RIEMANNIAN HIGH-ORDER POOLING FOR BRAIN FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electroencephalography (EEG) is a noninvasive technique for measuring brain electrical activity that supports a wide range of brain-computer interaction applications. Motivated by the breakthroughs of Large Language Models (LLMs), recent efforts have begun to explore Large EEG foundation Models trained on broad unlabeled corpora. However, most advances focus on improving the backbone while neglecting the classification head. Existing models often rely on a single class token, underutilizing the spatiotemporal structure and second-order statistics that are crucial for EEG decoding. We propose Riemannian High Order Pooling (RHOP), a plug-and-play module that injects principled Riemannian statistics into the classifier. RHOP maps each token to a quotient Gaussian jointly encoding mean and second-order information, yielding scale-invariant descriptors. Tokens are then aggregated by estimating a Riemannian Gaussian on the SPD manifold, where the Fréchet mean and covariance are embedded into an SPD descriptor. The resulting normalized vector is fused with the class token for prediction. RHOP is backbone-agnostic and integrates with modern EEG foundation models, *e.g.*, BIOT and LaBraM. Across diverse EEG benchmarks, it improves accuracy, robustness, and efficiency under full fine-tuning, linear probing, and from-scratch training settings.

1 INTRODUCTION

Electroencephalography (EEG), which records cortical electrical potentials with millisecond precision, provides dynamic insights into brain function. It has enabled advancements in seizure detection (Ahmad et al., 2022; Cherian & Kanaga, 2022), sleep staging (Aboalayon et al., 2016; Phan & Mikkelsen, 2022; Zhou et al., 2025), motor imagery (Altaheri et al., 2023; Ju & Guan, 2023; Roy et al., 2019), abnormality screening (Roy et al., 2019), emotion analysis (Suhaimi et al., 2020; Biesmans et al., 2016; Dadebayev et al., 2022), and auditory attention (Biesmans et al., 2016). However, EEG's practical deployment remains challenging due to issues like low signal-to-noise ratio, inter-subject variability, and task-dependent non-stationarity (Hine et al., 2017).

Early EEG decoding pipelines relied on traditional machine-learning methods (Lotte et al., 2007). With the rise of deep learning (LeCun et al., 2015; He et al., 2016), classic architectures such as Convolutional Neural Networks (CNNs) (Lawhern et al., 2018) and Long Short-Term Memory (LSTM) networks (Phan et al., 2019) were adapted to EEG tasks, followed by transformer backbones (Peh et al., 2022b). In parallel, geometric learning approaches have gained traction in EEG decoding tasks, most notably those leveraging Riemannian geometry. The power and spatial distribution of multi-channel EEG segments can be encoded into covariance matrices, which are symmetric positive definite (SPD) matrices. By operating on the SPD manifold, Riemannian methods exploit metrics that are robust to outliers and noise (Congedo et al., 2017), leading to broad success in practice (Pan et al., 2022; Kobler et al., 2022; Ju et al., 2024; Li et al., 2025). More recently, inspired by the rise of self-supervision and foundation models in vision and language, research has followed this paradigm Devlin et al. (2018); Radford et al. (2021); Bommasani et al. (2021); OpenAI (2023). EEG foundation models now pretrain on large unlabeled corpora with contrastive learning, masked reconstruction, or self-prediction and then transfer to diverse downstream tasks (Banville et al., 2021; Yang et al., 2023; Jiang et al., 2024; 2025).

Despite these advances, many foundation models still apply Global Average Pooling (GAP) or concatenate tokens before the final classification, which discards valuable second-order information and underuses global spatiotemporal dependencies. Intuitively, EEG features exhibit dependencies in-

herently across temporal and channel dimensions (Song et al., 2021a; Wang et al., 2025). Global Covariance Pooling (GCP) replaces GAP by summarizing activations with a covariance descriptor (Lin et al., 2015; Wang et al., 2017), partially closing this gap. However, typical GCP compresses all tokens into a single covariance matrix, thereby overlooking the intrinsic spatiotemporal structure of EEG features (Lin et al., 2015; Li et al., 2017b; 2018). This raises a central question for applying EEG foundation models to downstream decoding: Can we design a global pooling head that is both statistics-aware and geometry-aware while respecting the underlying spatiotemporal structure?

We answer this question with Riemannian High-Order Pooling (RHOP), a plug-and-play module for EEG foundation backbones and, to our knowledge, the first geometric pooling head tailored for this setting. RHOP is motivated by two empirical properties of EEG features: significant spatiotemporal structure and pervasive scale variation across temporal segments. First, we introduce a quotient-Gaussian embedding that normalizes per-token covariances to correlation form and jointly encodes first- and second-order statistics, which removes temporal scale discrepancies while preserving dependency structure (Lovrić et al., 2000a; Thanwerdas & Pennec, 2022). Second, we aggregate information across tokens by estimating a Riemannian Gaussian on the SPD manifold and embedding it into an SPD descriptor, efficiently capturing high-order interactions (Pennec, 2006). Finally, a sparse inverse-covariance layer emphasizes partial correlations and yields a compact vector, which is fused with the classification (CLS) token for prediction (Rahman et al., 2023). RHOP is architecture-agnostic and can be attached to modern EEG backbones such as BIOT and LaBraM (Yang et al., 2023; Jiang et al., 2024). In essence, RHOP bridges EEG foundation models with Riemannian statistics by embedding token-level representations into the SPD manifolds and preserving their spatiotemporal information and high-order dependencies in a pooling head. In summary, our contributions are threefold:

- **Quotient-Gaussian embedding.** A scale-invariant embedding that transforms per-token covariances into correlation form and jointly encodes first- and second-order statistics, addressing variance differences across temporal segments.
- **Riemannian High-Order Pooling.** A geometry-aware pooling head that preserves token-level spatiotemporal structure and captures high-order interactions via a Riemannian Gaussian-embedded SPD descriptor.
- **Comprehensive empirical validation.** RHOP delivers consistent gains and robust generalization across EEG decoding tasks under state-of-the-art EEG foundation backbones (e.g., BIOT and LaBraM) and three training regimes (full fine-tuning, linear probing, and train scratch).

2 Related Works

SPD manifold-based EEG decoding. Methods that operate on the SPD manifold achieve strong EEG decoding by respecting covariance geometry and improving robustness across subjects (Congedo et al., 2017; Ju et al., 2025). For example, MAtt (Pan et al., 2022) builds manifold attention directly on SPD manifolds and captures spatiotemporal dependencies, outperforming deep baselines. SPDDSMBN (Kobler et al., 2022) learns domain-invariant tangent-space mappings for unsupervised adaptation with an interpretable normalization scheme. DGCCA (Ju et al., 2024) introduces geodesic correlation with an SPD latent space to align paired covariance modalities. SPDIM (Li et al., 2025) addresses source-free adaptation with conditional and label shift through an SPD-constrained parameterization. Taken together, these works show that geometry-aware learning on SPD manifolds is powerful and motivate pairing such geometric bias with large-scale pretraining.

Brain foundation models. Foundation models are large self-supervised systems trained on broad data and adapted to many tasks via fine-tuning (Bommasani et al., 2021), with BERT (Devlin et al., 2018), CLIP (Radford et al., 2021), and GPT-4 (OpenAI, 2023). This paradigm is extending to brain signals. Recent EEG studies pretrain with contrastive learning, masked reconstruction, or self-prediction to learn transferable representations (Banville et al., 2021; Kostas et al., 2021; Chien et al., 2022; Wang et al., 2023; Zhang et al., 2023; Mohammadi Foumani et al., 2024). BIOT learns generic biosignal representations for joint pretraining and cross-dataset transfer (Yang et al., 2023), LaBraM predicts masked neural tokens for general EEG features (Jiang et al., 2024), CBraMod decouples spatial and temporal modeling with a criss-cross transformer and masked reconstruction for strong cross-dataset generalization (Wang et al., 2025), and NeuroLM tokenizes EEG signals and uses an LLM with multi-channel autoregression and instruction tuning to unify diverse tasks (Jiang et al., 2025). Despite progress, many backbones compress EEG features into a single CLS token or con-

catenate tokens before the final classification, leaving the global spatiotemporal structure underused, which motivates our geometry-aware high-order pooling head.

Gaussian embedding. Gaussian embedding offers a geometric route to compare distributions. The Fisher–Rao metric formalizes information geometry (Rao, 1945), but closed-form geodesics are challenging for multivariate Gaussians. A practical remedy identifies each Gaussian with an SPD matrix by viewing the Gaussian family as a Riemannian symmetric space, which enables affine-invariant tools for learning (Lovrić et al., 2000a; Pennec, 2006). Alternative lines embed Gaussians into the Siegel domain to define distances (Calvo & Oller, 1990), or relate Gaussians to affine matrix subspaces (Gong et al., 2009). Log–Euclidean mappings linearize SPD structure and yield efficient Gaussian descriptors for vision (Arsigny et al., 2005a; Li et al., 2017a). Deep architectures have adopted these ideas, for example a global Gaussian layer that maps images to SPD features (Wang et al., 2017) and a Lie-group embedding of Riemannian Gaussians for sets of SPD matrices (Nguyen, 2021). This geometric toolbox naturally connects to second-order pooling.

Global covariance pooling. GCP replaces global average pooling by summarizing activations with a covariance descriptor (Lin et al., 2015; Wang et al., 2017). Normalization is crucial for stability and discrimination, with matrix logarithm backpropagation (Ionescu et al., 2015), matrix power normalization such as MPN-COV (Li et al., 2017b), and iterative matrix square root as in iSQRT-COV (Li et al., 2018). Further advances address conditioning and scalability and introduce graph-aware GCP (Rahman et al., 2020; Zhu et al., 2024). On the classifier side, SoT fuses CLS and token features via global cross covariance with singular-value power normalization (Xie et al., 2021), while SICE estimates sparse inverse covariance to emphasize partial correlations (Rahman et al., 2023). Despite these advances, standard GCP on EEG underuses the spatiotemporal structure of the signal by flattening all spatiotemporal tokens into a single feature dimension before covariance pooling. This limitation motivates our Riemannian high-order pooling module.

3 Preliminary

This section provides a brief overview of the foundations of SPD geometry and the concept of Fréchet Mean (FM). For detailed discussions, please refer to Pennec (2006); Arsigny et al. (2005b).

