
Towards a General Attention Framework on Gyrovector Spaces for Matrix Manifolds

Rui Wang¹, Chen Hu¹, Xiaoning Song¹, Xiao-Jun Wu¹, Nicu Sebe², Ziheng Chen^{2*}

¹School of Artificial Intelligence and Computer Science, Jiangnan University

²Department of Information Engineering and Computer Science, University of Trento

{cs_wr, 6233112017, x.song, wu_xiaojun}@jiangnan.edu.cn

niculae.sebe@unitn.it, ziheng_ch@163.com

Abstract

Deep Neural Networks (DNNs) operating on non-Euclidean geometries have recently demonstrated impressive performance across various machine-learning applications. Several studies have extended the attention mechanism to different manifolds. However, most existing non-Euclidean attention models are tailored to specific geometries, limiting their applicability. On the other hand, recent studies show that several matrix manifolds, such as Symmetric Positive Definite (SPD), Symmetric Positive Semi-Definite (SPSD), and Grassmannian manifolds, admit gyrovector structures, which extend vector addition and scalar product into manifolds. Leveraging these properties, we propose a Gyro Attention (GyroAtt) framework over general gyrovector spaces, applicable to various matrix geometries. Empirically, we manifest GyroAtt on three gyro structures on the SPD manifold, three on the SPSD manifold, and one on the Grassmannian manifold. Extensive experiments on four Electroencephalography (EEG) datasets demonstrate the effectiveness of our framework. The code is available at <https://github.com/ChenHu-ML/GyroAtt>.

1 Introduction

Recently, DNNs over Riemannian manifolds, known as Riemannian neural networks, have garnered increasing attention in various applications [26, 11, 61, 41, 13, 73, 53, 72, 18]. Commonly encountered manifolds include vector manifolds, such as hyperbolic [68] and spherical spaces [65], and matrix manifolds, such as SPD [5], SPSD [9, 10], and Grassmannian manifolds [1]. Among these non-Euclidean spaces, hyperbolic manifolds stand out due to the rich algebraic structure of gyrovector spaces [67–69], which enables principled and convenient extensions of Euclidean deep learning to hyperbolic manifolds [26, 61, 7]. In contrast, matrix manifolds offer a compelling trade-off between structural expressiveness and computational feasibility [20]. As a result, neural networks defined on matrix manifolds have emerged as appealing alternatives to their hyperbolic counterparts across various applications [40, 51, 52, 14, 37]. Notably, recent studies [40, 50–52] have demonstrated that several matrix manifolds, including SPD, SPSD, and Grassmannian, admit gyro structures, facilitating the extension of existing neural network components to these manifolds [53].

Inspired by the success of attention mechanisms in DNNs [70, 32, 22], researchers have explored their extensions to non-Euclidean geometries. Wherein, Gulcehre et al. [26] introduced attention to the hyperbolic spaces based on the hyperboloid and Klein models, while Pan et al. [55] extended attention to the SPD manifolds under the Log-Euclidean Metric (LEM). Wang et al. [72] further adapted it to the Grassmannian manifolds using an extrinsic approach under the projection distance. However, existing manifold-attention designs exhibit two key challenges: i) these designs are tailored for specific manifolds and metrics, limiting their generalizability; ii) although their network layers,

*Corresponding author.

such as BiMap and ReEig in MAtt [55], and FrMap in GDLNet [72], preserve manifold constraints, they are fundamentally numerical and only partially respect the underlying Riemannian geometry.

As self-attention serves as the prototype of other attention variants, this work extends it to non-Euclidean settings. Leveraging the fact that several matrix manifolds admit gyro structures, we propose GyroAtt, a general framework for self-attention over gyrovector spaces. Unlike previous manifold attention approaches handcrafted for specific geometries [26, 55, 72], GyroAtt provides a unified building paradigm for attention design across different matrix manifolds. Notably, by intrinsically respecting Riemannian geometry and gyro algebra, GyroAtt generalizes fundamental attention components, such as linear transformations, attention computation, and feature aggregation, to gyrovector spaces. Specifically, we introduce *gyro homomorphisms*, extending linear transformations to gyrovector spaces. The attention scores are computed using a gyro distance-based function, while feature aggregation is performed via the Weighted Fréchet Mean (WFM) [25], the manifold analogue of the Euclidean weighted average. We instantiate GyroAtt on three gyro structures for the SPD manifold, three for the SPSD manifold, and one for the Grassmannian manifold. Extensive experiments on four EEG benchmarks show consistent improvements over geometry-specific manifold-attention baselines; *e.g.*, GyroAtt-SPSD outperforms MAtt [55] by **3.2%** on the MAMEM, while GyroAtt-SPD achieves **8.9%** and **7.8%** gains in inter-subject and inter-session settings on the BNCI2014001. Our **main contributions** are summarized below:

- **Generalizing attention to gyrovector spaces.** This is the first attention framework that unifies operations across diverse matrix manifolds through a common gyrovector space formulation, enabling flexible changes in the underlying geometry within a shared network structure.
- **Implementation on seven matrix gyrovector spaces.** We implement GyroAtt across three different matrix manifolds: *three gyro structures on the SPD manifold, one on the Grassmannian manifold, and three on the SPSD manifold*. To the best of our knowledge, we are the first to investigate attention mechanisms for the SPSD manifold.
- **Empirical validation under EEG signal classification.** We validate the effectiveness of GyroAtt on four EEG benchmarks. Apart from its good performance, the optimal geometries vary across different EEG tasks, demonstrating its efficacy and flexibility.

2 Preliminaries

This section reviews gyrogroups, gyrovector spaces, and their realizations on the SPD, Grassmannian, and SPSD manifolds, together with WFM. For further details, please refer to [68, 69, 56, 5, 10, 8].

Gyrogroups and gyrovector spaces. Gyrogroups and gyrovector spaces generalize groups and vector spaces, offering a powerful framework to analyze non-Euclidean geometries. A gyrogroup equips a set (G, \oplus) with a non-associative “addition” \oplus whose failure of associativity is controlled by a gyration $\text{gyr}[\cdot, \cdot]$. Adding a scalar multiplication \otimes that obeys vector-like axioms upgrades the structure to a *gyrovector space*. A review of gyrogroups and gyrovector spaces is given in App. D.1.

SPD geometries. Let \mathcal{S}_d^{++} denote the space of $d \times d$ SPD matrices. When equipped with the Affine-Invariant Metric (AIM) [56], LEM [5], and Log-Cholesky Metric (LCM) [45], \mathcal{S}_d^{++} induces three gyrovector spaces with corresponding binary operations [52] \oplus_{ai} , \oplus_{le} , and \oplus_{lc} , and associated gyro distances [19] $d_{\text{spd}}^{ai}(\cdot)$, $d_{\text{spd}}^{le}(\cdot)$, and $d_{\text{spd}}^{lc}(\cdot)$, given in Tab. 1. Given $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, we denote by $\text{logm}(\cdot)$ and $\text{expm}(\cdot)$ the matrix logarithm and exponential. $\mathcal{L}(\mathbf{P})$ means the Cholesky decomposition of \mathbf{P} , with $\mathcal{L}^{-1}(\cdot)$ denoting its inverse map. Besides, $[\mathcal{L}(\mathbf{P})]$ denotes the strictly lower triangular part, $\mathbb{D}(\mathbf{P})$ signifies the diagonal part, and then $\psi_{\text{LC}}(\mathbf{P}) = [\mathcal{L}(\mathbf{P})] + \text{logm}(\mathbb{D}(\mathbf{P}))$.

Grassmannian geometries. The Grassmannian manifold consists of all q -dimensional linear subspaces within \mathbb{R}^d . This study centers on the Orthonormal Basis (ONB) perspective, where a subspace is represented by $\mathbf{Y} \in \mathbb{R}^{d \times q}$ satisfying $\mathbf{Y}^\top \mathbf{Y} = \mathbf{I}_q$. We denote the set of such ONB representations as $\mathbf{Y} \in \tilde{\mathcal{G}}(q, d)$. A point in $\tilde{\mathcal{G}}(q, d)$ represents an equivalence class of orthonormal bases: $[\mathbf{Y}] = \{\tilde{\mathbf{Y}} \mid \tilde{\mathbf{Y}} = \mathbf{Y}\mathbf{O}, \mathbf{O} \in \text{O}(q)\}$. By abuse of notation, we use $[\mathbf{Y}]$ or \mathbf{Y} interchangeably. Under the ONB perspective, the Grassmannian manifold induces a nonreductive gyrovector structure [52]. The binary operation $\tilde{\oplus}_{gr}$ [52] and gyro distance $d_{gr}(\cdot)$ [19] for $\mathbf{U}, \mathbf{V} \in \tilde{\mathcal{G}}(q, d)$ are defined in Tab. 1.

SPSD geometries. Let $\mathcal{S}_{d,q}^+$ denote the set of $d \times d$ SPSD matrices of rank $q \leq d$. Each $\mathbf{P} \in \mathcal{S}_{d,q}^+$ admits a canonical decomposition $\mathbf{P} = \mathbf{U}_P \mathbf{S}_P \mathbf{U}_P^\top$, where $\mathbf{U}_P \in \tilde{\mathcal{G}}(q, d)$ and $\mathbf{S}_P \in \mathcal{S}_q^{++}$ [9, 10].

Table 1: Summary of the gyro additions and gyro distances over different manifolds.

