toes, and tongue activates corresponding body-part-specific motor representations. *Journal of Neurophysiology*, 90(5):3304–3316, 2003. doi: 10.1152/jn.01113.2002. URL https://doi.org/10.1152/jn.01113.2002. PMID: 14615433.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg - adapting to any setup with large-scale pretraining on 25,000 subjects.

In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 22541–22577. Curran Associates, Inc., 2025. doi: 10.52202/085713-0760. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/20a917f77773ac0fa8bea2bdd6606b66-Paper-Conference.pdf.

Michael D. Fox, Abraham Z. Snyder, Justin L. Vincent, Maurizio Corbetta, David C. Van Essen, and Marcus E. Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005. doi: 10.1073/pnas.0504136102. URL https://www.pnas.org/doi/abs/10.1073/pnas.0504136102.

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

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, Aline Herlopian, Ioannis Karakis, Jonathan J. Halford, Marcus C. Ng, Emily L. Johnson, Brian L. Appavu, Rani A. Sarkis, Gamaleldin Osman, Peter W. Kaplan, Monica B. Dhakar, Lakshman Arcot Jayagopal, Zubeda Sheikh, Olga Taraschenko, Sarah Schmitt, Hiba A. Haider, Jennifer A. Kim, Christa B. Swisher, Nicolas Gaspard, Mackenzie C. Cervenka, Andres A. Rodriguez Ruiz, Jong Woo Lee, Mohammad Tabaeizadeh, Emily J. Gilmore, Kristy Nordstrom, Ji Yeoun Yoo, Manisha G. Holmes, Susan T. Herman, Jennifer A. Williams, Jay Pathmanathan, Fábio A. Nascimento, Ziwei Fan, Samaneh Nasiri, Mouhsin M. Shafi, Sydney S. Cash, Daniel B. Hoch, Andrew J. Cole, Eric S. Rosenthal, Sahar F. Zafar, Jimeng Sun, and M. Brandon Westover. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17):e1750–e1762, 2023. doi: 10.1212/WNL.0000000000207127. URL https://www.neurology.org/doi/abs/10.1212/WNL.0000000000207127.

Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer Methods and Programs in Biomedicine*, 124:180–192, 2016. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2015.10.013. URL https://www.sciencedirect.com/science/article/pii/S0169260715002734.

Jinwoo Kim, Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 18582–18612. Curran Associates, Inc., 2023. doi: 10.52202/075280-0816. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3b5c7c9c5c7bd77eb73d0baec7a07165-Paper-Conference.pdf.

Arndt-Lukas Klaassen and Björn Rasch. Difficulty in artificial word learning impacts targeted memory reactivation and its underlying neural signatures. *eLife*, 12:RP90930, nov 2024. ISSN 2050-084X. doi: 10.7554/eLife.90930. URL https://doi.org/10.7554/eLife.90930.

Paul J. Laurienti, Jonathan H. Burdette, Mark T. Wallace, Yi-Fen Yen, Aaron S. Field, and Barry E. Stein. Deactivation of sensory-specific cortex by cross-modal stimuli. *Journal of Cognitive Neuroscience*, 14(3):420–429, 2002. doi: 10.1162/089892902317361930. URL https://pubmed.ncbi.nlm.nih.gov/11970801/.

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, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c. URL http://dx.doi.org/10.1088/1741-2552/aace8c.

Hongli Li, Man Ding, Ronghua Zhang, and Chunbo Xiu. Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. *Biomedical Signal Processing and Control*, 72:103342, 2022. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2021.103342. URL https://www.sciencedirect.com/science/article/pii/S1746809421009393.

Qi Mao, Hongke Zhu, Wenyao Yan, Yu Zhao, Xinhong Hei, and Jing Luo. Mcl-swt: Mirror contrastive learning with sliding window transformer for subject-independent eeg recognition. *Brain Sciences*, 15(5), 2025. ISSN 2076-3425. doi: 10.3390/brainsci15050460. URL https://www.mdpi.com/2076-3425/15/5/460.

Silvia Marchesotti, Michela Bassolino, Andrea Serino, Hannes Bleuler, and Olaf Blanke. Quantifying the role of motor imagery in brain-machine interfaces. *Scientific Reports*, 6(1):24076, 2016. doi: 10.1038/srep24076. URL https://doi.org/10.1038/srep24076.

Dennis J. McFarland, L. A. Miner, Theresa M Vaughan, and Jonathan Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, 12:177–186, 03 2000. ISSN 0896-0267. doi: 10.1023/A:1023437823106. URL http://www.ncbi.nlm.nih.gov/pubmed/10791681.

Kristen A. McKiernan, Jacqueline N. Kaufman, Jane Kucera-Thompson, and Jeffrey R. Binder. A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. *Journal of Cognitive Neuroscience*, 15(3):394–408, 2003. doi: 10.1162/089892903321593117. URL https://pubmed.ncbi.nlm.nih.gov/12729491/.

Wajid Mumtaz. Mdd patients and healthy controls eeg data (new), 11 2016. URL https://figshare.com/articles/dataset/EEG_Data_New/4244171.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

Adrian M. Owen, Kathryn M. McMillan, Angela R. Laird, and Ed Bullmore. N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25(1):46–59, 2005. doi: 10.1002/hbm.20131. URL https://pubmed.ncbi.nlm.nih.gov/15846822/.

Wei Yan Peh, Yuanyuan Yao, and Justin Dauwels. Transformer convolutional neural networks for automated artifact detection in scalp eeg. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3599–3602, 2022. doi: 10.1109/EMBC48229.2022.9871916.

Gert Pfurtscheller and Christa Neuper. Motor imagery activates primary sensorimotor area in humans. *Neuroscience Letters*, 239(2-3):65–68, 1997. ISSN 0304-3940. doi: https://doi.org/10.1016/S0304-3940(97)00889-6. URL https://www.sciencedirect.com/science/article/pii/S0304394097008896.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Edward J. Smith, Ishan Misra, Aditya Grover, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *CoRR*, abs/2110.03336, 2021. URL https://arxiv.org/abs/2110.03336.

Brady A. Riedner, Vladyslav V. Vyazovskiy, Reto Huber, Marcello Massimini, Steve Esser, Michael Murphy, and Giulio Tononi. Sleep homeostasis and cortical synchronization: Iii. a high-density eeg study of sleep slow waves in humans. *Sleep*, 30(12):1643–1657, 12 2007. ISSN 0161-8105. doi: 10.1093/sleep/30.12.1643. URL https://doi.org/10.1093/sleep/30.12.1643.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, aug 2019. doi: 10.1088/1741-2552/ab260c. URL https://doi.org/10.1088/1741-2552/ab260c.

Gerwin Schalk. EEG Motor Movement/Imagery Dataset. *PhysioNet*, September 2009. doi: 10.13026/C28G6P. URL https://doi.org/10.13026/C28G6P. Version 1.0.0.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for eeg decoding, 2021. URL https://arxiv.org/abs/2106.11170.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot Mueller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the bci competition iv. *Frontiers in Neuroscience*, 6, 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00055. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2012.00055.

Chao Wang, Sridhar Ramakrishnan, Srinivas Laxminarayan, Andrey Dovzhenok, J David Cashmere, Anne Germain, and Jaques Reifman. An attempt to identify reproducible high-density eeg markers of ptsd during sleep. *Sleep*, 43(1):zsz207, 01 2020. ISSN 0161-8105. doi: 10.1093/sleep/zsz207. URL https://doi.org/10.1093/sleep/zsz207.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 75310–75346, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bbbd6d915cb90be21c1254a82d45cedd-Paper-Conference.pdf.

Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Characterizing_and_Avoiding_Negative_Transfer_CVPR_2019_paper.html.

Dongrui Wu, Yifan Xu, and Bao-Liang Lu. Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):4–19, 2022. doi: 10.1109/TCDS.2020.3007453.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023a. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b30f3e2dd9cb53bbf2024402d02295-Paper-Conference.pdf.

Chaoqi Yang, Danica Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging, 2023b. URL https://arxiv.org/abs/2110.15278.

Huijun Yue, Zhuqi Chen, Wenbin Guo, Lin Sun, Yidan Dai, Yiming Wang, Wenjun Ma, Xiaomao Fan, Weiping Wen, and Wenbin Lei. Research and application of deep learning-based sleep staging: Data, modeling, validation, and clinical practice. *Sleep Medicine Reviews*, 74:101897, 2024. ISSN 1087-0792. doi: https://doi.org/10.1016/j.smrv.2024.101897. URL https://www.sciencedirect.com/science/article/pii/S1087079224000017.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1), 2019. ISSN 2306-5729. doi: 10.3390/data4010014. URL https://www.mdpi.com/2306-5729/4/1/14.

A New Open EEG Resources

HandMI. HandMI comprises EEG recordings of left- and right-hand motor imagery from 12 participants (6 female and 6 male; age 17.8 ± 0.4 years, mean $\pm$ SD). Each participant completed two blocks of 20 trials, with 20 trials per imagery condition and 480 trials overall. Trials comprised baseline, cue, imagery, and rest phases, including a 4 s imagery interval. EEG was acquired using a 32-channel NeuSen W wireless system (Neuracle) at 500 Hz with CPz as the reference. The resource includes continuous recordings and trial annotations specifying imagery conditions and event timing. These data support investigations of sensorimotor lateralization and interindividual variability, as well as the development of motor-imagery brain-computer interfaces.

SSVEP-EEG. SSVEP-EEG comprises EEG recordings during periodic visual stimulation from 12 participants (6 female and 6 male; age 19.7 ± 1.1 years, mean $\pm$ SD). Each participant completed two blocks of 32 trials, with eight trials per block at each stimulation frequency of 5, 7.5, 12, and 15 Hz. Trials consisted of a 2 s baseline, a 1 s cue, 4 s of visual stimulation, and a 1 s rest interval. EEG was acquired using a 32-channel NeuSen W wireless system (Neuracle) at 500 Hz with CPz as the reference. Continuous recordings and event annotations retain both stimulation and rest periods, supporting research on frequency-dependent visual responses, discrimination between stimulation and rest, and SSVEP-based brain-computer interfaces.

AesthEEG. AesthEEG integrates EEG recordings, image stimuli, aesthetic ratings, and participant questionnaires for research on aesthetic preference and visual perception. The EEG cohort comprises 60 participants aged 18–25 years, with 30 female and 30 male participants. Each participant evaluated 297 face images followed by 249 landscape images in two blocks. Images were preceded by a fixation interval of 0.8–1.2 s and displayed for 2 s. Participants subsequently provided a seven-point aesthetic rating, followed by a 0.5 s blank interval. EEG was acquired using a 32-channel NeuSen W wireless system (Neuracle) at 500 Hz with CPz as the reference. The recordings are accompanied by image identities, presentation timings, and individual ratings.

The release additionally includes the image library, image ratings from 1,300 individuals, and anonymized demographic and questionnaire data covering age, educational attainment, field of study, occupation, artistic training, AI use, and short-form video use, among other characteristics. The same questionnaire measures are provided for the EEG cohort. Together, these resources support investigations of how aesthetic judgments relate to neural responses, individual characteristics, and digital-media experience, providing a community resource for studying aesthetic preference and perception across faces and landscapes.

All three datasets and accompanying resources are available via Anonymous GitHub and Figshare.

B Dataset Details and Supplementary Results

B.1 Dataset Overview

Table 3 summarizes the nine datasets used in our evaluation, spanning six EEG decoding tasks and comprising six public benchmarks and three newly collected datasets.

Table 3: Overview of downstream EEG tasks and datasets. Duration denotes segment length, and samples denotes the total number of EEG segments.

Task	Dataset	Channels	Duration	Samples	Rate	Classes
Motor imagery	BCI-IV-2a	22	4 s	5,184	250 Hz	4
	PhysioNet-MI	64	4 s	9,837	160 Hz	4
	HandMI	32	2 s	1,440	500 Hz	2
Sleep staging	ISRUC	6	30 s	89,240	200 Hz	5
	HMC	4	30 s	137,243	256 Hz	5
Stress detection	MAT	20	5 s	1,707	500 Hz	2
Depression detection	Mumtaz	19	5 s	7,143	256 Hz	2
SSVEP classification	SSVEP-EEG	32	2 s	4,584	500 Hz	5
Aesthetic judgment	AesthEEG	32	3 s	25,970	500 Hz	4

B.2 MOTOR IMAGERY

BCIC-IV-2a. BCIC-IV-2a (Tangermann et al., 2012) provides two recording sessions for each of nine participants, with 22 EEG channels sampled at 250 Hz.

Table 4: BCIC-IV-2a results (mean ± SD; N as specified in Table 13). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	44.82±0.94	0.2693±0.0121	42.26±1.08
EEGConformer	46.96±1.06	0.2924±0.0141	45.33±1.28
SPaRCNet	46.35±1.17	0.2847±0.0147	44.32±1.26
ContraWR	46.78±1.25	0.2905±0.0160	44.13±1.42
CNN-Transformer	46.00±1.08	0.2800±0.0148	44.60±1.14
FFCL	44.70±1.43	0.2627±0.0176	42.38±1.39
ST-Transformer	45.75±1.45	0.2733±0.0198	44.71±1.42
BIOT	47.48±0.93	0.2997±0.0139	46.07±1.25
LaBraM-Base	48.69±0.85	0.3159±0.0154	47.58±1.03
CBraMod	51.38±0.66	0.3518±0.0094	49.84±0.85
REVE	<u>63.96</u>±0.95	<u>0.5194</u>±0.0126	<u>63.39</u>±1.10
SCoRE	**68.49*****±1.56	**0.5799*****±0.0208	**68.18*****±1.43

Each session contains 288 trials, equally distributed across left-hand, right-hand, both-feet, and tongue imagery. Following REVE (El Ouahidi et al., 2025), we retain the interval from 2 to 6 s after trial onset and apply a 0.5–99.5 Hz bandpass filter. Signals are resampled to 200 Hz, consistent with the sampling rate used in BIOT (Yang et al., 2023a), and Euclidean alignment is applied separately to each participant's recording session. This yields 5,184 segments of 4 s each. Participants 1–5, 6–7, and 8–9 form the training, validation, and test sets, respectively, with both sessions of each participant assigned to the same partition. Table 4 presents the comparative results.

PhysioNet-MI. PhysioNet-MI (Schalk, 2009) contains 64-channel EEG recordings from 109 participants, sampled at 160 Hz, covering imagery of the left fist, right fist, both fists, and both feet. Following REVE (El Ouahidi et al., 2025), we retain all participants and extract 4 s segments, yielding 9,837 samples. To accommodate differences in recording rates, signals are low-pass filtered at 64 Hz and converted to 128 Hz before resampling to 200 Hz, consistent with the sampling rate used in BIOT (Yang et al., 2023a). Participants 1–70, 71–89, and 90–109 constitute the training, validation, and test sets, respectively. Table 5 reports the comparative results.