SPD manifold. The space of $n \times n$ symmetric positive-definite matrices is denoted as $\mathcal{S}_n^+ = \{P \in \mathbb{R}^{n\times n} \mid P = P^\top,\ P \succ 0\}$. It forms a Riemannian manifold when endowed with the Affine-Invariant Metric (AIM) (Pennec, 2006). For two points $P, Q \in \mathcal{S}_n^+$, AIM distance is defined as

$$d_{\mathrm{AIM}}(P, Q) = \| \log(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}}) \|_F, \tag{1}$$

where $\| \cdot \|_F$ is the Frobenius norm and $\log(\cdot)$ denotes the matrix logarithm. The associated exponential and logarithm maps are given by

$$\mathrm{Exp}_P(S) = P^{\frac{1}{2}} \exp(P^{-\frac{1}{2}} S P^{-\frac{1}{2}}) P^{\frac{1}{2}}, \quad S \in T_P\mathcal{S}_n^+, \tag{2}$$

$$\mathrm{Log}_P(Q) = P^{\frac{1}{2}} \log(P^{-\frac{1}{2}} Q P^{-\frac{1}{2}}) P^{\frac{1}{2}}, \quad Q \in \mathcal{S}_n^+, \tag{3}$$

where $\log(\cdot)$ represents the matrix exponent and $T_P\mathcal{S}_n^+$ is the tangent space at P.

Fréchet Mean. Given a set of points $\{P_i\}_{i=1}^N \subset \mathcal{M}$, the FM is the point S that minimizes the weighted sum of squared distances to all points. the FM (Karcher, 1977b) is defined as

$$\mathrm{FM}(\{P_i\}) = \arg\min_{S\in\mathcal{M}} \frac{1}{N} \sum_{i=1}^{N} d(P_i, S)^2, \tag{4}$$

where $d(P_i, S)$ is the distance between them. The FM is locally unique on general manifolds (Afsari, 2011). In the case of $(\mathcal{S}_n^+, d_{\mathrm{AIM}})$, it is globally well-defined and unique (Chakraborty et al., 2022).

4 Riemannian High-order Pooling

In this section, we introduce Riemannian High-order Pooling (RHOP), a plug-and-play geometric pooling head for EEG foundation models. The overall RHOP framework consists of three parts: quotient Gaussian embedding, Riemannian Gaussian embedding, and a sparse inverse covariance estimation (iSICE) module (Rahman et al., 2023).

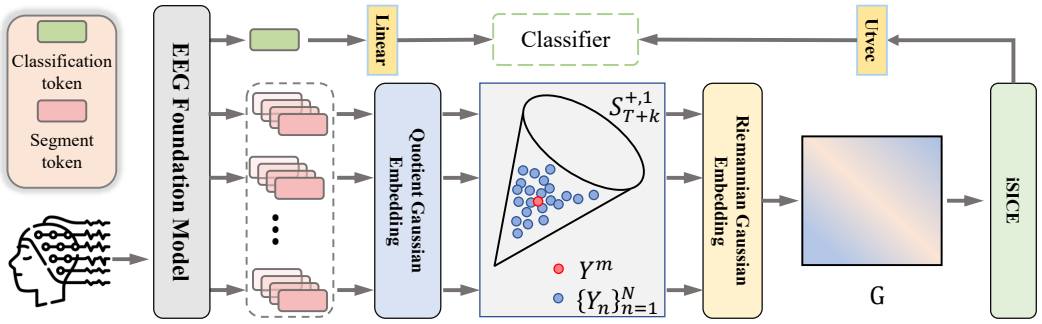

Figure 1: Overview of our RHOP framework. The EEG backbone outputs a CLS token and token-wise features. Each token is transformed into a quotient Gaussian and embedded as $Y_n \in \mathcal{S}^{+,1}_{T+k}$ on the SPD manifold. The set $\{Y_n\}$ is then aggregated into a Riemannian Gaussian, whose FM Y^m and covariance are jointly embedded into an SPD descriptor G. Finally, a iSICE + utvec block produces a sparse precision vector, which is fused with the CLS branch for classification.

4.1 Quotient Gaussian Embedding

Let $(\Sigma, \mu) \in \mathcal{N}(n)$ denote a Gaussian with covariance Σ and mean μ. Previous methods often used raw covariance matrices as EEG descriptors (Pan et al., 2022; Li et al., 2025), capturing second-order temporal dependencies but highly sensitive to scale variations. Two tokens with similar temporal dynamics may yield substantially different covariance matrices if their amplitudes differ. To address this, we introduce quotient Gaussian distributions, which normalize temporal covariances to eliminate scale dependence. This ensures that the representation focuses on correlation structures rather than raw magnitudes, thereby providing a scale-invariant descriptor of temporal dynamics.

Definition 4.1 (Quotient Gaussian Distributions). Let $(\Sigma, \mu) \in \mathcal{N}(n)$ be a Gaussian distribution. The Quotient Gaussian distribution $\mathcal{QN}(n)$ is defined as

$$\mathcal{QN}(n) \cong \mathcal{N}(n)/\mathrm{Diag}^+(n) = \left\{ [\Sigma, \mu] := \left\{ (\mathcal{D}^{-\frac{1}{2}} \Sigma \mathcal{D}^{-\frac{1}{2}}, \mu) \mid \mathcal{D} \in \mathrm{Diag}^+(n) \right\} \right\}. \tag{5}$$

For the quotient Gaussian $\mathcal{QN}(n)$, each element is an equivalence class of Gaussians $[\Sigma, \mu]$ sharing the same mean and whose covariances differ only by positive diagonal scalings. Within each class, the canonical representative $C = \mathrm{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \mathrm{diag}(\Sigma)^{-\frac{1}{2}}$, is exactly the correlation matrix, which is invariant to scaling. We thus denote each quotient element by (C, μ), where C is the normalized covariance, *i.e.*, a correlation matrix. Then, to facilitate end-to-end optimization, we embed the quotient Gaussian into SPD matrices.

Theorem 4.2 (Quotient Gaussian Embedding). *[Proof in App. I] Let $\mathcal{S}^{+,1}_{n+k}$ be the space of SPD $n \times n$ matrices with determinant 1. A quotient Gaussian $(C, \mu) \in \mathcal{QN}(n)$ can be identified as the matrix in $\mathcal{S}^{+,1}_{n+k}$:*

$$(\det C)^{-\frac{1}{n+k}} \begin{bmatrix} C + k\mu\mu^\top & \mu^{(k)} \\ \mu^{(k)\top} & I_k \end{bmatrix}, \tag{6}$$

where I_k is the $k \times k$ identity, and $\mu^{(k)}$ repeats μ across k identical columns.

This embedding maps quotient Gaussians into the SPD manifold, enabling the joint representation of mean and normalized covariance within a unified form. Following Nguyen (2021); Lovrić et al. (2000b), we endow this space with the AIM introduced in Sec. 3.

4.2 Riemannian Gaussians Embedding

Motivated by Euclidean covariance pooling, we extend first- and second-order statistics on $\mathcal{S}^+_n$ via Riemannian Gaussians. Given $\{P_i\}_{i=1}^N \in \mathcal{S}^+_n$, the FM (Fréchet, 1948) is given as

$$P^m = \mathrm{FM}(\{P_i\}_{i=1}^L) = \arg\min_{Y \in \mathcal{S}^+_n} \frac{1}{N} \sum_{i=1}^N d(Y, P_i)^2, \tag{7}$$

where $d_{\mathrm{AIM}}(\cdot, \cdot)$ is the AIM distance defined in Eq. (1). It can be computed via the Karcher flow (Karcher, 1977a), which iteratively maps P_i to $T_P\mathcal{S}^+_n$, calculates the mean in the tangent space,

Algorithm 1: RHOP over SPD manifolds

Input: EEG features $X \in \mathbb{R}^{D\times T\times N}$ from f_θ; CLS token y_0
Output: Class probabilities p
for $n \leftarrow 1$ **to** N **do**
  Compute temporal statistics (μ_n, Σ_n) across channels;
  $C_n \leftarrow \mathrm{diag}(\Sigma_n)^{-1/2}\Sigma_n\,\mathrm{diag}(\Sigma_n)^{-1/2}$;
  $Y_n \leftarrow$ quotient-Gaussian embedding via Eq. (11);
end
$Y^m \leftarrow \mathrm{FM}(\{Y_n\}_{n=1}^N)$; $Y^c \leftarrow$ covariance as in Eq. (12);
$G \leftarrow$ embed (Y^m, Y^c) using Eq. (9);
$g \leftarrow \mathrm{utvec}(\mathrm{SICE}(G; \lambda_{\mathrm{SICE}}))$;
$p \leftarrow \mathrm{softmax}\big(\mathrm{FC}([y_0; g])\big)$;
return p;

and maps back onto $\mathcal{S}_n^+$ until convergence (Moakher, 2005). Due to the computational cost of FM, we follow the previous work Brooks et al. (2019); Chen et al. (2024) to set the number of iterations to one. The implementation steps are provided in App. F.

Similarly, given FM P^m, the covariance of $\{P_i\}_{i=1}^N \in \mathcal{S}_n^+$ is defined as

$$P^c = \frac{1}{N-1}\sum_{i=1}^{N} f_v\left(\mathrm{Log}_{P^m}(P_i)\right) f_v\left(\mathrm{Log}_{P^m}(P_i)\right)^\top, \tag{8}$$

where $f_v(\cdot)$ vectorizes the lower-triangular entries of a symmetric matrix with off-diagonal terms scaled by $\sqrt{2}$ (Pennec et al., 2006). Then, A Riemannian Gaussian is parameterized by the pair $(P^m, P^c) \in \mathcal{S}_n^+ \times \mathcal{S}_{n'}^+$. As shown in Nguyen (2021), this pair lies on a product SPD manifold that forms a Lie group, and the following block-matrix construction provides a Lie-group–isomorphic embedding that preserves its algebraic and geometric structure. Let $P^c = LL^\top$ denote the Cholesky decomposition. The embedding is formulated as

$$(P^m, P^c) \;\mapsto\; \begin{bmatrix} L & 0_{n'\times k'} \\ \varphi^{k'}(P^m) & I_{k'} \end{bmatrix}, \tag{9}$$

where $I_{k'}$ is a $k'\times k'$ identity matrix and $0_{n'\times k'}$ is a zero block. Here φ is chosen as $\varphi = f_v \circ \log$, with $\log(P) = U\log(Z)U^\top$ representing the matrix logarithm by eigenvalue decomposition of P.