Manifold	Metric	Gyro addition	Gyro distance
SPD	AIM	$\mathbf{P} \oplus_{ai} \mathbf{Q} = \mathbf{P}^{\frac{1}{2}} \mathbf{Q} \mathbf{P}^{\frac{1}{2}}$	$\left\ \logm \left(\mathbf{Q}^{-\frac{1}{2}} \mathbf{P} \mathbf{Q}^{-\frac{1}{2}} \right) \right\ _{\mathbf{F}}$
	LEM	$\mathbf{P} \oplus_{le} \mathbf{Q} = \expm(\logm(\mathbf{P}) + \logm(\mathbf{Q}))$	$\left\ \logm(\mathbf{P}) - \logm(\mathbf{Q}) \right\ _{\mathbf{F}}$
	LCM	$\mathbf{P} \oplus_{lc} \mathbf{Q} = \mathcal{L}^{-1}([\mathcal{L}(\mathbf{P})] + [\mathcal{L}(\mathbf{Q})] + \mathbb{D}(\mathcal{L}(\mathbf{P})) \mathbb{D}(\mathcal{L}(\mathbf{Q})))$	$\left\ \psi_{LC}(\mathbf{P}) - \psi_{LC}(\mathbf{Q}) \right\ _{\mathbf{F}}$
Grassmannian	ONB perspective	$\mathbf{U} \widetilde{\oplus}_{gr} \mathbf{V} = \expm([\mathbf{Log}_{\mathbf{I}_{d,q}}^{gr}(\mathbf{U} \mathbf{U}^{\top}), \mathbf{I}_{d,q}]) \mathbf{V}$	$\left\ \arccos(\Sigma) \right\ $ $\mathbf{U}^{\top} \mathbf{V} \stackrel{\text{SVD}}{=} \mathbf{O} \boldsymbol{\Sigma} \mathbf{R}^{\top}$
SPSD	$(g_{gr}, \lambda g_{spd})$	$(\mathbf{U}_P, \mathbf{S}_P) \oplus_{psd,g} (\mathbf{U}_Q, \mathbf{S}_Q) = (\mathbf{U}_P \widetilde{\oplus}_{gr} \mathbf{U}_Q, \mathbf{S}_P \oplus_g \mathbf{S}_Q)$	$d_{gr}(\mathbf{U}_P, \mathbf{U}_Q) + \lambda d_{spd}^g(\mathbf{S}_P, \mathbf{S}_Q)$

Following Nguyen et al. [53], we represent \mathbf{P} in the product space $\mathcal{G}(q, d) \times \mathcal{S}_q^{++}$, as detailed in App. D.3. By equipping \mathcal{S}_q^{++} with a Riemannian metric $g \in \{ai, le, lc\}$, the space $\mathcal{S}_{d,q}^+$ naturally inherits a corresponding gyrovector structure [53]. The induced binary operation $\oplus_{psd,g}$ [53] and the related gyro distance $d_{psd,g}(\cdot)$ [19] are defined in Tab. 1, where $\lambda > 0$.

WFMs. The WFM [25] of points $\{\mathbf{P}_{i \dots N}\}$ is the point $\mathbf{S} \in \mathcal{M}$ that minimizes the weighted sum of squared distances to all points $\{\mathbf{P}_{i \dots N}\}$. Given weights $\{w_{1 \dots N}\}$ satisfying the convexity constraint, *i.e.*, $\forall i, w_i > 0$ and $\sum_i w_i = 1$, the WFM is expressed as:

$$\text{WFM}(\{w_i\}, \{\mathbf{P}_i\}) = \arg \min_{\mathbf{S} \in \mathcal{M}} \sum_{i=1}^N w_i d^2(\mathbf{P}_i, \mathbf{S}), \quad (1)$$

where $d(\mathbf{P}_i, \mathbf{S})$ is the distance between \mathbf{S} and \mathbf{P}_i . On Riemannian manifolds, WFM uniquely exists when samples are locally distributed [2], which is detailed in App. D.2.1. In this paper, we always assume WFM are well-defined.

3 Proposed method

Inspired by the success of attention mechanism [70, 32], recent studies have extended attention models to non-Euclidean settings, including hyperbolic [26], SPD [55], and Grassmannian [72] manifolds. However, these approaches typically rely on geometry-specific operations, such as BiMap layer [33] for SPD manifolds or FrMap layer [34] for Grassmannian manifolds, which limits their applicability and generalization across manifolds. In contrast, our proposed GyroAtt framework leverages gyro structures to provide a unified and principled formulation of attention across matrix manifolds. To motivate our design, we first revisit prior manifold-based attention mechanisms and summarize their core components in Tab. 2, before introducing our generalized gyro-based formulation.

3.1 Revisiting attention mechanisms on different geometries

Despite differences in underlying geometries, self-attention mechanisms generally follow a common three-stage pipeline: 1) feature transformations to compute query (\mathbf{q}_i), key (\mathbf{k}_i), and value (\mathbf{v}_i); 2) similarity computations, often based on distances or inner products; 3) weighted aggregations of the values (\mathbf{r}_i). Tab. 2 summarizes these components across geometry-aware attention models.

Euclidean. Standard Transformers [70] generate \mathbf{q}_i , \mathbf{k}_i , and \mathbf{v}_i through linear projections $\text{Linear}(\cdot)$, compute attention weights using scaled dot-product $\text{Softmax}(\langle \mathbf{q}_i, \mathbf{k}_j \rangle / \sqrt{d_k})$, and obtain outputs via $\mathbf{r}_i = \sum_{j=1}^N \mathcal{A}_{ij} \mathbf{v}_j$, where N signifies the number of value tokens.

Hyperbolic. HAN [26] extends attention to the hyperbolic space using the Klein (\mathbb{K}^d) and hyperboloid (\mathbb{H}^d) models. Euclidean features are first projected onto the manifold via $\pi_{\mathbb{R} \rightarrow \mathbb{K}}$ and $\pi_{\mathbb{R} \rightarrow \mathbb{H}}$. Attention scores are computed by $-\beta d(\mathbf{q}_i, \mathbf{k}_j) - c$, where $\beta > 0$ and c are learnable parameters. Value aggregations are performed using the Einstein midpoint [66].

SPD. MAtt [55] formulates attention on the SPD manifolds under LEM. Queries and keys are obtained via the BiMap function [33], attention scores are measured using the LEM-based geodesic distance, and LEM-based WFM is employed for value aggregation.

Grassmannian. GDLNet [72] adapts attention to the Grassmannian manifolds by employing FrMap and ReOrth layers for data transformation and activation [34]. Attention scores are computed based on the projection distance, while the extrinsic WFM [62] is employed for value aggregation.

In summary, the above manifold attention approaches are inherently tied to particular manifolds or metrics, limiting their direct generalization across different geometries.

Table 2: Summary of attention based on different geometries, where $f_s(\cdot)$ denotes the softmax.

Method	Geometries	Transformations	$d(\mathbf{q}_i, \mathbf{k}_i)$	Attention \mathcal{A}_{ij}	Aggregation $\mathbf{r}_i(\mathbf{R}_i)$
Transformer [70]	Euclidean	$\text{Linear}(\mathbf{x}_i)$	$\ \mathbf{q}_i - \mathbf{k}_j\ _F$	$f_s(\langle \mathbf{q}_i, \mathbf{k}_j \rangle / \sqrt{d_k})$	Arithmetic mean $\sum_j^N \mathcal{A}_{ij} \mathbf{v}_j$
HAN [26]	Hyperbolic	$\pi_{\mathbb{R} \rightarrow \mathbb{H}}(\text{Linear}(\mathbf{x}_i))$ $\pi_{\mathbb{R} \rightarrow \mathbb{K}}(\text{Linear}(\mathbf{x}_i))$	$\text{arccosh}(-\langle \mathbf{q}_i, \mathbf{k}_j \rangle_M)$	$f_s(-\beta d(\mathbf{q}_i, \mathbf{k}_j) - c)$	Einstein midpoint $\sum_j^N \left[\frac{\mathcal{A}_{ij} \gamma(\mathbf{v}_j)}{\sum_j^N \mathcal{A}_{ij} \gamma(\mathbf{v}_j)} \right] \mathbf{v}_j$
MAtt [55]	SPD under LEM	$\mathbf{W} \mathbf{X}_i \mathbf{W}^\top$	$\ \text{logm}(\mathbf{Q}_i) - \text{logm}(\mathbf{K}_i)\ _F$	$f_s((1 + \log(1 + d(\mathbf{Q}_i, \mathbf{K}_j)))^{-1})$	LEM-based WFM $\text{expm}(\sum_j^N \mathcal{A}_{ij} \text{logm}(\mathbf{V}_j))$
GDLNet [72]	Grassmannian under ONB	$\text{ReOrth}(\mathbf{W} \mathbf{X}_i)$	$\ \mathbf{Q}_i \mathbf{Q}_i^\top - \mathbf{K}_j \mathbf{K}_j^\top\ _F$	$f_s((1 + \log(1 + d(\mathbf{Q}_i, \mathbf{K}_j)))^{-1})$	Extrinsic WFM $\Phi^{-1}(\sum_j^N \mathcal{A}_{ij} \Phi(\mathbf{V}_j))$
GyroAtt (Ours)	Gyrovector spaces (SPD, SPSD, Grassmann)	Homomorphism Eq. (4)	Gyro distance	$f_s((1 + \log(1 + d(\mathbf{Q}_i, \mathbf{K}_j)))^{-1})$	WFM

3.2 Attention mechanisms over gyrovector spaces

In this part, we extend the basic attention operations illustrated in Tab. 2 to gyrovector spaces: 1) transformation through gyro homomorphisms, which preserves the gyrovector structure; 2) distance-based similarity computation; 3) value aggregation via gyro distance-based WFM.