Table 5: PhysioNet-MI results (mean ± SD; N as specified in Table 13). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	58.14±1.25	0.4468±0.0199	57.96±1.15
EEGConformer	60.49±1.04	0.4736±0.0171	60.62±0.95
SPaRCNet	59.32±1.52	0.4564±0.0234	59.37±1.47
ContraWR	58.92±1.33	0.4527±0.0248	59.18±1.16
CNN-Transformer	60.53±1.18	0.4725±0.0223	60.41±1.05
FFCL	57.26±0.92	0.4323±0.0182	57.01±0.79
ST-Transformer	60.35±0.81	0.4712±0.0199	60.53±0.75
BIOT	61.53±1.54	0.4875±0.0272	61.58±1.97
LaBraM-Base	61.73±1.22	0.4912±0.0192	61.77±1.41
CBraMod	<u>64.17</u>±0.91	<u>0.5222</u>±0.0169	<u>64.27</u>±1.00
REVE	**64.80**±1.40	**0.5306**±0.0187	**64.84**±1.70
SCoRE	63.95±0.77	0.5208±0.0103	63.00±0.69

HandMI. HandMI contains left- and right-hand motor imagery recordings from 12 participants, acquired with 32 EEG channels at 500 Hz (Appendix A). After synchronizing task events with EEG markers, signals are bandpass-filtered at 4–40 Hz and resampled to 200 Hz. Each 4 s imagery interval

is divided into three overlapping 2 s windows with a 1 s stride, yielding 1,440 labeled segments from 480 trials. Participants S001–S008, S009–S010, and S011–S012 are assigned to training, validation, and testing, respectively, yielding 960, 240, and 240 segments. All trials and windows from each participant remain within the same partition. Table 6 reports the comparative results.

Table 6: HandMI results (mean ± SD, $N = 10$). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	51.25±3.02	0.0250±0.0605	49.15±1.48
EEGConformer	65.00±1.77	0.3000±0.0354	64.61±1.94
SPaRCNet	64.58±3.12	0.2917±0.0624	63.56±3.77
ContraWR	59.17±4.58	0.1833±0.0915	54.79±8.08
CNN-Transformer	67.04±2.62	0.3408±0.0523	66.28±4.74
FFCL	<u>67.46</u>±2.49	<u>0.3492</u>±0.0498	66.23±3.70
ST-Transformer	51.13±3.41	0.0225±0.0682	45.73±7.70
BIOT	64.17±2.45	0.2833±0.0491	59.47±3.87
LaBraM-Base	65.96±0.52	0.3192±0.0104	65.31±0.93
CBraMod	67.38±2.39	0.3475±0.0478	<u>66.78</u>±2.86
REVE	62.38±2.77	0.2475±0.0554	61.35±3.67
SCoRE	**70.13***±2.77	**0.4025***±0.0554	**69.77***±3.38

B.3 Sleep Staging

ISRUC. We use Subgroup 1 of ISRUC (Khalighi et al., 2016), comprising polysomnographic recordings from 100 participants. Following REVE (El Ouahidi et al., 2025), we retain six EEG channels sampled at 200 Hz and divide the recordings into 89,240 epochs of 30 s each. Each epoch is assigned one of five AASM sleep stages: wake, N1, N2, N3, or REM. Epochs are organized into sequences of 20 consecutive epochs for sequence-to-sequence sleep staging. Participants 1–80, 81–90, and 91–100 form the training, validation, and test sets, respectively. Table 7 reports comparative results, with published baselines retaining the channel configurations documented by REVE, including possible chin-electrode inclusion (El Ouahidi et al., 2025).

Table 7: ISRUC results (mean ± SD; N as specified in Table 13). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	71.54±1.21	0.7040±0.0173	75.13±1.24
EEGConformer	74.00±1.33	0.7143±0.0162	76.34±1.51
SPaRCNet	74.87±0.75	0.7097±0.0132	76.24±0.92
ContraWR	74.02±1.26	0.7178±0.0156	76.10±1.37
CNN-Transformer	73.63±0.87	0.7129±0.0121	77.19±1.05
FFCL	72.77±1.82	0.7016±0.0291	76.14±1.97
ST-Transformer	73.81±2.05	0.7013±0.0352	76.81±1.75
BIOT	75.27±1.21	0.7192±0.0231	77.90±1.46
LaBraM-Base	76.33±1.02	0.7231±0.0182	78.10±1.33
CBraMod	<u>78.65</u>±1.10	<u>0.7442</u>±0.0152	**80.11**±0.99
REVE	78.19±0.78	**0.7500**±0.0156	<u>80.05</u>±1.35
SCoRE	**79.25**±0.37	0.7278±0.0047	78.60±0.47

HMC. The Haaglanden Medisch Centrum (HMC) Sleep Staging Database (Alvarez-Estevez & Rijsman, 2021) comprises 151 overnight polysomnographic recordings acquired at 256 Hz, with manual annotations for wake, N1, N2, N3, and REM sleep. Following REVE (El Ouahidi et al., 2025), we retain four EEG derivations (F4-M1, C4-M1, O2-M1, and C3-M2), apply a 0.1–75 Hz bandpass filter and a 50 Hz notch filter, and resample the signals to 200 Hz. The recordings are segmented into 137,243 epochs of 30 s each. Following REVE's filename ordering, the first 100 recordings are

used for training, the next 25 for validation, and the remaining 26 for testing. Table 8 reports the comparative results.

Table 8: HMC results (mean ± SD, $N = 10$). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	68.56±0.64	0.5050±0.0092	54.93±1.27
EEGConformer	72.10±1.02	0.5742±0.0153	63.84±1.88
SPaRCNet	47.56±11.09	0.3147±0.1315	41.08±13.10
ContraWR	42.42±5.41	0.2340±0.0554	29.87±2.88
CNN-Transformer	65.73±1.41	0.5961±0.0105	68.96±0.65
FFCL	44.27±7.02	0.2542±0.0654	29.02±4.85
ST-Transformer	25.59±1.41	0.0503±0.0183	14.28±1.22
BIOT	68.62±0.41	0.6295±0.0113	70.91±1.47
LaBraM-Base	72.86±1.01	0.6812±0.0073	75.54±0.24
CBraMod	73.35±0.47	0.6094±0.0134	68.29±1.62
REVE	<u>74.01</u>±0.75	<u>0.6982</u>±0.0078	<u>76.38</u>±0.74
SCoRE	**79.79*****±0.30	**0.7592*****±0.0066	**81.72*****±0.52

B.4 Stress Detection

MAT. MentalArithmetic (MAT) (Zyma et al., 2019) contains EEG recordings from 36 participants under mental arithmetic and resting conditions, corresponding to the stress and non-stress classes. Signals were acquired using 20 electrodes in the 10–20 system at 500 Hz. Following REVE (El Ouahidi et al., 2025), we apply a 0.5–45 Hz bandpass filter, resample the signals to 200 Hz, and extract 1,707 segments of 5 s each. Participants 1–28, 29–32, and 33–36 constitute the training, validation, and test sets, respectively. Table 9 reports the comparative results.