4.3 The Overall Framework of Riemannian High-order Pooling

In this section, we detail the whole framework of RHOP. An EEG foundation backbone f_θ (*e.g.*, BIOT or LaBraM) first extracts spatiotemporal features $X \in \mathbb{R}^{D\times T\times N}$, where D is the number of channels, T the number of temporal segments, and N the token length. For convenience we permute to $\widetilde{X} \in \mathbb{R}^{N\times T\times D}$. We then compute temporal first- and second-order statistics for each token n,

$$\mu_n = \frac{1}{D}\sum_{i=1}^{D}\widetilde{X}_{n,:,i} \in \mathbb{R}^T, \quad \Sigma_n = \frac{1}{D-1}\sum_{i=1}^{D}\big(\widetilde{X}_{n,:,i}-\mu_n\big)\big(\widetilde{X}_{n,:,i}-\mu_n\big)^\top \in \mathbb{R}^{T\times T}. \tag{10}$$

To ensure that Σ_n is an SPD matrix, we stabilize it by adding a small multiple of the identity matrix I, *i.e.*, $\Sigma_n \leftarrow \Sigma_n + \sigma I$, where we set $\sigma = 0.001$ in all experiments. Let $\mathcal{D}_n = \mathrm{diag}(\Sigma_n)$ and $C_n = \mathcal{D}_n^{-1/2}\Sigma_n\mathcal{D}_n^{-1/2}$ be the correlation matrix, which removes per-time scaling. According to Thm. 4.2, each token is then encoded as a quotient Gaussian embedding, given below

$$Y_n \;=\; (\det C_n)^{-\frac{1}{T+k}}\begin{bmatrix} C_n + k\,\mu_n\mu_n^\top & \mu_n^{(k)} \\ \mu_n^{(k)\top} & I_k \end{bmatrix} \in \mathcal{S}_{T+k}^{+,1}. \tag{11}$$

To aggregate information across tokens, we estimate a Riemannian Gaussian on the SPD manifold based on the set $\{Y_n\}_{n=1}^N$. The FM and empirical Riemannian covariance are given by

$$Y^m \;=\; \mathrm{FM}\left(\{Y_n\}_{n=1}^N\right), \quad Y^c \;=\; \frac{1}{N-1}\sum_{n=1}^{N} f_v\big(\mathrm{Log}_{Y^m}(Y_n)\big)\, f_v\big(\mathrm{Log}_{Y^m}(Y_n)\big)^\top. \tag{12}$$

The obtained product manifold (Y^m, Y^c) is embedded into an SPD matrix G using Eq. (9).

Finally, RHOP integrates global semantics with statistical structure to enhance prediction. The backbone outputs a CLS token $y_0 \in \mathbb{R}^L$ for global semantics, while the statistical branch applies iSICE (Rahman et al., 2023) to G, followed by upper-triangular extraction and vectorization , *i.e.*,

$$g = \text{utvec}\big(\text{iSICE}(G)\big). \tag{13}$$

The final prediction is obtained by concatenating y_0 and g and passing them through a linear layer with softmax activation. This design enriches the CLS token with quotient- and Riemannian-based high-order statistics, yielding a more discriminative representation.

5 Experiments

This section evaluates RHOP on four EEG benchmarks under training from scratch, full fine-tuning, and linear-head tuning settings, respectively. We compare our method against state-of-the-art foundation models and representative GCP heads.

5.1 Experiment Setup

Datasets. We evaluate RHOP on four EEG benchmarks spanning abnormal detection, event classification, motor imagery, and event-related potentials. **TUAB** (Obeid & Picone, 2016) contains 23-channel EEG at 256 Hz labeled as normal or abnormal, with 409,455 ten-second segments. **TUEV** (Obeid & Picone, 2016) includes 112,491 five-second segments from 23 channels at 256 Hz across six classes: spike & sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG). **BCIC2B** (Steyrl et al., 2016) records EEG from 10 subjects on three bipolar channels (C3, Cz, C4) at 250 Hz, with two-class motor imagery (left vs. right hand). Each subject has two no-feedback sessions (120 balanced trials each) and three feedback sessions. **PhysioP300** (Goldberger et al., 2000) is a Donchin-style 6×6 row/column speller, where rows/columns are flashed for 100 ms with 50 ms inter-stimulus intervals ($\approx$20 flashes each), and each subject spells 20 characters per run.

Backbones. We evaluate RHOP on two widely used EEG foundation models: BIOT (Yang et al., 2023) and LaBraM (Jiang et al., 2024). Since public checkpoints are only available for *LaBraM-Base*, we fine-tune this configuration in all experiments. To further assess robustness beyond foundation pretraining, we also train BIOT from scratch without any pretraining.

Preprocessing. For TUAB and TUEV, we strictly follow the original backbone preprocessing without any additional modifications. For BCIC2B, we use uniform units, apply a 0–38 Hz bandpass filter, resample the data to 200 Hz, and perform EA normalization (He & Wu, 2019) within each session. For PhysioP300, we also use uniform units, apply a 120 Hz low-pass filter, downsample to 200 Hz, and extract 2 s epochs starting at -0.7 s relative to the onset of the flicker stimulus. For more details, please refer to App. C

Baselines. We compare RHOP with both non-foundation and foundation baselines. For non-foundation baselines, we include widely used EEG models such as EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022), SPaRCNet (Jing et al., 2023), ContraWR (Yang et al., 2021), CNN-Transformer (Peh et al., 2022a), FFCL (Li et al., 2022), and ST-Transformer (Song et al., 2021a), implemented following the BIOT repository (Yang et al., 2023) unless official results are available. For foundation-model baselines, we evaluate BIOT (Yang et al., 2023) and LaBraM (Jiang et al., 2024), using their released code and checkpoints. Since only LaBraM-Base provides public weights, we fine-tune this variant in all experiments, while BIOT is tested both with and without pretraining. In addition, we compare RHOP against several representative global covariance pooling (GCP) heads, including iSICE (Rahman et al., 2023), iSQRT-COV (Li et al., 2018), and SVD-Padé (Song et al., 2021b). For fairness, given EEG inputs, all backbones output produce spatiotemporal tokens and a CLS token. GCP variants then compute covariance descriptors, which we vectorize via the upper triangular (including the diagonal) before classification with a linear layer.

Metrics. We report results using five widely adopted evaluation metrics: (1) **Balanced Accuracy**, defined as the mean recall across classes, applied to both binary and multi-class settings. (2) **AUC-PR**, the area under the precision–recall curve, for binary classification. (3) **AUROC**, the area under the receiver operating characteristic curve, also for binary classification. (4) **Cohen's Kappa**,

Table 1: Results on TUEV with different models and model complexity.

Methods	Model Size (#Params)	Time / Epoch (m)	Balanced Acc.	Cohen's Kappa	Weighted F1
SPaRCNet	0.79M	0.06	0.4161 ± 0.0262	0.4233 ± 0.0181	0.7024 ± 0.0104
ContraWR	1.6M	0.07	0.4384 ± 0.0349	0.3912 ± 0.0237	0.6893 ± 0.0136
CNN-Transformer	3.2M	0.12	0.4087 ± 0.0161	0.3815 ± 0.0134	0.6854 ± 0.0293
FFCL	2.4M	0.12	0.3979 ± 0.0104	0.3732 ± 0.0188	0.6783 ± 0.0120
ST-Transformer	3.5M	0.03	0.3984 ± 0.0228	0.3765 ± 0.0306	0.6823 ± 0.0190
BIOT (non-pretrained)					
BIOT	3.2M	0.12	0.4682 ± 0.0125	0.4482 ± 0.0285	0.7085 ± 0.0184
BIOT+iSQRT-COV	3.2M+Δ33.1K	0.81	0.4480 ± 0.0131	0.3544 ± 0.0221	0.6323 ± 0.0203
BIOT+SVD-Padé	3.2M+Δ33.1K	10.61	0.4347 ± 0.0133	0.4091 ± 0.0293	0.6906 ± 0.0137
BIOT+iSICE	3.2M+Δ33.1K	4.71	0.4630 ± 0.0143	0.4563 ± 0.0197	0.7140 ± 0.0167
BIOT+RHOP	3.2M+Δ1.3K	0.53	**0.5355 ± 0.0189**	**0.5177 ± 0.0252**	**0.7466 ± 0.0084**
BIOT (pretrained)					
BIOT	3.2M	0.12	0.5281 ± 0.0225	0.5273 ± 0.0249	0.7492 ± 0.0082
BIOT+iSQRT-COV	3.2M+Δ33.1K	0.81	0.4683 ± 0.0146	0.3563 ± 0.0218	0.6530 ± 0.0187
BIOT+SVD-Padé	3.2M+Δ33.1K	10.62	0.4372 ± 0.0176	0.4826 ± 0.0162	0.7399 ± 0.0171
BIOT+iSICE	3.2M+Δ33.1K	4.73	0.5358 ± 0.0241	0.5245 ± 0.0203	0.7534 ± 0.0094
BIOT+RHOP	3.2M+Δ1.3K	0.73	**0.5572 ± 0.0201**	**0.5460 ± 0.0212**	**0.7565 ± 0.0074**
LaBraM-Base (pretrained)					
LaBraM-Base	5.8M	1.03	**0.6409 ± 0.0065**	0.6637 ± 0.0093	0.8312 ± 0.0052
LaBraM-Base+iSQRT-COV	5.8M+Δ20.3K	3.23	0.6236 ± 0.0226	0.6147 ± 0.0234	0.8062 ± 0.0094
LaBraM-Base+SVD-Padé	5.8M+Δ20.3K	6.93	0.5605 ± 0.0217	0.5798 ± 0.0289	0.7900 ± 0.0101
LaBraM-Base+iSICE	5.8M+Δ20.3K	13.67	0.6405 ± 0.0239	0.6134 ± 0.0302	0.8182 ± 0.0091
LaBraM-Base+RHOP	5.8M+Δ0.9K	2.25	0.6380 ± 0.0056	**0.6785 ± 0.0079**	**0.8420 ± 0.0038**

Table 2: Results on TUAB with model complexity.