Definition 3.1 (Gyro Homomorphisms). Let $(\mathcal{M}, \oplus_{\mathcal{M}}, \otimes_{\mathcal{M}})$ and $(\mathcal{N}, \oplus_{\mathcal{N}}, \otimes_{\mathcal{N}})$ be two (nonreductive) gyrovector spaces, $\forall \mathbf{A}, \mathbf{B} \in \mathcal{M}$, and $\forall t \in \mathbb{R}$, the map $\text{hom}(\cdot) : (\mathcal{M}, \oplus_{\mathcal{M}}, \otimes_{\mathcal{M}}) \rightarrow (\mathcal{N}, \oplus_{\mathcal{N}}, \otimes_{\mathcal{N}})$ is a (nonreductive) gyrovector space homomorphism if it satisfies:

$$\text{hom}(\mathbf{A} \oplus_{\mathcal{M}} \mathbf{B}) = \text{hom}(\mathbf{A}) \oplus_{\mathcal{N}} \text{hom}(\mathbf{B}), \quad \text{hom}(t \otimes_{\mathcal{M}} \mathbf{A}) = t \otimes_{\mathcal{N}} \text{hom}(\mathbf{A}). \quad (2)$$

If we only consider (nonreductive) gyrogroups, $(\mathcal{M}, \oplus_{\mathcal{M}})$ and $(\mathcal{N}, \oplus_{\mathcal{N}})$, a map $\text{hom}(\cdot) : (\mathcal{M}, \oplus_{\mathcal{M}}) \rightarrow (\mathcal{N}, \oplus_{\mathcal{N}})$ satisfying Eq. (35) is called a (nonreductive) gyrogroup homomorphism, which has been introduced by Suksumran and Wibooton [63]. By abuse of notations, we call the above homomorphisms collectively gyro homomorphisms. The concept of gyro homomorphisms naturally generalizes linear maps from vector spaces to gyrovector spaces. Recall that a linear map $\text{Linear}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a homomorphism of vector spaces, meaning it preserves both vector addition and scalar multiplication: for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$ and $t \in \mathbb{R}$, it satisfies:

$$\text{Linear}(\mathbf{z}_1 + \mathbf{z}_2) = \text{Linear}(\mathbf{z}_1) + \text{Linear}(\mathbf{z}_2), \quad \text{Linear}(t\mathbf{z}_1) = t \text{Linear}(\mathbf{z}_1). \quad (3)$$

Therefore, we use $\text{hom}(\cdot)$ for feature transformation. While the above offers an algebraic definition, a natural question arises: *how can such mappings be identified on manifolds?*

Theorem 3.2 (Sufficient condition for gyro homomorphisms). [↓] Let \mathcal{M} be a Riemannian homogeneous space with isometry group G . If $(\mathcal{M}, \oplus, \otimes)$ forms a gyrovector space and an isometry $f \in G$ fixes the identity element, i.e., $f(e) = e$, then f is a gyro homomorphism.

This theorem provides a way to identify gyro homomorphisms from homogeneous space isometries. Then, we calculate the correlation between \mathbf{Q}_i and \mathbf{K}_j using their gyro distance, and map $d(\mathbf{Q}_i, \mathbf{K}_j)$ to a valid attention score, as defined in Eq. (5). For feature aggregation, we resort to WFM based on the gyro distance. Specifically, given a set of input data points $\{\mathbf{X}_{i \dots N} \in \mathcal{M}\}$, the key operations of GyroAtt are listed below:

$$\mathbf{Q}_i = \text{hom}(\mathbf{X}_i), \quad \mathbf{K}_i = \text{hom}(\mathbf{X}_i), \quad \mathbf{V}_i = \text{hom}(\mathbf{X}_i), \quad \text{(Transformation)} \quad (4)$$

$$\mathcal{A}_{ij} = \text{Softmax}((1 + \log(1 + d(\mathbf{Q}_i, \mathbf{K}_j)))^{-1}), \quad \text{(Attention)} \quad (5)$$

$$\mathbf{R}_i = \text{WFM}(\mathcal{A}_i, \mathbf{V}_{i \dots N}). \quad \text{(Aggregation)} \quad (6)$$

Here, \mathcal{A}_i denote the i -th row of \mathcal{A} , and \mathbf{R}_i is the resulting data under the manifold-valued weighted average (i.e., WFM) between \mathcal{A}_i and $\mathbf{V}_{i \dots N}$. While WFM offers a principled approach for aggregating manifold-valued data, its output \mathbf{R}_i is inherently restricted to the geodesic convex hull of the input points. This constraint may limit the expressiveness of the resulting representations. To mitigate this limitation, we introduce a bias term and a nonlinear transformation following the aggregation step:

$$\phi(\mathbf{R}_i) = \sigma(\mathbf{B} \oplus \mathbf{R}_i), \quad (7)$$

where \mathbf{B} is a learnable bias parameter, and σ represents a power-based activation function.

So far, all ingredients are in place to build attention over gyrovector spaces, as shown in Alg. 4.

4 Gyro attention mechanisms on matrix manifolds

This section presents our GyroAtt framework (Alg. 4) across various matrix gyrovector spaces, including three SPD gyro spaces, one Grassmannian gyro space, and three SPSD gyro spaces.

4.1 Gyro attention mechanism over SPD manifolds

As summarized in Tab. 1, three SPD gyrovector spaces are induced by the AIM, LEM, and LCM, respectively. The corresponding gyro distance (used for attention calculation) and gyro addition (used for biasing) have been well studied in prior works [5, 56, 45]. Therefore, our focus is on defining the gyro homomorphisms, WFs, and activation for these geometries. We begin by deriving the explicit forms of the gyro homomorphisms corresponding to each SPD geometry.

Theorem 4.1 (SPD Homomorphisms). [↓] Given $\mathbf{P} \in \mathcal{S}_d^{++}$, let $\text{hom}_g : \mathcal{S}_d^{++} \rightarrow \mathcal{S}_d^{++}$ be defined as

$$\text{hom}_g(\cdot) : \mathbf{P} \mapsto \begin{cases} \mathbf{O}\mathbf{P}\mathbf{O}^\top, & g = ai, \mathbf{O} \in \text{O}(d), \\ \text{expm}(\mathbf{M} \text{logm}(\mathbf{P})\mathbf{M}^\top), & g = le, \mathbf{M} \in \mathbb{R}^{d \times d}, \\ \mathcal{L}^{-1}([L(\mathbf{P})] + \exp(\mathbb{D}(L(\mathbf{P})))), & g = lc, \mathbf{M} \in \mathbb{R}^{d \times d}, \end{cases} \quad (8)$$

where $L(\mathbf{P}) = \mathbf{M}([L(\mathbf{P})] + [L(\mathbf{P})]^\top + \mathbb{D}(L(\mathbf{P})))\mathbf{M}^\top$. For each metric $g \in \{ai, le, lc\}$, $\text{hom}_g(\cdot)$ is a gyro homomorphism with respect to $(\mathcal{S}_d^{++}, \oplus_g, \otimes_g)$. Moreover, if \mathbf{M} is orthonormal, the LEM-based homomorphism is identical to the AIM one.

Transformation. As shown in Thm. 4.1, the forms of $\text{hom}_{le}(\cdot)$ and $\text{hom}_{lc}(\cdot)$ follow from the flat structures of their geometries. When a manifold is flat, its gyro homomorphism can be directly derived. The SPD manifold under LEM is flat in the logarithmic domain, where linear mappings act as valid homomorphisms. Orthogonal constraints can improve network generalization by imposing implicit regularization [30]. Therefore, we impose orthogonality on \mathbf{M} in both $\text{hom}_{lc}(\cdot)$ and $\text{hom}_{le}(\cdot)$.

WFMs. The WFs under LEM and LCM admit closed-form solutions, whereas the one under AIM requires iterative computation via the Karcher flow algorithm [38]. These are reviewed in App. D.2.2.

Activation. As demonstrated by Chen et al. [17, Fig. 1] and Chen et al. [15, Sec. 5.1], the matrix power can deform the latent SPD geometries. As a consequence, we use matrix power as the activation function to activate the underlying Riemannian geometry.

4.2 Gyro attention mechanisms over Grassmannian manifolds

We implement GyroAtt on the Grassmannian nonreductive gyrovector spaces under the ONB perspective. The gyro distance and gyro addition are shown in Tab. 1. Similar to the SPD gyro spaces, we use gyro homomorphism for transformation and WFM for aggregation. As shown by Nguyen and Yang [52, Sec. 2.3.2], the Grassmannian gyro addition can be viewed as a non-linear activation. Therefore, we do not use additional activation before the Grassmannian gyro biasing. In the following, we discuss gyro homomorphism and WFM over the Grassmannian manifold.

Theorem 4.2 (Grassmannian Homomorphisms). [↓] Let $\mathbf{U} \in (\tilde{\mathcal{G}}(q, d), \tilde{\oplus}_{gr}, \tilde{\otimes}_{gr})$, and $\mathbf{O} = \begin{bmatrix} \mathbf{O}_q & 0 \\ 0 & \mathbf{O}_{d-q} \end{bmatrix} \in \mathbb{R}^{d,d}$, where $\mathbf{O}_q \in \mathbb{R}^{q \times q}$ and $\mathbf{O}_{d-q} \in \mathbb{R}^{(d-q) \times (d-q)}$ are orthonormal. The map $\text{hom}_{gr}(\cdot) : (\tilde{\mathcal{G}}(q, d), \tilde{\oplus}_{gr}, \tilde{\otimes}_{gr}) \rightarrow (\tilde{\mathcal{G}}(q, d), \tilde{\oplus}_{gr}, \tilde{\otimes}_{gr})$ is a gyro homomorphism defined by

$$\text{hom}_{gr}(\mathbf{U}) = \mathbf{O}\mathbf{U}. \quad (9)$$

Eq. (9) plays the role of Grassmannian feature transformation. For the weighted aggregation, since the WFM on the Grassmannian manifold does not admit a closed-form solution, we employ the Karcher flow algorithm [1, 38]. More details are provided in App. D.2.3.