Table 9: MAT results (mean ± SD; N as specified in Table 13). All metrics are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	AUC-PR	AUROC
EEGNet	67.70±1.16	57.63±1.02	73.21±1.08
EEGConformer	68.05±1.23	58.29±1.34	74.24±1.28
SPaRCNet	68.79±1.07	58.25±1.93	74.18±1.32
ContraWR	66.31±0.97	57.87±1.64	73.32±0.82
CNN-Transformer	67.79±2.68	57.77±2.85	72.58±3.36
FFCL	67.98±1.42	57.86±2.66	73.30±1.98
ST-Transformer	66.31±1.73	56.72±2.59	71.32±1.74
BIOT	68.75±1.86	60.04±1.95	75.36±1.44
LaBraM-Base	69.09±1.25	59.99±1.55	77.21±0.93
CBraMod	72.56±1.32	62.67±0.99	79.05±0.73
REVE	<u>76.60</u>±3.55	<u>74.70</u>±8.07	<u>84.50</u>±5.14
SCoRE	**78.47**±2.62	**74.75**±2.77	**84.53**±5.71

B.5 Depression Detection

Mumtaz. Mumtaz (Mumtaz, 2016) comprises EEG recordings from 34 participants with major depressive disorder (MDD) and 30 healthy controls, acquired using 21 electrodes in the 10–20 system at 256 Hz. Following REVE (El Ouahidi et al., 2025), we retain the eyes-open and eyes-closed recordings, apply a 0.3–75 Hz bandpass filter and a 50 Hz notch filter, and resample the signals to 200 Hz. The recordings are divided into 7,143 segments of 5 s each. The training set includes 24 participants with MDD and 19 controls; the validation set includes 5 participants with MDD and 4 controls; and the test set includes 5 participants from each group. All recordings from each participant remain within a single partition. Table 10 reports the comparative results.

Table 10: Mumtaz results (mean ± SD; N as specified in Table 13). All metrics are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	AUC-PR	AUROC
EEGNet	92.32±1.04	96.26±0.95	96.39±0.93
EEGConformer	93.08±1.17	96.84±1.05	97.02±1.01
SPaRCNet	93.16±0.95	97.54±0.65	97.81±0.83
ContraWR	91.95±1.15	95.89±1.02	96.21±0.92
CNN-Transformer	93.05±0.68	97.57±0.74	97.42±0.59
FFCL	93.14±0.38	97.17±0.21	97.53±0.33
ST-Transformer	91.35±1.03	95.78±0.86	95.94±0.59
BIOT	93.58±0.52	97.36±0.34	97.58±0.42
LaBraM-Base	94.09±0.79	97.98±0.93	97.82±0.57
CBraMod	95.60±0.56	99.23±0.32	99.21±0.25
REVE	<u>96.44</u>±0.97	<u>99.61</u>±0.13	<u>99.57</u>±0.15
SCoRE	**97.44***±0.56	**99.66**±0.14	**99.60**±0.15

B.6 Aesthetic Judgment

AesthEEG. AesthEEG contains 32-channel EEG recordings acquired at 500 Hz from 60 participants evaluating faces and landscapes (Appendix A). Ratings of 1–3 and 5–7 on the seven-point scale define negative and positive aesthetic judgments, respectively; trials with neutral or missing ratings are excluded. Image category and aesthetic judgment jointly define four classes. Signals are notch-filtered at 50 Hz, bandpass-filtered at 0.5–45 Hz, and resampled to 250 Hz. Epochs span -0.5 to 2.5 s relative to image onset, with baseline correction over -0.5 to 0 s. The resulting benchmark contains 25,970 segments. We use a fixed subject-disjoint split of 40 participants for training, 10 for validation, and 10 for testing. Table 11 reports the comparative results.

Table 11: AesthEEG results (mean ± SD, $N = 10$). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	65.54±0.78	0.5412±0.0081	65.94±0.49
EEGConformer	66.13±1.36	0.5371±0.0216	65.40±1.78
SPaRCNet	48.50±5.92	0.3108±0.0651	47.96±5.52
ContraWR	65.22±1.05	0.5335±0.0115	65.28±0.85
CNN-Transformer	35.80±1.93	0.1490±0.0321	33.32±3.08
FFCL	44.47±0.44	0.2581±0.0073	42.99±2.06
ST-Transformer	51.95±1.00	0.3532±0.0089	51.16±1.12
BIOT	<u>66.33</u>±0.79	<u>0.5474</u>±0.0099	66.19±0.66
LaBraM-Base	65.61±0.75	0.5468±0.0019	<u>66.20</u>±0.14
CBraMod	55.47±0.34	0.3961±0.0059	54.10±0.59
REVE	63.79±0.43	0.5187±0.0040	63.82±0.38
SCoRE	**68.24*****±0.18	**0.5751*****±0.0025	**68.43*****±0.24

B.7 SSVEP Classification

SSVEP-EEG. SSVEP-EEG contains 32-channel EEG recordings acquired at 500 Hz from 12 participants, with five classes corresponding to stimulation at 5, 7.5, 12, and 15 Hz and rest (Appendix A). After synchronizing task events with EEG markers, signals are bandpass-filtered at 3–45 Hz and resampled to 200 Hz. We extract overlapping 2 s windows from the combined cue and stimulation intervals for the frequency classes. Rest windows are extracted from baseline and post-stimulation intervals, with adjacent intervals merged within each block. This yields 4,584 labeled segments. Participants S001–S008, S009–S010, and S011–S012 constitute the training, validation, and test sets, respectively, yielding 3,056, 764, and 764 segments. Table 12 reports the comparative results.

Table 12: SSVEP-EEG results (mean ± SD, $N = 10$). Balanced accuracy and weighted F1 are in %. Bold/underlined: best/second-best means; significance as in Table 1.

Method	Balanced Accuracy	Cohen's κ	Weighted F1
EEGNet	72.85±2.84	0.6537±0.0289	70.24±3.02
EEGConformer	47.58±7.70	0.3785±0.0804	50.47±5.80
SPaRCNet	39.34±4.13	0.2830±0.0510	43.62±3.79
ContraWR	67.01±1.28	0.5706±0.0134	66.02±1.31
CNN-Transformer	54.53±5.50	0.4518±0.0676	55.68±6.96
FFCL	49.10±5.76	0.3943±0.0571	51.33±6.60
ST-Transformer	45.07±4.88	0.3655±0.0542	47.95±5.62
BIOT	72.43±2.72	0.6447±0.0311	70.46±2.90
LaBraM-Base	71.92±0.30	0.6551±0.0035	71.89±0.95
CBraMod	<u>73.97</u>±0.21	0.6863±0.0072	<u>74.94</u>±0.81
REVE	73.08±1.71	<u>0.7003</u>±0.0156	73.84±1.30
SCoRE	**80.91*****±1.29	**0.7582*****±0.0164	**80.66*****±1.47

B.8 Result Provenance and Evaluation Protocols

For the six public benchmarks and TUAB, published baseline results are adopted from REVE (El Ouahidi et al., 2025), and *SCoRE* is evaluated directly on its released preprocessed data files using the corresponding train/validation/test partitions. REVE builds on the evaluation protocols of BIOT (Yang et al., 2023a), LaBraM (Jiang et al., 2024), and CBraMod (Wang et al., 2025), and serves as the most recent protocol reference among these studies. On HMC, EEGNet, EEGConformer, and CBraMod are additionally evaluated in this study over ten independent runs. For HandMI, SSVEP-EEG, and AesthEEG, all methods are evaluated in this study using identical preprocessed data, fixed subject-disjoint partitions, and a common evaluation protocol. Table 13 summarizes result sources, method-specific run counts, and evaluation settings.