Methods	Model Size (#Params)	Time / Epoch (m)	Balanced Acc.	AUC-PR	AUROC
SPaRCNet	0.79M	0.43	0.7896 ± 0.0018	0.8414 ± 0.0018	0.8676 ± 0.0012
ContraWR	1.6M	0.48	0.7746 ± 0.0041	0.8421 ± 0.0104	0.8456 ± 0.0074
CNN-Transformer	3.2M	0.76	0.7777 ± 0.0022	0.8433 ± 0.0039	0.8461 ± 0.0013
FFCL	2.4M	0.86	0.7848 ± 0.0038	0.8448 ± 0.0065	0.8569 ± 0.0051
ST-Transformer	3.5M	0.24	0.7966 ± 0.0023	0.8521 ± 0.0026	0.8707 ± 0.0019
BIOT (non-pretrained)					
BIOT	3.2M	1.36	0.7925 ± 0.0035	0.8707 ± 0.0087	0.8691 ± 0.0033
BIOT+iSQRT-COV	3.2M+Δ33.1K	4.78	0.7983 ± 0.0045	0.8684 ± 0.0051	0.8659 ± 0.0055
BIOT+SVD-Padé	3.2M+Δ33.1K	50.71	0.7503 ± 0.0051	0.8274 ± 0.0058	0.8270 ± 0.0068
BIOT+iSICE	3.2M+Δ33.1K	10.47	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043
BIOT+RHOP	3.2M+Δ1.0K	3.64	**0.7993 ± 0.0031**	**0.8719 ± 0.0084**	**0.8765 ± 0.0031**
BIOT (pretrained)					
BIOT	3.2M	1.09	0.7959 ± 0.0057	0.8792 ± 0.0023	0.8815 ± 0.0043
BIOT+iSQRT-COV	3.2M+Δ33.1K	4.78	0.7819 ± 0.0044	0.8590 ± 0.0028	0.8598 ± 0.0039
BIOT+SVD-Padé	3.2M+Δ33.1K	54.38	0.7532 ± 0.0064	0.8274 ± 0.0043	0.8270 ± 0.0056
BIOT+iSICE	3.2M+Δ33.1K	10.50	0.7976 ± 0.0097	0.8617 ± 0.0046	0.8739 ± 0.0040
BIOT+RHOP	3.2M+Δ1.3K	3.69	**0.8102 ± 0.0027**	**0.8833 ± 0.0079**	**0.8864 ± 0.0033**
LaBraM-Base (pretrained)					
LaBraM-Base	5.8M	11.54	0.8140 ± 0.0019	0.8965 ± 0.0016	0.9022 ± 0.0009
LaBraM-Base+iSQRT-COV	5.8M+Δ20.3K	19.07	0.8188 ± 0.0023	0.9039 ± 0.0018	0.9060 ± 0.0012
LaBraM-Base+SVD-Padé	5.8M+Δ20.3K	31.23	0.8202 ± 0.0017	0.9062 ± 0.0014	0.9072 ± 0.0011
LaBraM-Base+iSICE	5.8M+Δ20.3K	67.77	0.8183 ± 0.0018	0.9037 ± 0.0016	0.9059 ± 0.0016
LaBraM-Base+RHOP	5.8M+Δ4.6K	21.48	**0.8244 ± 0.0012**	**0.9078 ± 0.0012**	**0.9105 ± 0.0011**

which measures agreement beyond chance by comparing observed and expected accuracies in a contingency table, is used in multi-class tasks. (5) **Weighted F1**, the weighted harmonic mean of precision and recall, used for multi-class evaluation. For monitoring, AUROC is adopted in binary classification experiments, while Cohen's Kappa is used in multi-class settings.

5.2 Evaluations

Training from scratch. On TUEV, as reported in Tab. 1, RHOP yields larger gains: balanced accuracy increases from 46.82% to 53.55%, Cohen's Kappa from 44.82% to 51.77%, and Weighted F1 from 70.85% to 74.66%, with only 0.53 minutes per epoch. On TUAB, as shown in Tab. 2, BIOT with RHOP improves balanced accuracy from 79.25% to 79.93% and AUROC from 86.91% to 87.65%, while reducing per-epoch time to 3.64 minutes compared with 10.47 minutes for iSICE and 50.71 minutes for SVD Padé. These results highlight that RHOP is particularly effective without pretraining, as it preserves spatiotemporal structure and high-order statistics.

Full fine-tuning. On TUEV, as shown in Tab. 1, RHOP reaches a higher Cohen's Kappa of 67.85% and a Weighted F1 of 84.20%, while maintaining strong balanced accuracy. On TUAB, as reported in Tab. 2, LaBraM-Base with RHOP improves balanced accuracy from 81.40% to 82.44%

Table 3: Results on BCIC2B with different models and model complexity.

Methods	Model Size (#Params)	Time / Epoch (m)	Balanced Acc.	AUC-PR	AUROC
LaBraM-Base	5.8M	0.04	0.6840 ± 0.0059	0.7405 ± 0.0087	0.7472 ± 0.0054
LaBraM-Base+iSQRT-COV	5.8M+Δ20.3K	0.38	0.6642 ± 0.0061	0.7210 ± 0.0079	0.7285 ± 0.0058
LaBraM-Base+SVD-Padé	5.8M+Δ20.3K	0.15	0.6629 ± 0.0056	0.7198 ± 0.0083	0.7267 ± 0.0061
LaBraM-Base+iSICE	5.8M+Δ20.3K	0.97	0.6871 ± 0.0058	0.7433 ± 0.0082	0.7510 ± 0.0055
LaBraM-Base+RHOP	5.8M+Δ0.5K	0.05	$\mathbf{0.6901 \pm 0.0057}$	$\mathbf{0.7485 \pm 0.0079}$	$\mathbf{0.7587 \pm 0.0059}$

Table 4: Results on PhysioP300 with model complexity.

Methods	Model Size (#Params)	Time / Epoch (m)	Balanced Acc.	AUC-PR	AUROC
LaBraM-Base	5.8M	0.05	0.6327 ± 0.019	0.6565 ± 0.024	0.6893 ± 0.027
LaBraM-Base+iSQRT-COV	5.8M+Δ20.3K	0.08	0.6170 ± 0.016	0.6560 ± 0.019	0.6813 ± 0.012
LaBraM-Base+SVD-Padé	5.8M+Δ20.3K	0.16	0.6323 ± 0.015	0.6400 ± 0.020	0.6751 ± 0.014
LaBraM-Base+iSICE	5.8M+Δ20.3K	0.41	0.6138 ± 0.018	0.6350 ± 0.022	0.6427 ± 0.017
LaBraM-Base+RHOP	5.8M+Δ0.4K	0.08	$\mathbf{0.6517 \pm 0.018}$	$\mathbf{0.6630 \pm 0.016}$	$\mathbf{0.7044 \pm 0.024}$

Table 5: Ablation study on different components, where QGE denotes Quotient Gaussian Embedding, RGE denotes Riemannian Gaussian Embedding, and CLS indicates whether concatenation with the classification head is applied.

QGE	RGE	SICE	CLS	TUAB			TUEV		
				Balanced Acc.	AUC-PR	AUROC	Balanced Acc.	Cohen's Kappa	Weighted F1
✗	✗	✗	✗	0.8140 ± 0.0019	0.8965 ± 0.0016	0.9022 ± 0.0009	$\mathbf{0.6409 \pm 0.0065}$	0.6637 ± 0.0093	0.8312 ± 0.0052
✓	✗	✗	✗	0.8175 ± 0.0018	0.9002 ± 0.0015	0.9048 ± 0.0010	0.6325 ± 0.0062	0.6669 ± 0.0090	0.8331 ± 0.0051
✓	✓	✗	✗	0.8209 ± 0.0016	0.9031 ± 0.0014	0.9069 ± 0.0010	0.6355 ± 0.0060	0.6712 ± 0.0087	0.8365 ± 0.0049
✓	✓	✓	✗	0.8227 ± 0.0014	0.9056 ± 0.0013	0.9088 ± 0.0010	0.6341 ± 0.0058	0.6749 ± 0.0084	0.8391 ± 0.0047
✓	✓	✓	✓	$\mathbf{0.8244 \pm 0.0012}$	$\mathbf{0.9078 \pm 0.0012}$	$\mathbf{0.9105 \pm 0.0011}$	0.6380 ± 0.0056	$\mathbf{0.6785 \pm 0.0079}$	$\mathbf{0.8420 \pm 0.0038}$

and pushes AUROC to 91.05%. These improvements confirm that even strong pretrained backbones benefit from RHOP, which captures dependencies overlooked by conventional heads.

Linear-head tuning. With frozen backbones, RHOP still surpasses all baselines. On BCIC2B, LaBraM-Base with RHOP reaches 69.01% balanced accuracy and 75.87% AUROC, outperforming all GCP heads, as given in Tab. 3. On PhysioP300, it improves balanced accuracy from 63.27% to 65.17% and AUROC from 68.93% to 70.44%, as presented in Tab. 4. These results demonstrate RHOP's ability to deliver discriminative representations even when backbone weights are frozen, which makes it suitable for rapid adaptation with minimal compute.

Why RHOP outperforms GCP? Classical GCP methods such as iSQRT-COV, iSICE, and SVD Padé collapse all tokens into a single covariance matrix, which discards temporal and channel hierarchy. RHOP instead normalizes tokens' covariances to correlation form, embeds both means and normalized covariances on the SPD manifold, and aggregates via a Riemannian Gaussian defined by a Fréchet mean and tangent-space covariance. This preserves scale-invariant dependencies, aligns computation with manifold geometry, and emphasizes direct relationships through a sparse inverse-covariance layer. The result is a more faithful global representation that consistently outperforms GCP, especially in challenging setups such as TUEV with training from scratch.

Efficiency. RHOP delivers accuracy gains with minimal overhead. On TUAB, LaBraM-Base with RHOP trains in 21.48 min per epoch with only 4.6K additional parameters, compared with 67.77 min for iSICE and 31.23 minutes for SVD-Padé as reported in Tab. 2. BIOT with RHOP requires just 3.69 min per epoch when pretrained, and only 0.53 min per epoch when trained from scratch on TUEV, far below iSICE and SVD Padé, as shown in Tab. 1. On BCIC2B and PhysioP300, RHOP adds less than 1K parameters and only 0.01 min per epoch, while still improving all metrics, as shown in Tabs. 3 and 4. Overall, RHOP combines strong accuracy, scale robustness, and geometric fidelity with thousand-level parameter overhead and negligible training cost, making it a practical plug-and-play head for EEG foundation backbones.

5.3 Ablation Study

Quotient Gaussian Embedding vs Gaussian Embedding. Fig. 2 compares quotient Gaussian embedding with conventional Gaussian embedding. The only difference is that conventional embedding (Nguyen, 2021) directly uses the covariance matrix without normalization. On both TUAB

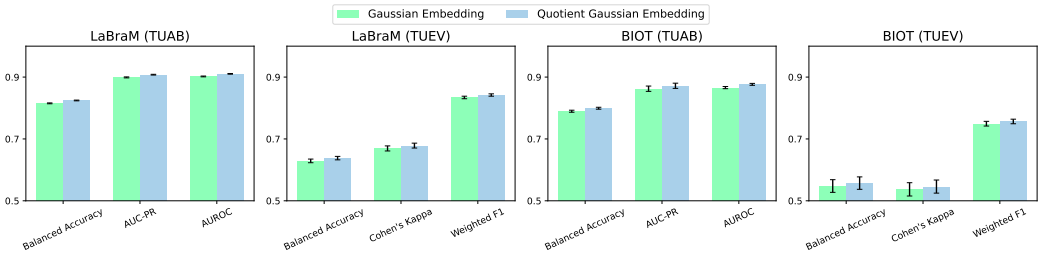

Figure 2: Quotient Gaussian embedding vs Gaussian embedding on TUEV and TUAB.