4.3 Gyro attention mechanisms over SPSD manifolds

As introduced in Sec. 2, any $\mathbf{P} \in \mathcal{S}_{d,q}^+$ can be canonically represented as a pair $(\mathbf{U}_P, \mathbf{S}_P) \in \tilde{\mathcal{G}}(q, d) \times \mathcal{S}_d^{++}$ within the structured product space. According to Tab. 1, both the distance and gyro addition operations in this space are defined component-wisely over the Grassmannian and SPD subspaces. Therefore, implementing GyroAtt in the SPSD gyrovector space reduces to specifying the gyro homomorphisms, WFs, and activation for the corresponding product structure.

Table 3: Key operators of GyroAtt over different matrix gyrovector spaces.

Manifold	SPD			Grassmannian	SPSD
Metric	AIM	LEM	LCM	ONB perspective	$(g_{gr}, \lambda g_{spd})$
Homomorphism	OPO^\top	$\text{expm}(\mathbf{M} \log(\mathbf{P}) \mathbf{M}^\top)$	$\mathcal{L}^{-1}([L(\mathbf{P})] + \exp(\mathbb{D}(L(\mathbf{P}))))$	OU	$(\text{hom}_{gr}(\mathbf{U}_P), \text{hom}_g(\mathbf{S}_P))$
WFM	Karcher flow Alg. 1	Closed-form Eq. (15)	Closed-form Eq. (16)	Karcher flow Alg. 2	$(\text{WFM}_{spd}, \text{WFM}_{gr})$
Bias and Non-linearity		$(\mathbf{B}_{spd} \oplus_g \mathbf{R}_i)^p$		$\mathbf{B}_{gr} \tilde{\oplus}_{gr} \mathbf{R}_i$	$(\mathbf{B}_{gr} \tilde{\oplus}_{gr} \mathbf{U}_{R_i}, (\mathbf{B}_{spd} \oplus_{spd} \mathbf{S}_{R_i})^p)$

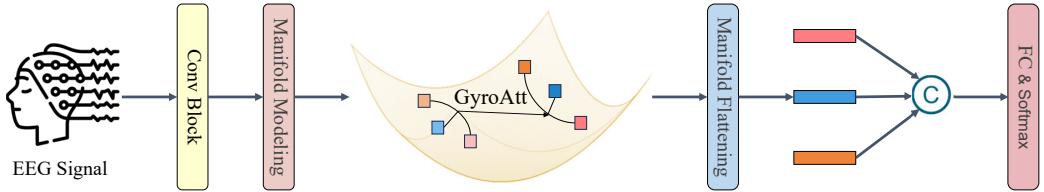


Figure 1: The GyroAtt network consists of three modules: A feature extraction module that applies convolution and manifold modeling to transform EEG signals into manifold-valued features; a gyro attention module that captures long-range dependencies among features; and a classification module that implements decision making with the flattened manifold-valued data.

Theorem 4.3 (SPSD Homomorphisms). \downarrow Let $g \in \{ai, le, lc\}$, and $(\mathbf{U}_P, \mathbf{S}_P) \in (\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}, \oplus_{psd,g}, \otimes_{psd,g})$. The map $\text{hom}_{psd,g}(\cdot) : (\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}, \oplus_{psd,g}, \otimes_{psd,g}) \rightarrow (\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}, \oplus_{psd,g}, \otimes_{psd,g})$ is a gyro homomorphism defined by

$$\text{hom}_{psd,g}(\mathbf{U}_P, \mathbf{S}_P) = (\text{hom}_{gr}(\mathbf{U}_P), \text{hom}_g(\mathbf{S}_P)). \quad (10)$$

For aggregation, we use the WFM induced by the product geometry, as detailed in App. D.2.4. The bias and non-linear activation are also formulated in the product space:

$$\phi_{psd}(\mathbf{U}_{R_i}, \mathbf{S}_{R_i}) = (\mathbf{B}_{gr} \tilde{\oplus}_{gr} \mathbf{U}_{R_i}, (\mathbf{B}_{spd} \oplus_{spd} \mathbf{S}_{R_i})^p), \quad (11)$$

where $\mathbf{B}_{gr} \in \tilde{\mathcal{G}}(q, d)$ and $\mathbf{B}_{spd} \in \mathcal{S}_d^{++}$.

In summary, our GyroAtt framework comprises several basic operations. First, the mapping $\text{hom}(\cdot)$ generates \mathbf{Q}_i , \mathbf{K}_i , and \mathbf{V}_i . Then, the attention score is computed using gyro distances between queries and keys, followed by the WFM-based aggregation on the values \mathbf{V}_i . Finally, the bias and non-linear activation enhance model expressiveness. GyroAtt offers a unified framework that supports multiple manifolds and metrics, demonstrating its superior generality and flexibility. Tab. 3 summarizes all the key ingredients for computing GyroAtt on the SPD, Grassmannian, and SPSD manifolds.

5 Experiments

Following prior works [55, 41], we evaluate the proposed GyroAtt on EEG decoding across four benchmarking datasets: BNCI2014001 [24], BNCI2015001 [64], MAMEM-SSVEP-II [54], and BCI-ERN [47] in this paper. For BNCI2014001 and BNCI2015001, we conduct both inter-session and inter-subject evaluations. In the inter-session evaluation, models are trained exclusively on data from the target subject. Additionally, the balanced accuracy, calculated as the average recall across all classes, is used as the primary metric [41]. For MAMEM-SSVEP-II, the overall accuracy is used, while for BCI-ERN, we report the Area Under the Curve (AUC) to address class imbalance. In the experiments, the first four sessions of each subject are used for training, with one for validation, and the fifth for testing. Further details regarding datasets and precessings are provided in App. E.1.

Implementation details. As shown in Fig. 1, the GyroAtt network consists of three components: a feature extraction module, a gyro attention module, and a classification head. In the feature extraction module, two convolutional blocks are first applied to the EEG signals to extract low-redundancy features, followed by pyramid-like temporal segmentation into s non-overlapping subparts, each producing a covariance matrix. For GyroAtt-SPD, these covariance matrices \mathbf{X}_i serve directly as inputs to the subsequent layers. In GyroAtt-SPSD and GyroAtt-Gr, each covariance matrix is transformed into its canonical form $(\mathbf{U}_X^i, \mathbf{S}_X^i)$ using Alg. 3, mapping them into the structure space $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}$. Here, \mathbf{U}_X^i is used as the input for GyroAtt-Gr, while both \mathbf{U}_X^i and \mathbf{S}_X^i are used

Table 4: The results of different methods on BNCI2014001 and BNCI2015001. Riemannian manifold attention models are highlighted in yellow. The top three results are marked in **red**, **blue**, and **cyan**.

Manifold	Method	BNCI2014001		BNCI2015001	
		Inter-session	Inter-subject	Inter-session	Inter-subject
Euclidean	FBCSP+SVM [4]	60.6 \pm 4.9	32.3 \pm 7.3	81.5 \pm 4.4	58.6 \pm 13.4
	EEGNet [44]	41.8 \pm 5.8	43.3 \pm 17.0	72.4 \pm 8.4	59.2 \pm 9.5
	ShConvNet [59]	51.3 \pm 2.3	42.2 \pm 16.2	74.1 \pm 4.2	58.7 \pm 5.8
	FBCSP+DSS+LDA [29]	71.3 \pm 1.8	48.3 \pm 14.3	84.6 \pm 4.8	67.7 \pm 14.3
SPD	TSM+SVM [6]	61.8 \pm 4.1	34.7 \pm 8.6	75.7 \pm 5.1	56.0 \pm 6.0
	FB+TSM+LR [42]	69.8 \pm 4.8	36.5 \pm 8.2	80.9 \pm 6.0	60.6 \pm 10.9
	URPA+MDM [57]	59.5 \pm 2.7	46.8 \pm 14.6	79.2 \pm 4.6	70.3 \pm 16.1
	SPDOT+TSM+SVM [74]	66.8 \pm 3.8	38.6 \pm 8.6	77.5 \pm 2.9	63.3 \pm 8.1
	TSMNet [41]	69.0 \pm 3.6	51.6 \pm 16.5	85.8 \pm 4.3	77.0 \pm 13.7
	Graph-CSPNet [36]	71.9 \pm 13.3	45.2 \pm 9.3	79.8 \pm 14.6	64.2 \pm 13.4
SPD Grassmann	MAtt [55]	66.5 \pm 8.9	45.3 \pm 11.3	80.8 \pm 14.8	63.1 \pm 10.1
	GDLNet [72]	58.1 \pm 8.9	46.3 \pm 5.1	76.9 \pm 13.6	63.3 \pm 14.2
SPD	GyroAtt-SPD-AIM	75.4 \pm 7.1	53.1 \pm 14.8	86.2 \pm 4.5	77.9 \pm 13.0
	GyroAtt-SPD-LEM	75.3 \pm 6.5	52.3 \pm 14.1	85.7 \pm 5.5	76.6 \pm 13.7
	GyroAtt-SPD-LCM	74.2 \pm 7.8	52.4 \pm 15.6	84.7 \pm 6.6	75.5 \pm 13.8
Grassmann	GyroAtt-Gr	72.5 \pm 7.3	52.1 \pm 14.2	85.0 \pm 7.7	75.3 \pm 13.7
SPSD	GyroAtt-SPSD-AIM	72.9 \pm 7.1	52.4 \pm 15.6	84.7 \pm 6.6	75.5 \pm 13.8
	GyroAtt-SPSD-LEM	72.8 \pm 6.9	50.5 \pm 13.2	85.3 \pm 5.3	76.0 \pm 14.1
	GyroAtt-SPSD-LCM	72.9 \pm 6.7	51.7 \pm 13.1	85.1 \pm 4.8	74.9 \pm 12.6

for GyroAtt-SPSD. We employ the corresponding GyroAtt block, as shown in Alg. 4, to capture long-range dependencies between different feature regions on the manifolds. In the classification module, we first flatten the manifold-valued representations by projecting them onto a tangent (or Euclidean) space, followed by vectorization. For GyroAtt-SPD, we follow previous works [71, 16] to apply matrix power normalization to the output matrix \mathbf{P} from the GyroAtt block, defined as $\psi_\theta(\mathbf{P}) = \frac{1}{\theta} \mathbf{P}^\theta$ with $\theta > 0$ and $\mathbf{P} \in \mathcal{S}_d^{++}$. The scaling factor $\frac{1}{\theta}$ ensures gradient stability during optimization. For GyroAtt-Gr, we project each element $\mathbf{Y}_i \in \mathcal{G}(q, d)$ into a Euclidean space using the operator $\Phi(\mathbf{Y}_i) = \mathbf{Y}_i \mathbf{Y}_i^\top$. For GyroAtt-SPSD, both \mathbf{U}_X^i and \mathbf{S}_X^i are processed accordingly within the classification module. Across all three models, the resulting matrices are vectorized, concatenated, and passed through a fully connected layer followed by a Softmax function for classification. For parameter optimization and detailed implementations, please refer to Apps. E.2.1 and E.2.2.