Table 13: Result provenance, run counts, and evaluation protocols. N denotes the number of independent runs.

Setting and datasets	**Published results**	**This study**	**Evaluation protocol**
Published baselines BCI-IV-2a, PhysioNet-MI, ISRUC, MAT, Mumtaz, TUAB	REVE ($N = 10$) All other baselines ($N = 5$)	*SCoRE* ($N = 10$)	REVE-released preprocessed files and original benchmark splits
Partially rerun HMC	SPaRCNet, ContraWR, CNN-Transformer, FFCL, ST-Transformer, BIOT, LaBraM-Base, REVE ($N = 10$)	EEGNet, EEGConformer, CBraMod, *SCoRE* ($N = 10$)	REVE-released preprocessed files and original benchmark splits
Fully rerun HandMI, SSVEP-EEG, AesthEEG	None	All eleven baselines and *SCoRE* ($N = 10$)	Identical preprocessing and evaluation across methods with fixed subject-disjoint splits

B.9 Statistical Significance of Main Results

Table 14 compares the methods with the highest (Rank 1) and second-highest (Rank 2) mean balanced accuracy on each dataset. Two-sided Welch's t-tests are computed from the reported means, standard deviations, and method-specific run counts in Table 13. N_1 and N_2 denote the run counts for Rank 1 and Rank 2, respectively. Δ denotes the balanced accuracy of Rank 1 minus Rank 2 in percentage points, and df denotes the degrees of freedom.

B.10 Session Statistics and Alignment

For offline cross-subject evaluation, *SCoRE* estimates alignment statistics separately for each predefined subject, session, or recording unit within its assigned data partition. Each A_s uses all pre-

Table 14: Statistical comparisons supporting Table 1. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$; n.s.: not significant.

Dataset	**Rank 1**	**Rank 2**	N_1/N_2	Δ (pp)	t	df	p	Sig.
BCI-IV-2a	*SCoRE*	REVE	10/10	+4.53	7.843	14.87	1.16×10^{-6}	$***$
PhysioNet-MI	REVE	CBraMod	10/5	+0.63	1.048	11.75	0.3159	n.s.
ISRUC	*SCoRE*	CBraMod	10/5	+0.60	1.187	4.46	0.2947	n.s.
HMC	*SCoRE*	REVE	10/10	+5.78	22.628	11.81	4.33×10^{-11}	$***$
MAT	*SCoRE*	REVE	10/10	+1.87	1.340	16.56	0.1983	n.s.
Mumtaz	*SCoRE*	REVE	10/10	+1.00	2.823	14.40	0.0132	$*$
AesthEEG	*SCoRE*	BIOT	10/10	+1.91	7.454	9.93	2.26×10^{-5}	$***$
HandMI	*SCoRE*	FFCL	10/10	+2.67	2.267	17.80	0.0361	$*$
SSVEP-EEG	*SCoRE*	CBraMod	10/10	+6.94	16.792	9.48	2.28×10^{-8}	$***$

processed EEG available for that unit and remains fixed during training and inference. The same label-free estimation rule applies to training, validation, and test units. Zero-shot evaluation uses the target unit's unlabeled EEG statistics without target-subject labels.

We retain the preprocessing specified for each dataset. BCI-IV-2a follows the session-wise EA included in the REVE protocol (El Ouahidi et al., 2025), with the resulting operators reused for reflection transport. For the other datasets, default EA in equation 1 follows their specified preprocessing, including the release-specific amplitude scaling for PhysioNet-MI and Mumtaz. HandMI and SSVEP-EEG estimate covariance before overlapping-window extraction, using complete imagery trials and disjoint stimulation/rest intervals, respectively. Baseline results follow the source protocols and data partitions documented in Appendix B.8.

C Supplementary Experiments

C.1 Subject-Dependent Evaluation

We evaluate subject-dependent performance on BCI-IV-2a, PhysioNet-MI, HMC, MAT, and Mumtaz. Within each recording, EEG segments are ordered chronologically and divided into training, validation, and test partitions with proportions of 70%, 15%, and 15%, respectively. The corresponding partitions are pooled across participants, allowing models to learn from and generalize to temporally held-out data from the same individuals. All methods use identical partitions.

We train *SCoRE*, EEGNet, and CNN-Transformer from scratch and fine-tune all parameters of CBraMod and REVE from their released pretrained weights. Checkpoints are selected by validation balanced accuracy. Test balanced accuracy is computed over the pooled test segments, and Table 15 reports the resulting means and standard deviations.

Table 15: Subject-dependent balanced accuracy (%, mean $\pm$ SD). Macro denotes the unweighted average across datasets. Best and second-best means are bold and underlined, respectively.

Method	BCI-IV-2a	PhysioNetMI	HMC	MAT	Mumtaz	Macro
EEGNet	66.04±1.85	63.00±0.46	68.47±0.69	<u>84.55</u>±0.11	96.47±1.42	75.71
CNN-Transformer	63.24±1.21	60.38±0.20	73.56±0.90	69.16±1.55	93.59±0.56	71.99
CBraMod	66.12±0.19	62.69±0.50	73.70±1.08	71.31±2.10	98.34±0.32	74.43
REVE	<u>70.89</u>±1.04	<u>66.93</u>±5.59	<u>77.26</u>±0.80	79.37±5.68	**99.53**±0.09	<u>78.80</u>
SCoRE	**77.56**±1.14	**67.06**±0.59	**78.52**±0.16	**86.04**±1.03	<u>99.44</u>±0.27	**81.72**

C.2 Personalization on Unseen Subjects

We evaluate personalization on BCI-IV-2a and HMC using models trained on source subjects and selected by validation balanced accuracy. For each target subject, we reserve a support pool containing 16 samples per class and use the remaining samples as a fixed query set. The annotation budget $n \in \{0, \ldots, 16\}$ specifies the number of labeled samples per class, corresponding to $4n$ trials on

BCI-IV-2a and $5n$ epochs on HMC. Support sets are nested across budgets, and all methods share the same support and query partitions.

Each subject and annotation budget starts from the same method-specific source checkpoint with a freshly initialized optimizer. At $n > 0$, adaptation updates all model parameters using AdamW and cross-entropy loss, with batch-normalization running statistics held fixed. The setting $n = 0$ corresponds to zero-shot evaluation on the same query set. Balanced accuracy is computed separately for each target subject and summarized as mean ± SD across subjects. Tables 16 and 17 report the results at each budget.

Table 16: Personalization on BCI-IV-2a. n: labeled samples per class per subject. Values: balanced accuracy (%, mean ± SD across subjects). Δ: gain from $n = 0$ to $n = 16$ in percentage points. Best and second-best means are bold and underlined.