Table 6: Ablation of (k, k') on LaBraM-Base.

k	k'	TUAB			TUEV		
		Balanced Acc.	**AUC-PR**	**AUROC**	**Balanced Acc.**	**Cohen's Kappa**	**Weighted F1**
1	1	0.8065 ± 0.0019	0.8973 ± 0.0017	0.8988 ± 0.0012	0.6457 ± 0.0121	0.5772 ± 0.0203	0.7876 ± 0.0114
	2	0.8193 ± 0.0021	0.9054 ± 0.0016	0.9064 ± 0.0013	0.6339 ± 0.0184	0.6078 ± 0.0227	0.8062 ± 0.0102
	3	0.8199 ± 0.0020	0.9054 ± 0.0015	0.9060 ± 0.0012	$\mathbf{0.6726 \pm 0.0137}$	0.6456 ± 0.0181	0.8224 ± 0.0097
2	1	0.8171 ± 0.0018	0.8975 ± 0.0018	0.9013 ± 0.0012	0.6631 ± 0.0116	0.6471 ± 0.0193	0.8252 ± 0.0089
	2	$\mathbf{0.8249 \pm 0.0022}$	0.9044 ± 0.0014	0.9071 ± 0.0011	0.6459 ± 0.0128	0.6008 ± 0.0251	0.7970 ± 0.0095
	3	0.8163 ± 0.0017	0.9058 ± 0.0016	0.9057 ± 0.0011	0.6440 ± 0.0137	0.6196 ± 0.0228	0.8145 ± 0.0117
3	1	0.8183 ± 0.0019	0.9035 ± 0.0015	0.9062 ± 0.0011	0.6659 ± 0.0123	0.6317 ± 0.0273	0.8176 ± 0.0135
	2	0.8193 ± 0.0020	0.9008 ± 0.0016	0.9039 ± 0.0012	0.6372 ± 0.0161	0.6220 ± 0.0234	0.8135 ± 0.0092
	3	0.8244 ± 0.0012	$\mathbf{0.9078 \pm 0.0012}$	$\mathbf{0.9105 \pm 0.0011}$	0.6380 ± 0.0056	$\mathbf{0.6785 \pm 0.0079}$	$\mathbf{0.8420 \pm 0.0038}$

Table 7: Comparison of different GCP heads on LaBraM-Base.

Normalization	TUAB			TUEV		
	Balanced Acc.	**AUC-PR**	**AUROC**	**Balanced Acc.**	**Cohen's Kappa**	**Weighted F1**
SVD-Padé	0.8190 ± 0.0016	0.9012 ± 0.0014	0.9058 ± 0.0010	0.6368 ± 0.0071	0.6710 ± 0.0085	0.8365 ± 0.0049
iSQRT-COV	0.8181 ± 0.0017	0.9001 ± 0.0015	0.9051 ± 0.0011	0.6375 ± 0.0068	0.6692 ± 0.0091	0.8350 ± 0.0047
iSICE	$\mathbf{0.8244 \pm 0.0012}$	$\mathbf{0.9078 \pm 0.0012}$	$\mathbf{0.9105 \pm 0.0011}$	0.6380 ± 0.0056	$\mathbf{0.6785 \pm 0.0079}$	$\mathbf{0.8420 \pm 0.0038}$

and TUEV, the quotient form consistently outperforms the raw covariance representation, confirming that scale normalization is essential for robust EEG descriptors.

Component analysis. Tab. 5 evaluates the contribution of each RHOP component. QGE alone improves performance, RGE further enhances high-order modeling, and SICE highlights partial correlations. The best results are obtained when CLS fusion is added, which combines semantic and statistical cues. Overall, each component provides complementary gains, and the full RHOP yields the strongest improvements.

Embedding dimensions. The parameters k and k' control the augmentation of mean vectors and the embedding of Riemannian statistics in Eqs. (9) and (11). Tab. 6 reports their influence on TUAB and TUEV. Non-trivial choices with $k, k' > 0$ consistently outperform the baseline, and the optimal configuration varies across datasets. TUAB achieves the highest AUROC with $k = 3, k' = 3$, while TUEV obtains the best Cohen's Kappa with $k = 3, k' = 3$ and the best balanced accuracy with $k = 1, k' = 3$. Increasing k or k' expands the SPD embedding dimension and amplifies the mean-related terms, which reduces the relative influence of the covariance structure and introduces additional computational and numerical burdens. As a result, moderate settings such as $(k = 3, k' = 3)$ achieve the best trade-off between representational expressiveness and model stability.

Covariance normalization. Tab. 7 examines why iSICE is selected as the final covariance normalization method. Replacing iSICE with iSQRT-COV or SVD-Padé reduces performance, while iSICE achieves the best results on TUAB and TUEV. The reason is that iSICE imposes sparsity during normalization, which regularizes high-dimensional features. This property is particularly effective for EEG decoding, where signals are noisy and correlations are often spurious, making iSICE a more robust choice than dense normalization approaches.

Table 8: Gaussianity statistics of token features after random projection. SW-p: Shapiro–Wilk p-value; Prop.: proportion of projections with $p > 0.05$. $p > 0.05$ indicates non-rejection of Gaussianity.

(a) Gaussianity statistics of RHOP temporal-segment features.

Model	Dataset	SW-p	Prop.($p > 0.05$)	Skew	Kurt
LaBraM	TUAB	0.49	0.94	−0.13	0.19
	TUEV	0.47	0.89	−0.10	0.15
BIOT	TUAB	0.46	0.91	−0.07	0.10
	TUEV	0.44	0.85	−0.09	0.12

(b) Gaussianity statistics of GCP token-level aggregated features.

Model	Dataset	SW-p	Prop.($p > 0.05$)	Skew	Kurt
LaBraM	TUAB	0.01	0.05	0.14	18.94
	TUEV	0.01	0.03	0.25	19.54
BIOT	TUAB	0.02	0.03	0.22	19.16
	TUEV	0.01	0.00	0.14	19.97

5.4 Gaussianity Analysis

In this section, we evaluate the Gaussianity of token features extracted from LaBraM and BIOT on TUAB and TUEV, aiming to understand the statistical behavior of foundation model activations.

Testing procedure. For each dataset, we randomly sample 10 000 data samples, extract their token representations, and assess the distributional properties of the temporal feature vectors. Each token feature is projected onto multiple random one-dimensional directions, a standard approach for testing multivariate normality. For each projection, we compute the Shapiro–Wilk p-value, skewness, and excess kurtosis, and then average these quantities.

Discussion. As summarized in Tab. 8(a), the Gaussianity statistics reflect the distribution of RHOP's temporal-segment features, that is, the feature vectors obtained along the temporal dimension on which RHOP computes covariance. Along this axis, the features exhibit near-Gaussian behavior: the Shapiro–Wilk p-values are moderate, 85%–94% of projections satisfy $p > 0.05$, skewness is close to zero, and excess kurtosis remains small. These results indicate that RHOP's temporal-segment representations follow a Gaussian suitable for second-order modeling.

In contrast, the GCP features in Tab. 8(b) exhibit strong non-Gaussian behavior. Unlike RHOP, GCP computes covariance after flattening and aggregating both spatial and temporal tokens, thereby mixing heterogeneous EEG components within a single feature vector. When Gaussianity is assessed along this mixed token axis, the resulting statistics deviate substantially from a normal distribution: the Shapiro–Wilk p-values are near zero, fewer than 5% of projections satisfy $p > 0.05$, and the excess kurtosis reaches extremely high values. These results indicate heavy-tailed, asymmetric, and multimodal structure, reflecting the statistical complexity introduced by combining spatial and temporal information in a single representation.

Since RHOP evaluates covariance within coherent temporal segments—rather than across mixed token features—it operates on statistically well-behaved representations. This explains the greater stability of RHOP's covariance estimates and its superior downstream performance.

6 Conclusion

This work introduced Riemannian High-Order Pooling, a geometry-aware classification head designed to complement large EEG foundation backbones. By embedding per-token quotient Gaussians on the SPD manifold and aggregating them into a Riemannian Gaussian descriptor, RHOP preserves scale-invariant dependencies and captures high-order spatiotemporal interactions. Extensive experiments across abnormal detection, epileptic event classification, motor imagery, and event-related potentials demonstrated that RHOP consistently improves accuracy, robustness, and efficiency over global average pooling and classical covariance pooling. These gains persist across different training regimes, including learning from scratch, full fine-tuning, and linear probing, confirming that RHOP provides a principled and effective bridge between Riemannian statistics and foundation models. More broadly, this study underscores the importance of incorporating geometric inductive bias into large-scale EEG systems and points toward scalable, geometry-aware modeling of brain signals.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. All experiments were conducted on publicly available datasets with appropriate licenses, and no personally identifiable or sensitive information was involved.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure reproducibility. The datasets used are publicly available. Model architectures, training procedures, and evaluation metrics are detailed in Sections 4 and 5, with additional implementation details and hyperparameters provided in the appendix. Upon acceptance, we will release the complete source code and instructions for reproducing all experiments.

REFERENCES

K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour. Sleep stage classification using eeg signal analysis: A comprehensive survey and new investigation. *Entropy*, 18(9): 272, 2016.

Bijan Afsari. Riemannian l^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 2011.

Ijaz Ahmad, Xin Wang, Mingxing Zhu, Cheng Wang, Yao Pi, Javed Ali Khan, Siyab Khan, Oluwarotimi Williams Samuel, Shixiong Chen, and Guanglin Li. Eeg-based epileptic seizure detection via machine/deep learning approaches: A systematic review. *Computational Intelligence and Neuroscience*, 2022:9814248, 2022.

H. Altaheri, G. Muhammad, M. Alsulaiman, S. U. Amin, G. A. Altuwaijri, W. Abdul, M. A. Bencherif, and M. Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Fast and simple computations on tensors with log-euclidean metrics. Research Report RR-5584, INRIA, Sophia Antipolis, France, 2005a.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. *Fast and simple computations on tensors with log-Euclidean metrics.* PhD thesis, INRIA, 2005b.

Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering*, 18(4):046020, 2021. doi: 10.1088/1741-2552/abca18.

Wouter Biesmans, Neetha Das, Tom Francart, and Alexander Bertrand. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(5): 402–412, 2016.

Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *NeurIPS*, pp. 15463–15474, 2019.

Miquel Calvo and Josep M. Oller. A distance between multivariate normal distributions based in an embedding into the siegel group. *Journal of Multivariate Analysis*, 35(2):223–242, 1990. doi: 10.1016/0047-259X(90)90026-E.