Main results. We evaluated the proposed GyroAtt framework on four EEG classification datasets, with 10-fold cross-validation results summarized in Tabs. 4 and 5. Our models—GyroAtt-SPD/Gr/SPSD are compared against state-of-the-art baselines. The most effective manifold choice in GyroAtt varies by dataset. To be specific, GyroAtt-SPSD-LCM achieves best accuracy on the MAMEM and BCI-ERN datasets, surpassing GDLNet by **3.2%** and **0.9%**, respectively. On the BNCI2014001 and BNCI2015001 datasets, GyroAtt-SPD-AIM delivers the highest scores, outperforming TSMNet by **6.4%**, **1.5%**, **0.4%**, and **0.9%**, respectively across different evaluation settings. Notably, GyroAtt-Gr consistently ranks second across all datasets and outperforms GDLNet in every case. These experimental observations highlight the generality and effectiveness of our GyroAtt. The superior performance of GyroAtt can be attributed to the underlying geometry—

Table 5: Results on the MAMEM-SSVEP-II (MAMEM) and BCI-ERN datasets. Manifold attention models are highlighted in yellow. The top three results are marked in **red**, **blue**, and **cyan**.

Manifold	Method	MAMEM	BCI-ERN
Euclidean	EEGNet [44]	53.7 \pm 7.2	74.3 \pm 2.5
	ShallowCNet [59]	56.9 \pm 6.7	71.9 \pm 2.6
	EEG-TCNet [35]	55.5 \pm 7.7	77.1 \pm 2.5
	FBCNet [46]	53.1 \pm 5.7	60.5 \pm 3.1
	TCNet-Fusion [48]	45.0 \pm 6.6	70.5 \pm 2.9
	MBEEGSE [3]	56.5 \pm 7.3	75.5 \pm 2.3
SPD Grassmann	MAtt [55]	65.2 \pm 3.1	75.7 \pm 2.2
	GDLNet [72]	65.5 \pm 2.9	78.2 \pm 2.5
SPD	GyroAtt-SPD-AIM	66.3 \pm 2.2	75.8 \pm 3.3
	GyroAtt-SPD-LEM	66.2 \pm 2.5	76.1 \pm 4.2
	GyroAtt-SPD-LCM	65.1 \pm 2.5	75.4 \pm 3.7
Grassmann	GyroAtt-Gr	67.1 \pm 1.6	78.4 \pm 1.4
SPSD	GyroAtt-SPSD-AIM	66.5 \pm 2.3	78.2 \pm 1.9
	GyroAtt-SPSD-LEM	66.5 \pm 2.9	79.1 \pm 1.7
	GyroAtt-SPSD-LCM	68.7 \pm 1.5	78.4 \pm 1.6

Table 6: The impact of matrix power activation on the accuracy of GyroAtt-SPD/SPSD.

Geometry	Activation	BNCI2014001		BNCI2015001		MAMEM
		Inter-session	Inter-subject	Inter-session	Inter-subject	
SPD	w/o	74.8 \pm 6.7	51.2 \pm 15.7	85.5 \pm 5.0	75.4 \pm 12.9	64.3 \pm 2.4
	w/	75.4 \pm 7.1	53.1 \pm 14.8	86.2 \pm 4.5	77.9 \pm 13.0	66.3 \pm 2.2
SPSD	w/o	72.2 \pm 7.2	49.2 \pm 13.7	83.9 \pm 5.1	74.2 \pm 14.4	66.5 \pm 2.4
	w/	72.9 \pm 7.1	52.4 \pm 15.6	85.3 \pm 5.3	76.0 \pm 14.1	68.7 \pm 1.5

Table 8: Replacing $\text{hom}(\cdot)$ and power activation with BiMap and ReEig in GyroAtt-SPD-LEM

Transformation	Activation	BNCI2014001		BNCI2015001	
		Inter-session	Inter-subject	Inter-session	Inter-subject
BiMap	Power	74.0 \pm 6.5	52.3 \pm 15.0	85.2 \pm 7.2	77.2 \pm 13.2
Homomorphisms	ReEig	75.1 \pm 6.3	52.6 \pm 14.2	85.9 \pm 5.3	76.4 \pm 12.8
BiMap	ReEig	73.6 \pm 6.8	52.2 \pm 15.2	85.4 \pm 7.8	76.8 \pm 13.0
Homomorphisms	Power	75.4 \pm 7.1	53.1 \pm 14.8	86.2 \pm 4.5	77.9 \pm 13.0

aware attention mechanism, which effectively captures long-range dependencies and spatiotemporal fluctuations inherent in EEG signals.

Ablations on the matrix power-based nonlinear activation $\sigma(\cdot)$. Tab. 6 illustrates the impact of the nonlinear activation (as defined in Tab. 3) on the accuracy of GyroAtt-SPD/SPSD. Generally, removing nonlinear activation results in performance degradation, *e.g.*, GyroAtt experiences a 2.0% and 2.4% accuracy drop on the MAMEM dataset. These experimental findings emphasize the importance of matrix power activation in enhancing model expressiveness by introducing nonlinearity into the metric space of the underlying feature manifold.

Gyro distance vs. gyro inner product. The Euclidean inner product is a natural similarity metric in flat space, but a global inner product is generally unavailable on curved manifolds. Hence, most manifold-attention models rely on distance-based similarity [26, 55, 72]. Gyrovector spaces are an exception because a *gyro inner product* can be defined in the tangent space at the identity [52][Defs. 2.9, 2.15]: $\langle \mathbf{P}, \mathbf{Q} \rangle^{\text{gyr}} = \langle \text{Log}_{\mathbf{I}}(\mathbf{P}), \text{Log}_{\mathbf{I}}(\mathbf{Q}) \rangle_{\mathbf{I}}$. To evaluate its suitability, we replaced the distance-based similarity in Eq. (5) with gyro inner product and conducted an ablation study, as reported in Tab. 7. Across all datasets, the gyro distance consistently outperforms the inner product in accuracy. An essential reason is that the gyro inner product is defined within a single tangent space, which may not reflect the geometric structure faithfully. In contrast, gyro distance better preserves the intrinsic manifold geometry. The detailed implementations are found in App. E.6.

GyroAtt-SPD-LEM vs. MAtt. Although both GyroAtt-SPD-LEM and MAtt [55] implement attention mechanisms under the SPD manifold with LEM, GyroAtt-SPD-LEM differs from MAtt in two key components: linear transformation $\text{hom}(\cdot)$, and the nonlinear activation $\sigma(\cdot)$. While BiMap and ReEig used in MAtt preserve SPD property, they are loosely connected to the underlying Riemannian geometry. Hence, we replace $\text{hom}(\cdot)$ and $\sigma(\cdot)$ with BiMap and ReEig and made an ablation study, as reported in Tab. 8. Either replacement leads to a performance drop, while combining both further degrades accuracy. From a theoretical standpoint, gyro homomorphisms generalize linear maps and naturally reduce to them in the Euclidean setting. This indicates that our transformation layer is a principled extension of Euclidean attention. In contrast, the ReEig function merely rescales eigenvalues to ensure positive definiteness, lacking geometric interpretability. As shown in [17][Fig. 1], matrix power can deform any given Riemannian metric, effectively acting as a geometry-aware activation. These findings confirm that integrating gyro-based transformations and nonlinearities is not only theoretically justified but also empirically beneficial.

Table 7: Accuracy comparison under different measures.

Methods	Similarity	BNCI2014001		MAMEM
		Inter-session	Inter-subject	
GyroAtt-SPD	Gyro inner product	74.7 \pm 6.8	85.6 \pm 5.4	63.9 \pm 3.2
	Gyro distance	75.4 \pm 7.1	86.2 \pm 4.5	66.3 \pm 2.2
GyroAtt-Gr	Gyro inner product	72.4 \pm 7.3	83.4 \pm 5.9	65.7 \pm 3.1
	Gyro distance	72.5 \pm 7.3	85.0 \pm 7.7	67.1 \pm 1.6
GyroAtt-SPSD	Gyro inner product	71.6 \pm 6.3	83.3 \pm 5.4	65.0 \pm 2.6
	Gyro distance	72.9 \pm 6.2	85.3 \pm 5.3	68.7 \pm 1.5

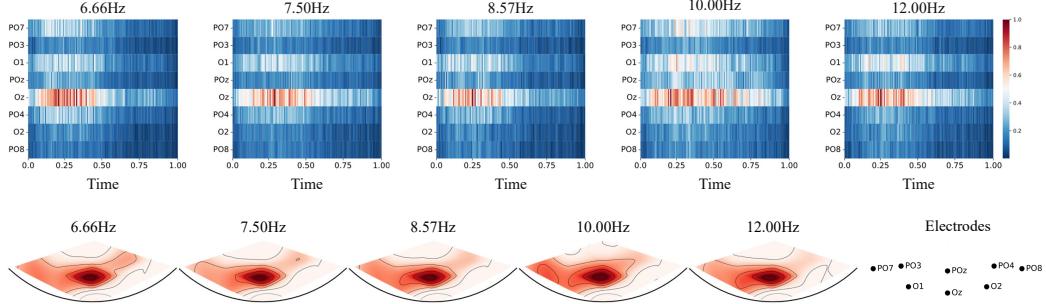


Figure 2: Spatial topomaps and gradient heatmaps of GyroAtt-SPSD on the MAMEM (subject S11) dataset. Strong responses appear around the Oz electrode, especially within 0–0.5s. In heatmaps, the x-axis denotes time and the y-axis denotes EEG channels.