Method	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
EEGNet	44.87±0.56	46.88±0.83	47.46±0.28	48.10±0.29	50.10±0.56	52.39±0.40	53.47±0.29	58.40±0.42	59.18±0.16
CNN-Transformer	45.90±3.59	56.05±0.55	60.55±0.45	59.67±1.79	61.43±2.07	60.45±0.96	62.01±3.18	62.30±2.21	61.91±1.38
CBraMod	52.29±0.97	55.96±2.76	60.16±3.18	62.50±3.31	63.09±1.38	63.38±1.24	63.77±2.07	64.26±1.10	64.26±2.49
REVE	63.28±1.52	63.92±3.04	64.40±3.45	64.16±3.18	65.28±1.93	65.53±0.69	65.63±0.28	66.99±1.66	65.92±1.52
SCoRE	**69.19**±6.74	**71.97**±2.13	**72.61**±4.59	**77.05**±2.65	**77.78**±2.13	**78.08**±2.76	**80.08**±2.57	**81.25**±3.04	**81.35**±1.66

Method	$n=9$	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$	$n=16$	Δ (pp)
EEGNet	60.30±0.29	62.30±0.55	63.18±0.41	63.09±0.28	63.18±0.69	63.48±0.55	63.67±0.28	63.48±0.28	+18.61
CNN-Transformer	61.72±2.49	61.23±2.07	60.94±4.97	61.82±0.96	62.50±2.76	61.23±1.52	61.91±1.38	60.84±0.69	+14.94
CBraMod	63.67±3.87	64.16±3.45	65.04±2.76	65.33±2.35	65.23±3.04	66.21±2.21	66.02±2.49	66.50±1.79	+14.21
REVE	66.31±0.41	66.21±1.10	66.70±1.24	66.50±1.52	69.14±2.76	69.04±1.52	69.24±1.52	69.04±2.07	+5.76
SCoRE	**82.71**±1.24	**82.52**±2.82	**82.37**±2.62	**82.76**±2.04	**83.69**±1.16	**83.30**±1.88	**83.69**±0.11	**83.64**±0.65	+14.45

Table 17: Personalization on HMC. n: labeled samples per class per subject. Values: balanced accuracy (%, mean ± SD across subjects). Δ: gain from $n = 0$ to $n = 16$ in percentage points. Best and second-best means are bold and underlined.

Method	$n=0$	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
EEGNet	72.08±5.40	72.84±4.53	73.44±4.11	73.52±4.22	73.83±4.42	74.12±4.79	74.17±4.21	74.25±4.27	74.64±4.37
CNN-Transformer	65.09±4.48	68.37±5.05	71.68±5.41	73.37±4.06	76.87±3.83	80.96±4.37	80.47±3.24	80.72±2.95	80.21±2.96
CBraMod	77.49±3.76	78.23±4.85	78.86±2.85	79.62±2.87	79.84±3.12	80.52±2.89	81.06±2.69	81.09±2.07	81.95±2.50
REVE	77.29±4.98	**79.67**±3.02	80.39±2.44	80.80±2.29	80.89±2.23	81.29±2.56	81.38±2.23	81.80±2.39	81.63±2.59
SCoRE	**79.40**±3.38	79.35±4.75	**80.62**±4.72	**82.02**±4.39	**82.59**±4.28	**82.52**±4.23	**82.76**±3.60	**82.82**±3.66	**83.17**±3.50

Method	$n=9$	$n=10$	$n=11$	$n=12$	$n=13$	$n=14$	$n=15$	$n=16$	Δ (pp)
EEGNet	74.61±4.12	74.89±4.01	74.90±4.06	74.87±3.82	73.89±4.95	74.34±4.61	74.53±4.79	74.56±4.49	+2.48
CNN-Transformer	80.46±2.46	80.79±2.83	79.23±2.67	80.18±2.92	80.19±3.20	80.78±3.24	80.72±3.12	80.72±2.97	+15.63
CBraMod	82.10±2.46	82.01±2.28	82.13±2.75	82.12±2.82	81.92±2.95	82.21±2.83	82.73±3.08	82.69±2.59	+5.20
REVE	81.94±2.78	82.09±2.62	82.21±2.60	82.18±2.58	82.57±2.61	82.56±2.85	82.73±2.59	82.71±2.63	+5.42
SCoRE	**83.28**±3.50	**83.70**±3.51	**84.12**±3.39	**83.87**±3.45	**84.08**±3.43	**83.24**±3.23	**84.11**±2.45	**84.27**±2.32	+4.87

Across both datasets, *SCoRE* and the baselines generally benefit from additional labeled repetitions. Personalization gains are larger overall on BCI-IV-2a, where *SCoRE* improves by 14.45 percentage points, compared with 4.87 on HMC. This difference is consistent with individual variation in motor-imagery ability and sensorimotor rhythm modulation, which makes individual calibration particularly valuable (Marchesotti et al., 2016). Sleep staging, in comparison, draws on shared physiological hallmarks, including sleep spindles, K-complexes, and slow waves (Yue et al., 2024). On BCI-IV-2a, *SCoRE* achieves 69.19% balanced accuracy without target-subject labels, exceeding all baselines at $n = 16$, including REVE at 69.04%. Its strong zero-shot performance and further gains from limited supervision support both deployment without individual calibration and effective subsequent personalization.

C.3 Warm-up Sensitivity

We vary the duration of supervised classifier warm-up on BCI-IV-2a while keeping the architecture, data partitions, and subsequent joint-training configuration fixed. Warm-up budgets range from 0 to 1,800 minibatch steps. With 2,880 training trials and a batch size of 64, one epoch corresponds to 45 steps. After warm-up, the output contrast is initialized from training-set logit differences, followed by joint optimization of the model and learned reflections.

Table 18 shows that as few as five warm-up steps (0.11 epochs) improve balanced accuracy from 64.86% to 66.33% ($p < 0.05$). Performance subsequently rises toward a plateau across 5–20 epochs before declining at 40 epochs. This trajectory suggests that brief warm-up establishes discriminative logit structure for contrast initialization, whereas prolonged classifier-only optimization can hinder subsequent adaptation to the joint learning objective. Five warm-up epochs already achieve 68.80% balanced accuracy, within 0.06 percentage points of the highest mean. The broad plateau supports a modest initialization budget with limited sensitivity to the precise warm-up duration.

Table 18: Warm-up sensitivity on BCI-IV-2a (mean ± SD). Balanced accuracy and weighted F1 are in %. Best and second-best means are bold and underlined.

Warm-up steps	Equivalent epochs	Balanced accuracy	Weighted F1	Cohen's κ
0	0.000	64.86±1.36	66.51±1.68	0.5315±0.0181
5	0.111	66.33±1.27	66.26±0.93	0.5510±0.0169
11	0.244	66.61±1.19	66.32±1.13	0.5549±0.0159
23	0.511	66.55±1.46	66.33±1.35	0.5541±0.0195
45	1.000	67.38±1.54	67.14±1.62	0.5650±0.0205
90	2.000	67.52±0.60	67.84±0.83	0.5669±0.0080
135	3.000	68.11±1.35	**68.73**±1.31	0.5748±0.0180
225	5.000	<u>68.80</u>±1.18	68.51±1.15	<u>0.5840</u>±0.0157
450	10.000	**68.86**±1.70	<u>68.60</u>±1.76	**0.5848**±0.0227
900	20.000	68.49±1.56	68.18±1.43	0.5799±0.0208
1800	40.000	66.79±1.02	66.35±0.77	0.5572±0.0136

C.4 Computational Efficiency

We measure parameter counts, multiply-accumulate operations (MACs), and forward-pass latency using 22×800 inputs, corresponding to four-second EEG segments sampled at 200 Hz, with four output classes. All models run in FP32 evaluation mode with a batch size of one. Measurements use an Intel Xeon Platinum 8470Q CPU with four threads and an NVIDIA RTX PRO 6000 Blackwell Server Edition GPU. Inputs reside on the respective device before timing. Following at least ten untimed forward passes and 0.25 s of runtime warm-up, latency is measured over five timing blocks, with GPU synchronization around each forward pass.