Rudrasis Chakraborty, Jose Bouza, Jonathan H Manton, and Baba C Vemuri. Manifoldnet: A deep neural network for manifold-valued data with applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. doi: 10.1109/TPAMI.2020.3003846. URL https://doi.org/10.1109/TPAMI.2020.3003846.

Ziheng Chen, Yue Song, Yunmei Liu, and Nicu Sebe. A Lie group approach to Riemannian batch normalization. In *ICLR*, 2024.

R. Cherian and E. G. Kanaga. Theoretical and methodological analysis of eeg based seizure detection and prediction: An exhaustive review. *Journal of Neuroscience Methods*, 369:109483, 2022.

Hsiang-Yun Sherry Chien, Hanlin Goh, Christopher M. Sandino, and Joseph Y. Cheng. MAEEG: Masked auto-encoder for EEG representation learning. *arXiv preprint arXiv:2211.02625*, 2022.

Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4(3):155–174, 2017.

D. Dadebayev, W. W. Goh, and E. X. Tan. Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques. *Journal of King Saud University-Computer and Information Sciences*, 34(7):4385–4401, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, 1948.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Liyu Gong, Tianjiang Wang, and Fang Liu. Shape of gaussians as feature descriptors. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2366–2371. IEEE, 2009. doi: 10.1109/CVPR.2009.5206506.

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, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gabriel Emile Hine, Emanuele Maiorana, and Patrizio Campisi. Resting-state eeg: A study on its non-stationarity for biometric applications. In *Proceedings of the 2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2017.

Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2965–2973, 2015. doi: 10.1109/ICCV.2015.339.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv preprint arXiv:2405.18765*, 2024.

Weibang Jiang, Yansen Wang, Bao-liang Lu, and Dongsheng Li. NeuroLM: A universal multitask foundation model for bridging the gap between language and EEG signals. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Ce Ju and Cuntai Guan. Graph neural networks on spd manifolds for motor imagery classification: A perspective from the time-frequency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10955–10969, 2023.

Ce Ju, Reinmar J Kobler, Liyao Tang, Cuntai Guan, and Motoaki Kawanabe. Deep geodesic canonical correlation analysis for covariance-based neuroimaging data. In *ICLR*, 2024.

Ce Ju, Reinmar J Kobler, Antoine Collas, Motoaki Kawanabe, Cuntai Guan, and Bertrand Thirion. Spd learning for covariance-based neuroimaging analysis: Perspectives, methods, and challenges. *arXiv preprint arXiv:2504.18882*, 2025.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30:509–541, 1977a.

Hermann Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977b.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. *NeurIPS*, 35: 6219–6235, 2022.

Dimitrios Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

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. doi: 10.1016/j.bspc.2021.103342.

Peihua Li, Qilong Wang, Hui Zeng, and Lei Zhang. Local log-euclidean multivariate gaussian descriptor and its application to image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):803–817, 2017a. doi: 10.1109/TPAMI.2016.2560816.

Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2070–2078, 2017b. doi: 10.1109/ICCV.2017.228.

Peihua Li, Jiangtao Xie, Qilong Wang, and Zilin Gao. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 947–955, 2018. doi: 10.1109/CVPR.2018.00105.

Shanglin Li, Motoaki Kawanabe, and Reinmar Kobler. Spdim: Source-free unsupervised conditional and label shift adaptation in eeg. In *International Conference on Representation Learning*, volume 2025, pp. 101907–101931, 2025.

Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1449–1457, 2015. doi: 10.1109/ICCV.2015.170.

Fabien Lotte, Emmanuel Congedo, Aurélien Lécuyer, François Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 4(2):R1, 2007.

Miroslav Lovrić, Maung Min-Oo, and Ernst A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000a. doi: 10.1006/jmva.1999.1853.

Miroslav Lovrić, Maung Min-Oo, and Ernst A. Ruh. Multivariate normal distributions parametrized as a riemannian symmetric space. *Journal of Multivariate Analysis*, 74(1):36–48, 2000b.

Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005. doi: 10.1137/S0895479803436937.

Navid Mohammadi Foumani, Geoffrey Mackellar, Soheila Ghane, Saad Irtza, Nam Nguyen, and Mahsa Salehi. EEG2Rep: Enhancing self-supervised EEG representation through informative masked inputs. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, pp. 5544–5555, Barcelona, Spain, 2024. ACM. doi: 10.1145/3637528.3671600.

Xuan Son Nguyen. Geomnet: A neural network based on riemannian geometries of spd matrix space and cholesky space for 3d skeleton-based interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13379–13389, 2021.

Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10(196), 2016. doi: 10.3389/fnins.2016.00196.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A manifold attention network for EEG decoding. In *NeurIPS*, pp. 31116–31129, 2022.

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

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

Xavier Pennec. Statistical computing on manifolds for computational anatomy. Habilitation à diriger des recherches, Université Nice Sophia Antipolis, December 2006. Mémoire présenté pour l'obtention de l'habilitation à diriger des recherches.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

H. Phan and K. Mikkelsen. Automatic sleep staging of eeg signals: Recent development, challenges, and future directions. *Physiological Measurement*, 43(4):04TR01, 2022.

Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chen, and Maarten De Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(3):400–410, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Saimunur Rahman, Lei Wang, Changming Sun, and Luping Zhou. Redro: Efficiently learning large-sized SPD visual representation. In *Computer Vision – ECCV 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pp. 1–17. Springer, 2020. doi: 10.1007/978-3-030-58555-6_1.

Saimunur Rahman, Piotr Koniusz, Lei Wang, Luping Zhou, Peyman Moghadam, and Changming Sun. Learning partial correlation based deep visual representation for image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6231–6240, 2023. doi: 10.1109/CVPR52729.2023.00603.

C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

Subhrajit Roy, Isabell Kiral-Kornek, and Stefan Harrer. Chrononet: A deep recurrent neural network for abnormal eeg identification. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019*, pp. 47–56. Springer, 2019.

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

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, 2022. doi: 10.1109/TNSRE.2022.3230250.

Yue Song, Nicu Sebe, and Wei Wang. Why approximate matrix square root outperforms accurate svd in global covariance pooling? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1115–1123, 2021b.

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

Nazmi Sofian Suhaimi, James Mountstephens, Jason Teo, et al. Eeg-based emotion recognition: A state-of-the-art review of current trends and opportunities. *Computational Intelligence and Neuroscience*, 2020:9814248, 2020.

Yann Thanwerdas and Xavier Pennec. Theoretically and computationally convenient geometries on full-rank correlation matrices. *SIAM J. Matrix Anal. Appl.*, 2022.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations (ICLR), Poster*, 2023.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations (ICLR)*, 2025.

Qilong Wang, Peihua Li, and Lei Zhang. G2DeNet: Global gaussian distribution embedding network and its application to visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2730–2739, 2017.

Jiangtao Xie, Ruiren Zeng, Qilong Wang, Ziqi Zhou, and Peihua Li. Sot: Delving deeper into classification head for transformer. *arXiv preprint arXiv:2104.10935*, 2021. doi: 10.48550/arXiv.2104.10935.

Chaoqi Yang, Danica Xiao, M. Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021. URL https://arxiv.org/abs/2110.15278. Preprint.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 78240–78260, 2023.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 26304–26321, 2023.

Y. Zhou, S. Zhao, J. Wang, H. Jiang, S. Li, B. Luo, T. Li, and G. Pan. Personalized sleep staging leveraging source-free unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14529–14537, 2025.

Pengfei Zhu, Jialu Li, Zhaowei Dong, Ling Shao, et al. CCP-GNN: Competitive covariance pooling for improving graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6395–6406, 2024. doi: 10.1109/TNNLS.2024.3390249.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

We used large language models (LLMs) only for limited assistance in language polishing and for implementing minor parts of the code. All research ideas, experiment design, analysis, and conclusions were conceived and validated by the authors, who remain fully responsible for the content of this work.

B NOTATIONS

For better clarity, we summarize all the notations used in this paper in Tab. 9.

Table 9: Notations used in this paper.

Symbol	Meaning
$\mathcal{S}_n^+$	Set of $n \times n$ symmetric positive definite matrices
$d_{\text{AIM}}(P, Q)$	Affine invariant Riemannian distance on $\mathcal{S}_n^+$
$\text{Exp}_P(S)$	Exponential map at $P \in \mathcal{S}_n^+$ for tangent vector $S \in T_P\mathcal{S}_n^+$
$\text{Log}_P(Q)$	Logarithm map at $P \in \mathcal{S}_n^+$ for $Q \in \mathcal{S}_n^+$
$T_P\mathcal{S}_n^+$	Tangent space of $\mathcal{S}_n^+$ at P
$\|\cdot\|_F$	Frobenius norm
$\text{WFM}(\{w_i\}, \{P_i\})$	Weighted Fréchet mean on $\mathcal{S}_n^+$ with weights $\{w_i\}$
w_i	Nonnegative weights with $\sum_i w_i = 1$
$\mathcal{N}(n)$	Family of n-dimensional Gaussians parameterized by (Σ, μ)
Σ, μ	Covariance matrix and mean vector of a Gaussian
k, k'	Hyperparameters for quotient and Riemannian embedding dimensions

C DATASET DESCRIPTION AND PREPROCESSING

In this section, we provide additional implementation details to ensure full reproducibility of our experiments, including dataset descriptions, preprocessing steps, and training configurations that were not covered in the main paper.

C.1 THE TUAB DATASET

Description. The TUAB dataset[1] is a large clinical EEG corpus in which each recording is labeled as either normal or abnormal. The raw EEGs are 23-channel clinical recordings sampled at 256 Hz. After segmentation into non-overlapping 10-second windows, the corpus provides 409,455 samples for binary normal/abnormal classification.

Preprocessing. For experiments using the BIOT model, we adopt the preprocessing pipeline provided in BIOT (Yang et al., 2023). All recordings are processed using 16 bipolar montage channels defined under the 10–20 system. Each EEG signal is first resampled to 200 Hz and then transformed into a standardized set of 16 bipolar derivations (Yang et al., 2023), which specify fixed electrode pairs (e.g., FP1–F7, F7–T3, T3–T5, etc.). Segments shorter than 10 seconds are discarded. For experiments using LaBraM, we adopt the preprocessing protocol defined in (Jiang et al., 2024). The EEG signals are first band-pass filtered between 0.1 Hz and 75 Hz to remove low-frequency drift, followed by a 50 Hz notch filter to suppress power-line interference. All recordings are then resampled to 200 Hz. Since raw EEG values typically lie within $[-0.1, 0.1]$ mV, we normalize the amplitude by scaling the unit to 0.1 mV such that the resulting signals fall approximately within $[-1, 1]$. Each valid 10-second window inherits the normal/abnormal label of its source recording and is stored as an individual sample for downstream training, validation, and evaluation.