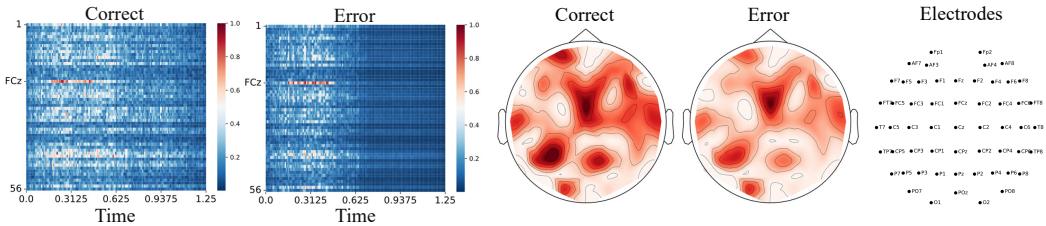


Figure 3: Spatial topomaps and gradient heatmaps of GyroAtt-SPSD, showing ‘correct’ vs. ‘error’ trials. Strong activations consistently emerge around the FCz electrode within 0.1–0.4s.

Training efficiency. GyroAtt demonstrates superior or comparable efficiency to existing manifold attention networks such as MAtt and GDLNet. In particular, GyroAtt-SPD under LCM achieves efficiency on par with MAtt, benefiting from the computational advantages of the Cholesky decomposition. More detailed discussions and comparisons are provided in App. E.8.

Table 9: Running time (s/epoch) comparison.

Method	BCI2014001		BCI2015001	
	Inter-session	Inter-subject	Inter-session	Inter-subject
MAtt	4.86	89.12	2.74	56.78
GDLNet	4.55	88.59	1.71	47.66
GyroAtt-SPD-LCM	4.11	87.44	2.42	49.86

EEG interpretability analysis. For the MAMEM dataset, as shown in Fig. 2, GyroAtt shows strong gradient responses around the Oz electrode across five stimulus frequencies, especially within 0–0.5s time window. This pattern aligns well with the established findings on the relationship between SSVEP signals and the Oz region [31, 28]. This is likely due to the Oz central location in the primary visual cortex, which leads to stronger evoked potentials and signal-to-noise ratio. On the BCI-ERN dataset, as illustrated in Fig. 3, gradient responses for both ‘correct’ and ‘error’ trials centered around the FCz electrode. This observation aligns with substantial empirical evidence that the anterior cingulate cortex, a central medial prefrontal cortex region connected to limbic and frontal areas, underlies ERN generation. Notably, consistent gradient responses around FCz were observed for both feedback types within the 0.1–0.4s interval, reinforcing Event-Related Potential (ERP) waveform differences between correct and error trials as reported in [27].

6 Conclusion

In this paper, we propose GyroAtt, a principled framework that generalizes the Euclidean attention mechanism to gyrovector spaces. Specifically, we adopt gyro homomorphism, gyro distance-based attention, and WFM as counterparts to the transformation, attention, and aggregation operations in Euclidean attention. Notably, we identify the concrete non-trivial expressions of gyro homomorphisms on different matrix gyro spaces. The principled construction of GyroAtt enables a direct assessment of the impact of geometry on a given task while keeping the network architecture constant. Extensive experiments and ablation studies on four EEG benchmarking datasets certify the effectiveness and flexibility of our proposed framework.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (62306127, 62020106012, 62332008), the Natural Science Foundation of Jiangsu Province (BK20231040, BK20221535), the Fundamental Research Funds for the Central Universities (JUSRP124015), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX25_1319), the Key Project of Wuxi Municipal Health Commission (Z202318), the EU Horizon projects ELIAS (No. 101120237), the ELLIOT (No. 101214398), and the National Key R&D Program of China (2023YFF1105102, 2023YFF1105105).

References

- [1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80:199–220, 2004.
- [2] Bijan Afsari. Riemannian l^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 2011.
- [3] Ghadir Ali Altuwaijri, Ghulam Muhammad, Hamdi Altaheri, and Mansour Alsulaiman. A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification. *Diagnostics*, page 995, 2022.
- [4] Kai Keng Ang, Zheng Yang Chin, Haihong Zhang, and Cuntai Guan. Filter bank common spatial pattern (FBCSP) in brain-computer interface. In *IJCNN*, pages 2390–2397. IEEE, 2008.
- [5] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. *Fast and simple computations on tensors with log-Euclidean metrics*. PhD thesis, INRIA, 2005.
- [6] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain-computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2011.
- [7] Ahmad Bdeir, Kristian Schwegel, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.
- [8] Thomas Bendokat, Ralf Zimmermann, and P-A Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Adv. Comput. Math.*, pages 1–51, 2024.
- [9] Silvere Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070, 2010.
- [10] Silvere Bonnabel, Anne Collard, and Rodolphe Sepulchre. Rank-preserving geometric means of positive semi-definite matrices. *Linear Algebra and its Applications*, 438(8):3202–3216, 2013.
- [11] Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *NeurIPS*, pages 15463–15474, 2019.
- [12] 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, 2020.
- [13] Ziheng Chen, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Zhiwu Huang, and Josef Kittler. Riemannian local mechanism for SPD neural networks. In *AAAI*, 2023.
- [14] Ziheng Chen, Yue Song, Gaowen Liu, Ramana Rao Kompella, Xiao-Jun Wu, and Nicu Sebe. Riemannian multinomial logistics regression for SPD neural networks. In *CVPR*, 2024.
- [15] Ziheng Chen, Yue Song, Yunmei Liu, and Nicu Sebe. A Lie group approach to Riemannian batch normalization. In *ICLR*, 2024.

[16] Ziheng Chen, Yue Song, Xiao-Jun Wu, Gaowen Liu, and Nicu Sebe. Understanding matrix function normalizations in covariance pooling through the lens of Riemannian geometry. *arXiv preprint arXiv:2407.10484*, 2024.

[17] Ziheng Chen, Yue Song, Xiaojun, and Nicu Sebe. RMLR: Extending multinomial logistic regression into general geometries. *arXiv preprint arXiv:2409.19433*, 2024.

[18] Ziheng Chen, Yue Song, Tianyang Xu, Zhiwu Huang, Xiao-Jun Wu, and Nicu Sebe. Adaptive Log-Euclidean metrics for SPD matrix learning. *IEEE TIP*, 2024.

[19] Ziheng Chen, Yue Song, Xiaojun Wu, and Nicu Sebe. Gyrogroup batch normalization. In *ICLR*, 2025.

[20] Calin Cruceru, Gary Bécigneul, and Octavian-Eugen Ganea. Computationally Tractable Riemannian Manifolds for Graph Embeddings. In *AAAI*, pages 7133–7141, 2021.

[21] Thibault de Surrel, Sylvain Chevallier, Fabien Lotte, and Florian Yger. Geometry-aware visualization of high dimensional symmetric positive definite matrices. *Transactions on Machine Learning Research Journal*, 2025.

[22] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[24] Josef Faller, Carmen Vidaurre, Teodoro Solis-Escalante, Christa Neuper, and Reinhold Scherer. Autocalibration and recurrent adaptation: towards a plug and play online ERD-BCI. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 20(3):313–319, 2012.

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

[26] Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018.

[27] Greg Hajcak. What we've learned from mistakes: Insights from error-related brain activity. *Curr. Dir. Psychol.*, 2012.

[28] Chengcheng Han, Guanghua Xu, Jun Xie, Chaoyang Chen, and Sicong Zhang. Highly interactive brain-computer interface based on flicker-free steady-state motion visual evoked potential. *Sci. Rep.*, 2018.

[29] Lea Hehenberger, Reinmar J Kobler, Catarina Lopes-Dias, Nitikorn Srisrisawang, Peter Tumhart, John B Uroko, Paul R Torke, and Gernot R Müller-Putz. Long-term mutual training for the CYBATHLON BCI race with a tetraplegic pilot: a case study on inter-session transfer and intra-session adaptation. *Frontiers in Human Neuroscience*, 15:635777, 2021.

[30] Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International Conference on Machine Learning*, pages 1969–1978. PMLR, 2018.

[31] Christoph S Herrmann. Human EEG responses to 1–100 hz flicker: resonance phenomena in visual cortex and their potential correlation to cognitive phenomena. *Exp. Brain Res.*, 2001.

[32] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[33] Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. In *AAAI*, pages 2036–2042, 2017.

[34] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on Grassmann manifolds. In *AAAI*, pages 1137–1145, 2018.

[35] Thorir Mar Ingolfsson, Michael Hersche, Xiaying Wang, Nobuaki Kobayashi, Lukas Cavigelli, and Luca Benini. EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces. In *SMC*, pages 2958–2965, 2020.

[36] 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*, pages 1–15, 2023. doi: 10.1109/TNNLS.2023.3307470.

[37] Ce Ju, Reinmar J Kobler, Liyao Tang, Cuntai Guan, and Motoaki Kawanabe. Deep geodesic canonical correlation analysis for covariance-based neuroimaging data. In *The Twelfth International Conference on Learning Representations*, 2024.