Table 19: Computational efficiency for 22×800 inputs and four output classes (FP32, batch size 1). Latency is reported as mean ± SD across five timing blocks.

Method	Params (M)	MACs (G)	CPU latency (ms)	GPU latency (ms)
EEGNet	0.003	0.014	0.886±0.009	0.204±0.026
EEGConformer	0.646	0.051	2.917±0.055	1.818±0.256
SPaRCNet	0.993	0.051	5.724±0.936	4.134±0.045
ContraWR	1.576	0.061	3.310±0.026	1.533±0.125
CNN-Transformer	3.159	0.084	11.260±0.081	7.289±0.055
FFCL	2.408	0.425	11.351±0.074	8.288±0.026
ST-Transformer	3.431	0.041	2.169±0.051	1.329±0.009
BIOT	3.189	0.499	7.690±0.089	3.766±0.330
LaBraM-Base	5.821	0.552	9.400±0.201	3.256±0.134
CBraMod	4.885	0.436	9.387±0.106	3.538±0.022
REVE	69.417	6.261	51.412±0.347	6.194±0.041
SCoRE	0.295	0.276	6.575±0.344	3.003±0.111

As shown in Table 19, *SCoRE* requires 0.295M parameters and 0.276G MACs, with CPU and GPU latencies of 6.575 ms and 3.003 ms, respectively. Its parameter count and inference latencies are lower than those of all four pretrained baselines. Relative to REVE, it uses approximately $1/235$ as many parameters and provides a $7.8\times$ CPU inference speedup. These reductions in model storage

and per-segment processing time support continuous EEG analysis under limited computational resources.

C.5 Mechanistic Analysis Protocols

We analyze *SCoRE* using 1,152 BCI-IV-2a trials from subjects 6–9 in the validation and test sets, covering both recording sessions. Model outputs are computed in evaluation mode. We denote the unreflected and reflected route logits by z and z^r, respectively.

Learned reflection matrices. Figure 3(a) visualizes the learned input and output reflection matrices R and Q. Electrodes in R are ordered by left, midline, and right scalp locations. Both matrices use a shared, sign-preserving nonlinear color scale to improve the visibility of smaller coefficients.

Q-weighted spectral topographies. Motor imagery induces event-related desynchronization of sensorimotor β rhythms, providing a physiological basis for examining β-power suppression (Pfurtscheller & Neuper, 1997; McFarland et al., 2000). We estimate low-β power (13–20 Hz) in native electrode coordinates using Welch's method with 1 s Hann windows and 50% overlap. For trial n and channel c, suppression is quantified as $s_{nc} = 10\log_{10}(P_{nc}^{\text{base}}/P_{nc}^{\text{MI}})$, using the $[-2, 0]$ s baseline and $[0, 4]$ s imagery intervals relative to cue onset. The contribution of Q is quantified by $a_n = [\frac{1}{2}v_{y_n}^\top(Q - I)z_n^r]_+$, where $v_k = e_k - \frac{1}{3}\sum_{j\neq k} e_j$, e_k is the kth standard basis vector, and $[u]_+ = \max(u, 0)$. These weights express the positive Q-mediated gain in correct-class contrast. Weighted suppression maps are computed within each class, session, and run, then averaged equally across sessions and runs. Each class map is contrasted with the mean of the other three classes, spatially centered, and normalized by its maximum absolute interpolated value (Figure 3(b)).

Reflection-even and reflection-odd evidence. We center the logits using $H = I - \mathbf{1}\mathbf{1}^\top/4$ and compute the reflection-even and reflection-odd components as $z_\pm = \frac{1}{4}(I \pm Q)H(z \pm z^r)$. For hand-imagery trials, we measure the left-minus-right logit contrast in each component; for feet- and tongue-imagery trials, we measure the feet-minus-tongue contrast. Contrasts are averaged within each true class and run, then summarized as mean $\pm$ SD across ten runs. Two-sided paired t-tests compare odd contrasts between the hand classes and even contrasts between feet and tongue, with Holm correction across the two comparisons (Figure 3(c)).

Route representations. We extract the 96-dimensional embeddings preceding the shared classification head from both routes. The resulting 2,304 vectors are jointly standardized featurewise, reduced to 50 principal components, and embedded using two-dimensional t-SNE with perplexity 80, 1,200 iterations, and random seed 23. Both routes share the same projection, with 95% covariance ellipses summarizing their distributions (Figure 3(d)).

Single-trial analysis. Figure 3(e) illustrates a tongue-imagery trial misclassified by the unreflected route and correctly classified by the reflected route. Class probabilities are obtained by applying softmax separately to z, z^r, and Qz^r. The input topography depicts spatially centered $10\log_{10}$ power in the 13–30 Hz band, estimated in native electrode coordinates over the four-second imagery interval using the same Welch settings.

C.6 Quantitative Analysis of Physiological Relevance

We quantify *SCoRE*'s sensitivity to physiological EEG rhythms and relate its learned reflections to spectral biomarkers (Figure 4). Analyses use three independently trained models and HMC test recordings.

Biomarker masking (left). We examine wakeful posterior α and N2 spindle-related σ activity (De Gennaro et al., 2001), N3 slow-wave activity (Riedner et al., 2007), and REM central low-frequency activity (Baird et al., 2018). Masks cover 8–13 Hz at O2–M1 during Wake, 11–16 Hz across all channels during N2, 1–4 Hz and 0.5–1 Hz across all channels during N3, and 2–6 Hz at C4–M1/C3–M2 during REM. Target Fourier coefficients are multiplied by 0.2 before inverse transformation. Each mask is compared with 20 random contiguous bands within 0.5–30 Hz, outside the target band and matched in channel coverage, frequency-bin count, and attenuation. Flip rates quantify the fraction of initially correct predictions that become incorrect after perturbation within the corresponding sleep stage, measuring sensitivity to the targeted rhythms. Control rates are averaged over the 20 masks; both conditions are then averaged across models and equally across recordings.

Reflection–biomarker associations (right). We relate reflection magnitudes to occipital α power (8–12 Hz), central fast-spindle-band power (12–16 Hz) (Klaassen & Rasch, 2024), slow-wave activity (0.5–4 Hz) (Riedner et al., 2007), central θ power (4–8 Hz), and low-β power (15–20 Hz) (Wang et al., 2020). Occipital and central measurements use O2–M1 and C4–M1/C3–M2, respectively; slow-wave and low-β power use all channels. Fourier power is averaged over the corresponding channels and normalized by their 0.5–30 Hz power. We define reflection magnitudes as $d_R(X) = \|RX - X\|_{\mathrm{F}}/\sqrt{CT}$ and $d_Q(X) = \|QHf(X) - Hf(X)\|_2/\sqrt{K}$, where $f(X)$ denotes the final logits, $H = I - \mathbf{1}\mathbf{1}^\top/K$, and $K = 5$. Within-recording Spearman correlations span all sleep stages and are aggregated across models and recordings in Fisher-z space. For both panels, 95% confidence intervals are obtained from 10,000 recording-level bootstrap resamples. Two-sided exact sign-flip tests assess paired masking effects and Fisher-z correlations against zero, with Holm correction within each analysis.