Experimental Configuration. We adopt a subject-independent evaluation protocol. Since the official subject-wise split already provides separate training and test partitions, we further divide the training portion into an 80%–20% train–validation split, following prior work (Yang et al., 2023; Jiang et al., 2024). The model achieving the highest validation performance is selected for final testing. Each method is evaluated using its own preprocessing pipeline and corresponding channel configuration: 16 channels for BIOT and 23 channels for LaBraM. To improve statistical reliability, all experiments are repeated three times with different random seeds, and we report the mean and standard deviation across runs.

[1] https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml

C.2 THE TUEV DATASET

Description. The TUEV[2] corpus is a clinically annotated subset of TUEG that categorizes EEG segments into six event types: spike and sharp wave (SPSW), generalized periodic epileptiform discharges (GPED), periodic lateralized epileptiform discharges (PLED), eye movement (EYEM), artifact (ARTF), and background (BCKG). All recordings contain 23 EEG channels sampled at 256 Hz, and the dataset provides 112,491 non-overlapping 5-second segments for multiclass classification.

Preprocessing. We adopt the same preprocessing pipeline as in TUAB. Each EEG recording is resampled to 200 Hz and transformed into the standardized 16-channel bipolar montage used in BIOT Yang et al. (2023).Segments shorter than 5 s are discarded. Each 5-second segment inherits the event label of its source annotation and is stored as an individual sample for downstream training, validation, and evaluation.

Experimental Configuration. Following the official subject-wise split, the training portion is further divided into an 80%–20% train–validation split, and the model with the best validation accuracy is used for final testing. All experiments are repeated three times with different random seeds, and the mean and standard deviation are reported.

C.3 THE BCIC2B DATASET

Description. The BCIC2B motor-imagery dataset[3] contains recordings from 9 participants and includes both EEG and EOG modalities. EEG signals were collected from three bipolar channels (C3, Cz, C4) at a sampling rate of 250 Hz, while ocular activity was monitored using three monopolar EOG electrodes with a dynamic range of ± 1 mV. Each participant completed five experimental runs: the first two were screening sessions without feedback, and the remaining three provided feedback. The motor imagery task involved imagining either left-hand or right-hand movements (class 1 and class 2). For the screening phase, each subject was recorded on two separate days within a two-week interval, and every session consisted of 120 trials evenly distributed across the two classes.

Preprocessing. We discard all ocular channels (`EOG:ch01`, `EOG:ch02`, `EOG:ch03`) prior to further processing. The EEG signals are band-pass filtered within 0–38 Hz, resampled to 200 Hz, and subsequently normalized using EA normalization He & Wu (2019), which is widely adopted for motor-imagery datasets.

Experimental Configuration. We employ a subject-independent evaluation. Since the BCIC2B dataset contains recordings from 9 subjects, the experiments follow a 9-fold LOSO cross-validation protocol. In each fold, one subject (e.g., `sub1`) is designated as the test set, and the remaining subjects (`sub2--sub9`) are used for training. To enhance statistical reliability, every experiment is repeated 3 times with different random seeds, and the mean and standard deviation are reported.

C.4 THE PHYSIOP300 DATASET

Description. The PhysioNetP300 dataset[4] includes recordings from 12 participants performing a P300-based character-spelling task using a Donchin speller. EEG was acquired with a 64-channel BioSemi ActiveTwo system at a sampling rate of 2048 Hz. Each subject was asked to spell 20 characters. For each run, a target character was randomly chosen, and the rows and columns of a standard 6×6 matrix were intensified for 100 ms with 50 ms intervals (SOA of 150 ms), resulting in roughly 20 flashes of the target per run. During the sequence, subjects focused on the designated character and counted its occurrences.

Preprocessing. We retain all 64 EEG channels from the BioSemi montage and first convert the raw signals into uniform units. The data are band-pass filtered within 0–120 Hz and resampled to 200 Hz. For each trial, we use a 2-second window beginning 0.7 s before the flash onset, which corresponds approximately to a $[-0.1, 2]$ s peri-stimulus interval after resampling. Each extracted epoch is assigned a binary label (target or non-target) depending on whether the intensified row or column contains the designated character.

[2] `https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml`
[3] `https://www.bbci.de/competition/iv/#datasets`
[4] `https://physionet.org/content/erpbci/1.0.0/`

Table 10: Dataset–backbone specific hyperparameters for downstream training.

Hyperparameters	TUAB/TUEV		BCIC2B		PhysioNetP300	
	LaBraM	BIOT	LaBraM	BIOT	LaBraM	BIOT
Batch size	64	64	64	64	64	64
LR scheduler	Cosine	Cosine	OneCycle	OneCycle	OneCycle	OneCycle
Start learning rate	—	—	1.5e-5	1.5e-5	3.0e-5	3.0e-5
Peak/Max learning rate	5e-4	2.5e-3	4.0e-4	4.0e-4	8.0e-4	8.0e-4
Minimal learning rate	1e-6	1e-6	1.5e-7	1.5e-7	3.0e-7	3.0e-7
Optimizer	AdamW	Adam	AdamW	AdamW	AdamW	AdamW
Adam β	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999	0.9, 0.999
Weight decay	0.05	0.0005	0.05	0.05	0.05	0.05
Total epochs	50	50	100	100	100	100
Warmup epochs	5	0	—	—	—	—
Dropout	0	0.2	0	0.2	0	0.2
Drop path	0.1	0	0	0	0	0
Layer-wise LR decay	0.65	—	-	—	-	—
Label smoothing (multi-class)	0.1	—	0.1	—	0.1	—
Sliding-window step	—	—	—	—	125 ms	125 ms

Table 11: Comparison on TUAB using the CBraMod backbone.

Methods	Model Size (#Params)	Time / Epoch (m)	Balanced Acc.	Cohen's Kappa	Weighted F1
CBraMod (Original)*	4.0M	10.42	0.7973 ± 0.0024	0.8805 ± 0.0043	0.8745 ± 0.0031
CBraMod + GAP*	4.0M	10.02	0.7978 ± 0.0031	0.8749 ± 0.0165	0.8774 ± 0.0143
CBraMod + iSQRT-COV	4.0M+Δ20.3K	17.89	0.8046 ± 0.0047	0.8873 ± 0.0064	0.8841 ± 0.0022
CBraMod + SVD-Padé	4.0M+Δ20.3K	29.97	0.8069 ± 0.0051	0.8910 ± 0.0053	0.8885 ± 0.0025
CBraMod + iSICE	4.0M+Δ20.3K	65.38	0.8027 ± 0.0054	0.8858 ± 0.0049	0.8823 ± 0.0021
CBraMod + RHOP	4.0M+Δ4.6K	20.14	$\mathbf{0.8131 \pm 0.0019}$	$\mathbf{0.8972 \pm 0.0052}$	$\mathbf{0.8913 \pm 0.0024}$

* Re-implemented following the official CBraMod preprocessing protocol.

Experimental Configuration. We follow the configuration used in BENDR Kostas et al. (2021) and exclude subjects 8, 10, and 12, leaving data from the remaining 9 participants. A subject-independent evaluation is employed, using a 9-fold LOSO cross-validation protocol in which one subject is held out for testing while the others are used for training. During fine-tuning, the PhysioNetP300 data are processed using a 125 ms sliding window to extract temporally localized ERP segments for classification.

D Hyperparameters Setting

Unless otherwise stated, we use the dataset–backbone specific configurations summarized in Tab. 10. For TUAB and TUEV, models are fine-tuned for 50 epochs with batch size 64 and cosine scheduling: LaBraM uses AdamW with peak learning rate 5e-4, minimal 1e-6, 5 warmup epochs, drop path 0.1, and layer-wise decay 0.65; BIOT uses Adam with peak learning rate 2.5e-3, minimal 1e-6, no warmup, and dropout 0.2. For BCIC2B and PhysioNetP300, we adopt LOSO validation and OneCycle schedules for 100 epochs with batch size 64. BCIC2B starts at 1.5e-5, peaks at 4.0e-4, and decays to 1.5e-7, while PhysioNetP300 starts at 3.0e-5, peaks at 8.0e-4, and decays to 3.0e-7. Other optimizer settings follow Tab. 10 to ensure fair comparison across heads.

E More Results

E.1 Comparisons Using the CBraMod Backbone

The original CBraMod (Wang et al., 2025) employs a multi-layer nonlinear classifier rather than a GAP head. To enable a fair comparison, we replaced this classifier with GAP, RHOP, and several existing covariance-based pooling heads, while keeping all remaining components unchanged. Because the exact subject-level data partitions used in CBraMod cannot be faithfully reproduced, we follow the authors' official preprocessing protocol and re-implement the model according to their

Table 12: Ablation of FM iterations on TUAB.

Methods	#Iter	Time/Epoch (min)	Balanced Acc.	AUC-PR	AUROC
LaBraM-Base	1	21.48	0.8244 ± 0.0012	0.9078 ± 0.0012	0.9105 ± 0.0011
LaBraM-Base	2	24.59	0.8261 ± 0.0011	$\mathbf{0.9086 \pm 0.0011}$	0.9110 ± 0.0010
LaBraM-Base	3	27.71	$\mathbf{0.8263 \pm 0.0012}$	0.9085 ± 0.0011	$\mathbf{0.9111 \pm 0.0010}$
LaBraM-Base	∞	40.19	0.8239 ± 0.0013	0.9077 ± 0.0011	0.9107 ± 0.0010

Table 13: Ablation of FM iterations on TUEV.

Methods	#Iter	Time/Epoch (min)	Balanced Acc.	Cohen's Kappa	Weighted F1
LaBraM-Base	1	2.25	0.6380 ± 0.0056	0.6785 ± 0.0079	0.8420 ± 0.0038
LaBraM-Base	2	3.38	$\mathbf{0.6401 \pm 0.0054}$	0.6802 ± 0.0077	0.8447 ± 0.0036
LaBraM-Base	3	5.02	0.6400 ± 0.0053	$\mathbf{0.6803 \pm 0.0076}$	$\mathbf{0.8451 \pm 0.0037}$
LaBraM-Base	∞	10.21	0.6360 ± 0.0054	0.6783 ± 0.0076	0.8414 ± 0.0037

recommendations.[5] As shown in Tab. 11, RHOP consistently outperforms both the original classifier and other pooling baselines across all evaluation metrics, highlighting its clear advantage when integrated into the CBraMod backbone.

E.2 Ablation on the Number of Fréchet Mean (FM) Iterations

We evaluate the impact of the number of Fréchet Mean (FM) iterations on both TUAB and TUEV. The results are summarized in Tabs. 12 and 13, where the iteration count is set to 1, 2, 3, and ∞, with ∞ denoting iteration until convergence (i.e., the matrix norm change falling below 1×10^{-5}).