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

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

[40] Sejong Kim. Ordered Gyrovector Spaces. *Symmetry*, 12(6), 2020.

[41] 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.

[42] Reinmar J Kobler, Jun-Ichiro Hirayama, Lea Hehenberger, Catarina Lopes-Dias, Gernot R Müller-Putz, and Motoaki Kawanabe. On the interpretation of linear Riemannian tangent space model parameters in M/EEG. In *EMBC*, pages 5909–5913. IEEE, 2021.

[43] Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian Optimization in PyTorch, 2020. [_eprint: 2005.02819](#).

[44] 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. *J. Neural Eng.*, page 056013, 2018.

[45] Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.

[46] Ravikiran Mane, Effie Chew, Karen Chua, Kai Keng Ang, Neethu Robinson, A Prasad Vinod, Seong-Whan Lee, and Cuntai Guan. FBCNet: A multi-view convolutional neural network for brain-computer interface. *arXiv preprint arXiv:2104.01233*, 2021.

[47] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during P300-based spelling. *Adv. Hum-Comput. Interact.*, pages 1–13, 2012.

[48] Yazeed K Musallam, Nasser I AlFassam, Ghulam Muhammad, Syed Umar Amin, Mansour Alsulaiman, Wadood Abdul, Hamdi Altaheri, Mohamed A Bencherif, and Mohammed Al-gabri. Electroencephalography-based motor imagery classification using temporal convolutional network fusion. *Biomed. Signal Process. Control*, page 102826, 2021.

[49] Galileo Namata, Ben London, Lise Getoor, Bert Huang, and U Edu. Query-driven active surveying for collective classification. In *10th international workshop on mining and learning with graphs*, 2012.

[50] Xuan Son Nguyen. A Gyrovector Space Approach for Symmetric Positive Semi-definite Matrix Learning. In *ECCV*, pages 52–68, 2022.

[51] Xuan Son Nguyen. The Gyro-Structure of Some Matrix Manifolds. In *NeurIPS*, pages 26618–26630, 2022.

[52] Xuan Son Nguyen and Shuo Yang. Building Neural Networks on Matrix Manifolds: A Gyrovector Space Approach. In *ICML*, volume abs/2305.04560, 2023.

[53] Xuan Son Nguyen, Shuo Yang, and Aymeric Histace. Matrix manifold neural networks++. *arXiv preprint arXiv:2405.19206*, 2024.

[54] Spiros Nikolopoulos. MAMEM EEG SSVEP Dataset II (256 channels, 11 subjects, 5 frequencies presented simultaneously). Dataset, 2016. URL <https://api.semanticscholar.org/CorpusID:63262529>.

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

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

[57] Pedro Luiz Coelho Rodrigues, Christian Jutten, and Marco Congedo. Riemannian procrustes analysis: transfer learning for brain–computer interfaces. *IEEE Transactions on Biomedical Engineering*, 66(8):2390–2401, 2018.

[58] Takashi Sakai. *Riemannian geometry*, volume 149. American Mathematical Soc., 1996.

[59] Robin Tibor Schirrmeyer, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.*, pages 5391–5420, 2017.

[60] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 2008.

[61] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. *arXiv preprint arXiv:2006.08210*, 2020.

[62] Anuj Srivastava and Eric Klassen. Bayesian and geometric subspace tracking. *Adv. Appl. Prob.*, pages 43–56, 2004.

[63] Teerapong Suksumran and Keng Wibootton. Isomorphism theorems for gyrogroups and l-subgyrogroups. *arXiv e-prints*, pages arXiv–1406, 2014.

[64] Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot Müller-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.

[65] William P HG Thurston. *Three-Dimensional Geometry and Topology, Volume 1: Volume 1*. Princeton university press, 1997.

[66] Abraham A Ungar. *Analytic hyperbolic geometry: Mathematical foundations and applications*. World Scientific, 2005.

[67] Abraham Albert Ungar. *Beyond the Einstein Addition Law and Its Gyroscopic Thomas Precession: The Theory of Gyrogroups and Gyrovector Spaces*. Fundamental Theories of Physics, vol. 117, Springer, Netherlands, 2002.

[68] Abraham Albert Ungar. *Analytic Hyperbolic Geometry: Mathematical Foundations and Applications*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.

[69] Abraham Albert Ungar. *Analytic Hyperbolic Geometry in N Dimensions: An Introduction*. CRC Press, 2014.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[71] Qilong Wang, Jiangtao Xie, Wangmeng Zuo, Lei Zhang, and Peihua Li. Deep CNNs meet global covariance pooling: Better representation and generalization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2582–2597, 2020.

- [72] Rui Wang, Chen Hu, Ziheng Chen, Xiao-Jun Wu, and Xiaoning Song. A Grassmannian manifold self-attention network for signal classification. In *IJCAI*, 2024.
- [73] Rui Wang, Xiao-Jun Wu, Ziheng Chen, Cong Hu, and Josef Kittler. SPD manifold deep metric learning for image set classification. *IEEE TNNLS*, 2024.
- [74] Or Yair, Felix Dietrich, Ronen Talmon, and Ioannis G Kevrekidis. Domain adaptation with optimal transport on the manifold of spd matrices. *arXiv preprint arXiv:1906.00616*, 2019.
- [75] Wei Zhao, Federico Lopez, J Maxwell Riestenberg, Michael Strube, Diaaeldin Taha, and Steve Trettel. Modeling graphs beyond hyperbolic: Graph neural networks in symmetric positive definite matrices. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2023.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: **[Yes]**

Justification: The main claims in the abstract and introduction are consistent with the theoretical contributions and experimental validations.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: **[Yes]**

Justification: We discuss limitations in App. A.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full assumptions and formal proofs for Thm. 4.1, Thm. 4.2, and Thm. 4.3 are provided in App. F.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide sufficient details in Sec. 5 and App. E to reproduce all main experiments, including datasets, model settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: We will release the code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We describe all necessary training details in Sec. 5. Additional implementation details are provided in App. E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report mean \pm standard deviation over 10 independent runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments were conducted on a machine with an Intel i9-14900 CPU, 64GB RAM, and two NVIDIA RTX 4080 Super GPUs. Further details are provided in App. E.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed and adhered to the NeurIPS Code of Ethics. Our research involves publicly available datasets and does not pose foreseeable risks related to privacy, safety.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work focuses on foundational research and is not tied to particular applications. Therefore, societal impacts are not directly applicable at this stage.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not involve pretrained language models, generative models, or scraped datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: **[Yes]**

Justification: All datasets and code used in this work are publicly available and properly cited in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: **[NA]**

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix Contents

A Limitations	24
B Notations	24
C Abbreviations	24
D Preliminaries	25
D.1 Gyrogroups and gyrovector spaces	25
D.2 Weighted Fréchet Mean	25
D.2.1 Existence and uniqueness of the Weighted Fréchet Mean	25
D.2.2 Weighted Fréchet Mean on SPD manifolds	26
D.2.3 Weighted Fréchet Mean on Grassmannian manifolds	27
D.2.4 Weighted Fréchet Mean on PSD manifolds	27
D.3 Canonical representation in PSD manifolds	28
E Implementation details and additional experiments	28
E.1 Datasets and EEG preprocessing	28
E.2 Additional implementation details	30
E.2.1 Network architectures	30
E.2.2 Optimization	31
E.3 Statistical Significance Analysis	31
E.4 Graph node classification and radar recognition tasks	32
E.5 Additional ablations studies	32
E.5.1 Ablations on the Riemannian metrics and matrix power-based nonlinear activation	32
E.5.2 Ablations on the matrix power normalization	33
E.5.3 Ablation on the number of GyroAtt blocks	34
E.6 Implementation details of gyro inner product	34
E.7 Implementation details of replacing geometric components in GyroAtt	34
E.8 Computational complexity analysis and comparison	35
E.8.1 Number of matrix functions in manifold attention mechanisms	35
E.8.2 Experimental comparison of time efficiency	36
E.9 Model compactness and parameter efficiency	37
E.10 Long-range dependencies	37
E.11 Visualizing learned representations via Rie-SNE	37

F Proofs of the Theorems in the Main Paper	37
F.1 Proof of the Thm. 3.2	37
F.2 Proof of the Thm. 4.1	38
F.3 Proof of the Thm. 4.2	40
F.4 Proof of the Thm. 4.3	41

A Limitations

While several widely used manifolds, such as hyperbolic, SPD, Grassmannian, and SPSD manifolds, admit gyrovector structures, this property does not hold universally across all manifolds. As a result, our method is not directly applicable to manifolds lacking such a structure. Extending GyroAtt to accommodate more general non-gyrovector manifolds constitutes a promising direction for future research.

B Notations

To enhance readability, we summarize the main mathematical symbols and operators used throughout this paper in Tab. 10.

Table 10: Summary of notations.