C.7 Abnormal EEG Detection

TUAB.TUAB (Obeid & Picone, 2016) contains clinical EEG recordings labeled as normal or abnormal, acquired with 23 channels at 256 Hz. We use the preprocessed files released with REVE (El Ouahidi et al., 2025). Signals are bandpass-filtered at 0.5–99.5 Hz and resampled to 200 Hz, retaining all channels. The resulting dataset comprises 409,455 segments of 10 s each. We retain the provided training/test split and divide the training portion into 80% for training and 20% for validation, following BIOT (Yang et al., 2023a). Table 20 summarizes the comparative results, with baseline values taken from REVE.

Table 20: TUAB results (mean ± SD; N as specified in Table 13). All metrics are in %. Bold/underlined: best/second-best means.

Method	Balanced Accuracy	AUC-PR	AUROC
EEGNet	76.42±0.36	82.99±0.43	84.12±0.31
EEGConformer	77.58±0.49	84.27±0.54	84.45±0.38
SPaRCNet	78.96±0.18	84.14±0.18	86.76±0.12
ContraWR	77.46±0.41	84.21±1.04	84.56±0.74
CNN-Transformer	77.77±0.22	84.33±0.39	84.61±0.13
FFCL	78.48±0.38	84.48±0.65	85.69±0.51
ST-Transformer	79.66±0.23	85.21±0.26	87.07±0.19
BIOT	79.59±0.57	87.92±0.23	88.15±0.43
LaBraM-Base	81.40±0.19	89.65±0.16	90.22±0.09
CBraMod	<u>82.89</u>±0.22	<u>92.58</u>±0.08	<u>92.27</u>±0.11
REVE	**83.15**±0.14	**92.81**±0.09	**92.45**±0.13
SCoRE	79.64±0.43	87.94±0.38	88.01±0.31

SCoRE achieves 79.64% balanced accuracy, close to the strongest non-pretrained baseline, ST-Transformer (79.66%), and below CBraMod and REVE by 3.25 and 3.51 percentage points, respectively. A plausible explanation is a weaker correspondence between TUAB's discriminative features and the interregional structure targeted by our reflection prior. Pathological waveforms, such as epileptiform discharges (Obeid & Picone, 2016), can retain their diagnostic meaning under spatial reflection. This mismatch would diminish the gains from physiological alignment, leaving performance comparable to strong spatiotemporal classifiers.

Figure 6 shows that original and reflected inputs preserve similar separation between normal and abnormal recordings (left and middle), whereas output alignment and fusion cancel shared discriminative evidence and increase class overlap (right). With the class-swap Q in Remark 1, fusion takes half the difference between the two route contrasts. Similar discriminative evidence is therefore attenuated, illustrating how weak correspondence between task-relevant neural activity and the reflection prior limits the gains from modeling physiological coupling on TUAB.

Although TUAB exposes a limitation, this exception further supports our central argument that the additional gains of *SCoRE* derive from modeling intrinsic, task-relevant interregional coupling. The reduced gain when physiological alignment is weak reinforces this connection between neural structure and model performance. Such coordinated neural activity is widespread across motor, sensory,

and cognitive functions, providing shared physiological structure that can serve as an effective inductive bias across subjects.

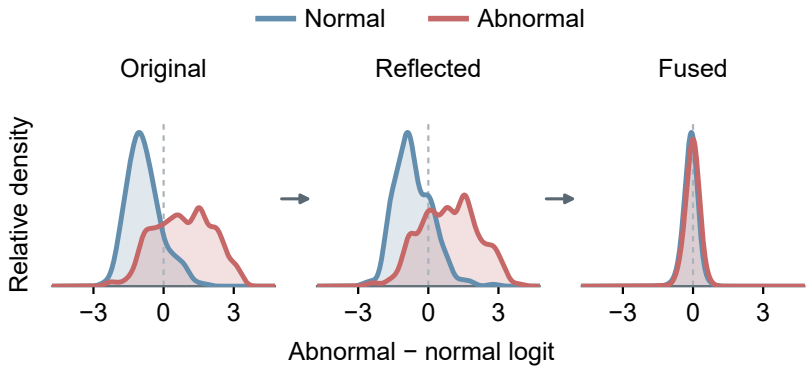


Figure 6: Evidence cancellation under reflection on TUAB. Abnormal-minus-normal logit contrasts for original (left), reflected (middle), and fused (right) outputs. Distributions summarize recording-level contrasts averaged across model seeds.

D Empirical Verification of Equivariance

We evaluate the equivariance relation in equation 10 on BCI-IV-2a. For each EEG input, we compare the logits produced after input reflection, $F(G_s\widetilde{X})$, with those obtained by applying the learned output reflection, $QF(\widetilde{X})$.

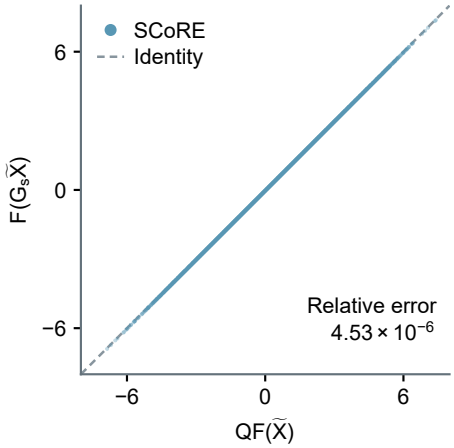


Figure 7: Numerical verification of logit equivariance on BCI-IV-2a. Each point compares corresponding coordinates of $QF(\widetilde{X})$ and independently evaluated $F(G_s\widetilde{X})$. The dashed line indicates equality.

As shown in Figure 7, the two outputs closely follow the identity line, with a relative error of 4.53×10^{-6}. This agreement provides empirical support for the equivariance guarantee, showing that the learned transformation of EEG inputs is accompanied by the corresponding transformation of class evidence.

E Related Work

Group-equivariant CNNs (Cohen & Welling, 2016) improve sample efficiency through weight sharing across group transformations. Frame Averaging (Puny et al., 2021) achieves exact equivariance

by averaging over small equivariant frames, reducing computational cost while preserving backbone expressivity. Learning Probabilistic Symmetrization (Kim et al., 2023) jointly learns input-conditioned transformation distributions and backbone representations, reducing the sampling requirements of symmetrization. These constructions demonstrate how structured transformations can support efficient learning within expressive architectures.

In EEG, MCL-SWT (Mao et al., 2025) incorporates contralateral sensorimotor organization through fixed bilateral channel exchanges and corresponding hand-label mappings. Mirror contrastive learning and aligned prediction fusion improve subject-independent motor-imagery recognition. TS-ception (Ding et al., 2023) improves emotion recognition by combining hemispheric asymmetry with global spatial features. Physiological relationships also extend beyond bilateral organization. Resting-state fMRI reveals correlated activity within distributed networks and anticorrelations between task-positive and default-mode systems (Fox et al., 2005). Visual and auditory stimulation can also activate the corresponding sensory cortex while deactivating the other modality's cortex (Laurienti et al., 2002). These findings motivate structural priors that accommodate diverse forms of intrinsic coupling across brain regions.

SCoRE extends these methodological and physiological principles to a reflection-equivariant EEG network for modeling intrinsic couplings across the brain.