Across both datasets, performance improves noticeably from 1 to 2 iterations, but the gains saturate quickly thereafter: configurations using 2, 3, or ∞ iterations yield nearly identical results. This indicates that a small number of FM iterations is sufficient for stable and effective Riemannian aggregation. Additional iterations incur significant computational overhead without providing further benefits. Consequently, we adopt 1 iteration as the default setting, achieving a strong balance between accuracy and efficiency. This observation is also consistent with prior studies Brooks et al. (2019); Chen et al. (2024); Chakraborty et al. (2022).

E.3 More Ablation on (k, k')

We further investigate the influence of the embedding dimensions k and k' on model performance. As shown in Tab. 14, in the Quotient-Gaussian embedding, increasing k expands the SPD dimension from d to $d + k$, which amplifies the first-order term $k\mu\mu^\top$. As k grows, the embedding becomes more mean-dominated, weakening the relative contribution of second-order statistics that are known to be highly informative for EEG decoding. Moreover, enlarging $d + k$ introduces quadratic growth in matrix size, leading to increased computational burden and potential numerical instability. These effects collectively explain why excessively large k tends to degrade performance.

A parallel interpretation applies to k', which controls the dimensionality of the embedded mean vector: larger values introduce additional first-order information but also increase representation size, resulting in diminishing returns and eventual performance decline. As summarized in Tab. 14, performance improves initially as (k, k') increases, but the gains plateau and then decrease for overly large values. Notably, $(k = 2, k' = 2)$ and $(k = 3, k' = 3)$ consistently lie at or near the optimum across TUAB and TUEV, indicating that moderate dimensionality provides the balance between representational richness, computational efficiency, and numerical stability.

F Fréchet Mean on SPD manifolds

As shown in Alg. 2, the Karcher flow algorithm computes the weighted Fréchet mean (WFM) on the SPD manifold through an iterative process. In each iteration, the data points are projected onto the tangent space at the current estimate G_{k-1} using the logarithmic map, a weighted average is calculated in this tangent space, and the result is mapped back to the manifold using the exponential map. This algorithm is guaranteed to converge on manifolds with non-positive curvatures, such as

[5] https://github.com/wjq-learning/CBraMod/blob/main/preprocessing/README.md

Table 14: Ablation on embedding dimensions (k, k') for LaBraM-Base.

k	k'	TUAB			TUEV		
		Balanced Acc.	**AUC-PR**	**AUROC**	**Balanced Acc.**	**Cohen's Kappa**	**Weighted F1**
1	1	0.8065±0.0019	0.8973±0.0017	0.8988±0.0012	0.6457±0.0121	0.5772±0.0203	0.7876±0.0114
	2	0.8193±0.0021	0.9054±0.0016	0.9064±0.0013	0.6339±0.0184	0.6078±0.0227	0.8062±0.0102
	3	0.8199±0.0020	0.9054±0.0015	0.9060±0.0012	**0.6726±0.0137**	0.6456±0.0181	0.8224±0.0097
	4	0.8181±0.0019	0.9017±0.0016	0.9042±0.0012	0.6325±0.0142	0.6359±0.0208	0.8184±0.0105
2	1	0.8171±0.0018	0.8975±0.0018	0.9013±0.0012	0.6631±0.0116	0.6471±0.0193	0.8252±0.0089
	2	**0.8249±0.0022**	0.9044±0.0014	0.9071±0.0011	0.6459±0.0128	0.6008±0.0251	0.7970±0.0095
	3	0.8163±0.0017	0.9058±0.0016	0.9057±0.0011	0.6440±0.0137	0.6196±0.0228	0.8145±0.0117
	4	0.8164±0.0018	0.9043±0.0015	0.9046±0.0012	0.6409±0.0135	0.6302±0.0221	0.8129±0.0110
3	1	0.8183±0.0019	0.9035±0.0015	0.9062±0.0011	0.6659±0.0123	0.6317±0.0273	0.8176±0.0135
	2	0.8193±0.0020	0.9008±0.0016	0.9039±0.0012	0.6372±0.0161	0.6220±0.0234	0.8135±0.0092
	3	0.8244±0.0012	**0.9078±0.0012**	**0.9105±0.0011**	0.6380±0.0056	**0.6785±0.0079**	**0.8420±0.0038**
	4	0.8187±0.0021	0.9024±0.0017	0.9041±0.0013	0.6484±0.0138	0.6116±0.0217	0.8123±0.0109
4	1	0.8181±0.0020	0.9017±0.0016	0.9042±0.0012	0.6475±0.0125	0.6341±0.0208	0.8188±0.0107
	2	0.8164±0.0019	0.9043±0.0015	0.9046±0.0011	0.6362±0.0141	0.6193±0.0221	0.8127±0.0112
	3	0.8187±0.0021	0.9024±0.0017	0.9041±0.0013	0.6484±0.0138	0.6116±0.0217	0.8123±0.0109
	4	0.8190±0.0020	0.9031±0.0016	0.9050±0.0012	0.6400±0.0132	0.6250±0.0225	0.8150±0.0110

Algorithm 2: Karcher Flow Algorithm on the SPD Manifold under AIM

Input : A set of SPD matrices $X_{1\ldots N} \in \mathcal{S}_d^+$
Number of iterations K
Output : The FM $G_K \in \mathcal{S}_d^+$
Initialize $G_0 = \frac{1}{N}\sum_{i=1}^{N} X_i$
for $k \leftarrow 1$ **to** K **do**
  $G_k \leftarrow \mathrm{Exp}_{G_{k-1}}\left(\sum_{i=1}^{N} \mathrm{Log}_{G_{k-1}}(X_i)\right)$
end

$\mathcal{S}_d^+$ (Karcher, 1977a). In practice, we initialize G_0 with the arithmetic mean and set the number of iterations to $K = 1$, which provides a stable and efficient approximation.

G Limitation

RHOP involves batched SVD to compute matrix logarithms and update sparse inverse covariance. While such operations are generally less optimized than convolutions and attention on current GPUs, in our case, the matrix size is very small (at most 15×15), making the actual overhead negligible. Thus, this limitation is mainly an implementation detail rather than a practical concern.

H Future Work

RHOP is primarily motivated by EEG decoding, but its basic framework and statistics-based principles are general. The Quotient Gaussian Embedding and Riemannian Gaussian Embedding capture two key characteristics that are particularly salient in EEG signals, i.e., spatiotemporal dependency and scale variation. RHOP employs second-order pooling across temporal segments, which preserves fine-grained temporal relations that are lost in first-order pooling schemes. More importantly, the quotient Gaussian provides scale-invariant representations, allowing the model to focus on relational structure rather than absolute amplitude. These theoretical properties make RHOP naturally well-suited for EEG foundation models, which typically rely on mean-based pooling.

However, the RHOP head can also be extended to other multi-channel temporal modalities that exhibit similar structural characteristics, such as ECG, MEG, and fMRI-derived time series. In these domains, the high-order pooling framework may enhance representation ability, while the degree of quotient Gaussian can be adapted to reflect the discriminative role of signal amplitude. Investigating these directions may help understand when quotient-based modeling is advantageous and shed light on how RHOP could generalize beyond EEG.

I PROOF OF THE THM. 4.2

Proof of Thm. 4.2 . We construct $\Psi : \mathcal{QN}(n) \to \mathcal{S}^{+,1}_{n+k}$ and show it is a smooth embedding. Fix $k \geq 1$ and define

$$\Phi(C, \mu) = \begin{bmatrix} C + k\mu\mu^\top & \mu^{(k)} \\ \mu^{(k)\top} & I_k \end{bmatrix}, \qquad \mu^{(k)} = \underbrace{[\mu, \ldots, \mu]}_{k \text{ columns}}. \tag{14}$$

Observe that $\Phi(C, \mu)$ is SPD and $\det \Phi(C, \mu) = \det C$. Indeed, with

$$T(\mu) = \begin{bmatrix} I_n & \mu^{(k)} \\ 0 & I_k \end{bmatrix}, \tag{15}$$

a direct multiplication gives

$$\Phi(C, \mu) = T(\mu) \begin{bmatrix} C & 0 \\ 0 & I_k \end{bmatrix} T(\mu)^\top. \tag{16}$$

Since $C \in \mathcal{S}^+_n$ and $T(\mu)$ is invertible, $\Phi(C, \mu)$ is SPD. Also $\det T(\mu) = 1$, hence $\det \Phi(C, \mu) = \det C$.

Set $s(C) = (\det C)^{-\frac{1}{n+k}}$ and define

$$\Psi(C, \mu) = s(C)\, \Phi(C, \mu). \tag{17}$$

Then $\det \Psi(C, \mu) = 1$, so $\Psi(C, \mu) \in \mathcal{S}^{+,1}_{n+k}$, matching Eq. (6).

For injectivity and a smooth inverse on the image, suppose $\Psi(C_1, \mu_1) = \Psi(C_2, \mu_2) =: X$, and write $X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$. By construction $X_{22} = s(C_i) I_k$, so $s(C_1) = s(C_2) =: s$. Then $X_{12} = s\, \mu_i^{(k)}$ implies $\mu_1 = \mu_2 =: \mu$, and $X_{11} = s(C_i + k\mu\mu^\top)$ yields $C_1 = C_2$.

Conversely, given X in the image, recover $s = \frac{1}{k} \operatorname{tr}(X_{22})$, then $\mu = s^{-1} X_{12} e_1$ (with e_1 the first basis vector in $\mathbb{R}^k$), and finally $C = s^{-1} X_{11} - k\mu\mu^\top$. These depend smoothly on X and satisfy $\Psi(C, \mu) = X$. Hence Ψ is a smooth embedding.

It remains to situate this within the affine action. Let $N(n)$ be the space of Gaussians (Σ, μ) with $\Sigma \in \mathcal{S}^+_n$. As shown in Nguyen (2021); Lovrić et al. (2000b), Eqs. (41)–(47), the transitive $Aff^+(n)$-action on $N(n)$ and the embedding $j : Aff^+(n) \hookrightarrow SL(n + k)$ induce

$$(\Sigma, \mu) \longmapsto (\det \Sigma)^{-\frac{1}{n+k}} \begin{bmatrix} \Sigma + k\mu\mu^\top & \mu^{(k)} \\ \mu^{(k)\top} & I_k \end{bmatrix} \in \mathcal{S}^{+,1}_{n+k}. \tag{18}$$

A quotient Gaussian identifies (Σ, μ) up to a positive scalar multiple on Σ. Choosing the canonical representative $C = (\det \Sigma)^{-\frac{1}{n}} \Sigma$ yields exactly $\Psi(C, \mu)$. Therefore, Ψ is the natural embedding for quotient Gaussians. □