Notations	Explanation
(G, \oplus)	A gyrogroup G with a binary operation \oplus
\mathcal{S}_d^{++}	Space of $d \times d$ SPD matrices
\mathcal{S}^d	Space of $d \times d$ symmetric matrices
$\mathcal{S}_{d,q}^+$	Space of $d \times d$ SPSD matrices with rank $q \leq d$
$\mathcal{G}(q, d)$	Grassmannian in the projector perspective
$\tilde{\mathcal{G}}(q, d)$	Grassmannian in the ONB perspective
$\oplus_{ai}, \ominus_{ai}, \otimes_{ai}$	Binary, inverse, and scalar multiplication operations in \mathcal{S}_d^{++} under AIM
$\oplus_{le}, \ominus_{le}, \otimes_{le}$	Binary, inverse, and scalar multiplication operations in \mathcal{S}_d^{++} under LEM
$\oplus_{lc}, \ominus_{lc}, \otimes_{lc}$	Binary, inverse, and scalar multiplication operations in \mathcal{S}_d^{++} under LCM
$\tilde{\oplus}_{gr}, \tilde{\ominus}_{gr}, \tilde{\otimes}_{gr}$	Binary, inverse, and scalar multiplication operations in $\tilde{\mathcal{G}}(q, d)$
$\oplus_{gr}, \ominus_{gr}, \otimes_{gr}$	Binary, inverse, and scalar multiplication operations in $\mathcal{G}(q, d)$
$\oplus_{psd,g}, \ominus_{psd,g}, \otimes_{psd,g}$	Binary, inverse, and scalar multiplication operations in $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_d^{++}$ under metrics g
$\langle \mathbf{P}, \mathbf{Q} \rangle^g$	Inner product in \mathcal{S}_d^{++} under metrics g
$\langle \mathbf{U}, \mathbf{V} \rangle^{gr}$	Inner product in $\tilde{\mathcal{G}}(q, d)$
$\langle (\mathbf{U}_P, \mathbf{S}_P), (\mathbf{U}_Q, \mathbf{S}_Q) \rangle^{psd,g}$	Inner product in $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_d^{++}$ under metrics g
$\ \ominus_g \mathbf{P} \oplus_g \mathbf{Q}\ _g^{spd}$	the gyrodistance in \mathcal{S}_d^{++} under metrics g
$\ \tilde{\ominus}_{gr} \mathbf{U} \tilde{\oplus}_{gr} \mathbf{V}\ ^{gr}$	the gyrodistance in $\tilde{\mathcal{G}}(q, d)$
$\ (\tilde{\ominus}_{gr} \mathbf{U}_P \tilde{\oplus}_{gr} \mathbf{U}_Q, \ominus_g \mathbf{S}_P \oplus_g \mathbf{S}_Q)\ _{psd}^g$	the gyrodistance in $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_d^{++}$ under metrics g
$[\cdot, \cdot]$	the matrix commutator
$\text{expm}(\cdot), \text{logm}(\cdot)$	Matrix exponentiation and logarithm
$\mathcal{L}(\cdot), \mathcal{L}^{-1}(\cdot)$	Cholesky decomposition and its inverse
$\mathbb{D}(\cdot)$	A diagonal matrix with diagonal elements from a square matrix
$[\cdot]$	The strictly lower triangular part of a square matrix
$\text{Log}_{\mathbf{P}}^{gr}(\mathbf{Q})$	Logarithmic map of \mathbf{Q} at \mathbf{P} in $\mathcal{G}(q, d)$
\mathcal{M}, \mathcal{N}	Matrix manifold
WFM	the weighted Fréchet mean
$\text{hom}_{ai}(\cdot), \text{hom}_{le}(\cdot), \text{hom}_{lc}(\cdot)$	the maps in \mathcal{S}_d^{++} under AIM, LEM, and LCM satisfying gyro homomorphism
$\text{hom}_{gr}(\cdot)$	the maps in $\tilde{\mathcal{G}}(q, d)$ satisfying gyro homomorphism
$\text{hom}_{psd,g}(\cdot)$	the maps in $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_d^{++}$ under metrics g satisfying gyro homomorphism
$\ \cdot\ _{\mathbf{F}}$	The norm induced by the standard Frobenius inner product
$O(d)$	The special orthogonal group
$\text{Exp}_{\mathbf{P}}^{ai}(\mathbf{A})$	Exponential map of \mathbf{A} at \mathbf{P} in \mathcal{S}_d^{++} under AIM
$\text{Log}_{\mathbf{P}}^{ai}(\mathbf{Q})$	Logarithmic map of \mathbf{Q} at \mathbf{P} in \mathcal{S}_d^{++} under AIM
$\text{Exp}_{\mathbf{P}}^{gr}(\mathbf{W})$	Exponential map of \mathbf{W} at \mathbf{P} in $\mathcal{G}(q, d)$
$\tilde{\text{Exp}}_{\mathbf{P}}^{gr}(\mathbf{H})$	Exponential map of \mathbf{H} at \mathbf{P} in $\tilde{\mathcal{G}}(q, d)$
$\tilde{\text{Log}}_{\mathbf{P}}^{gr}(\mathbf{Q})$	Logarithmic map of \mathbf{Q} at \mathbf{P} in $\tilde{\mathcal{G}}(q, d)$

C Abbreviations

For completeness, the abbreviations appearing in this paper are listed below for easy reference.

List of Abbreviations

AIM	Affine-Invariant Metric	2, 3, 5, 6
AUC	Area Under the Curve	6

DNN	Deep Neural Network	1
EEG	Electroencephalography	1, 2, 6–9
ERP	Event-Related Potential	9
GyroAtt	Gyro Attention	1–7, 9
LCM	Log-Cholesky Metric	2, 3, 5, 6, 9
LEM	Log-Euclidean Metric	1–6
ONB	Orthonormal Basis	2–6
SPD	Symmetric Positive Definite	1–9
SPSD	Symmetric Positive Semi-Definite	1–9
WFM	Weighted Fréchet Mean	2–6

D Preliminaries

D.1 Gyrogroups and gyrovector spaces

Gyrogroups and gyrovector spaces generalize groups and vector spaces, offering a powerful framework to analyze non-Euclidean geometries. Below, we formally present their definitions.

Definition D.1 (Gyrogroups [69]). A gyrogroup is a generalization of groups. Let G be a nonempty set with a binary operation \oplus and an identity element $\mathbf{E} \in G$. A pair (G, \oplus) is a gyrogroup if it satisfies the following axioms:

- (G1) There exists an identity element $\mathbf{E} \in G$ such that for all $\mathbf{A} \in G$, $\mathbf{E} \oplus \mathbf{A} = \mathbf{A}$.
- (G2) For each $\mathbf{A} \in G$, there exists a left inverse $\ominus \mathbf{A} \in G$ satisfying $\ominus \mathbf{A} \oplus \mathbf{A} = \mathbf{E}$.
- (G3) For all $\mathbf{A}, \mathbf{B}, \mathbf{C} \in G$, there exists an automorphism $\text{gyr}[\mathbf{A}, \mathbf{B}](\cdot) : G \rightarrow G$, satisfying

$$\mathbf{A} \oplus (\mathbf{B} \oplus \mathbf{C}) = (\mathbf{A} \oplus \mathbf{B}) \oplus \text{gyr}[\mathbf{A}, \mathbf{B}](\mathbf{C}). \quad (12)$$

Here, the map $\text{gyr}[\mathbf{A}, \mathbf{B}](\cdot)$ is called the gyroautomorphism, or the gyration of G generated by \mathbf{A}, \mathbf{B} .

- (G4) For all $\mathbf{A}, \mathbf{B} \in G$, The map $\text{gyr}[\mathbf{A}, \mathbf{B}]$ generated by each \mathbf{A}, \mathbf{B} satisfies the left loop property: $\text{gyr}[\mathbf{A}, \mathbf{B}] = \text{gyr}[\mathbf{A} \oplus \mathbf{B}, \mathbf{B}]$.

Definition D.2 (Gyrocommutative Gyrogroups [69]). A gyrogroup (G, \oplus) is gyrocommutative if it satisfies the gyrocommutative law: $\mathbf{A} \oplus \mathbf{B} = \text{gyr}[\mathbf{A}, \mathbf{B}](\mathbf{B} \oplus \mathbf{A})$ for all $\mathbf{A}, \mathbf{B} \in G$.

The following definition of gyrovector spaces is derived from Nguyen [51, Def. 2.3], which is slightly different from in Ungar [69, Def. 3.2].

Definition D.3 (Gyrovector Spaces [51]). A gyrocommutative gyrogroup (G, \oplus) equipped with a scalar multiplication $\otimes : \mathbb{R} \times G \rightarrow G$ is a gyrovector space if the following axioms are satisfied:

- (V1) $1 \otimes \mathbf{A} = \mathbf{A}$, $0 \otimes \mathbf{A} = t \otimes \mathbf{E} = \mathbf{E}$, and $(-1) \otimes \mathbf{A} = \ominus \mathbf{A}$.
- (V2) $(s + t) \otimes \mathbf{A} = s \otimes \mathbf{A} \oplus t \otimes \mathbf{A}$.
- (V3) $(st) \otimes \mathbf{A} = s \otimes (t \otimes \mathbf{A})$.
- (V4) $\text{gyr}[\mathbf{A}, \mathbf{B}](t \otimes \mathbf{C}) = t \otimes \text{gyr}[\mathbf{A}, \mathbf{B}]\mathbf{C}$.
- (V5) $\text{gyr}[s \otimes \mathbf{A}, t \otimes \mathbf{A}] = \text{Id}$, where Id is the identity map.

D.2 Weighted Fréchet Mean

D.2.1 Existence and uniqueness of the Weighted Fréchet Mean

The WFM is a central tool in manifold-based learning for aggregating features that reside on non-Euclidean spaces. Formally, given a set of points $\{x_i\}$ on a Riemannian manifold \mathcal{M} with associated non-negative weights $\{w_i\}$ summing to one, the WFM is defined as the minimizer of the weighted sum of squared distances.

Table 12: Uniqueness and Solution of WFM.

Geometry	SPD-AIM	SPD-LEM	SPD-LCM	Grassmann	SPSD
Uniqueness of WFM Solution of WFM Reference	Global Karcher Flow [56]	Global Closed Form [5]	Global Closed Form [45]	Within $r < \frac{\pi}{4\sqrt{2}}$ Karcher Flow [2]	The Grassmann component within $r < \frac{\pi}{4\sqrt{2}}$ WFM _{spd} , WFM _{gr} [10]

Uniqueness. The uniqueness of WFM is generally guaranteed when all sample points $\{x_i\}$ lie within a geodesically convex hull of \mathcal{M} [2]. In our paper, we focus on several commonly used manifolds, including SPD manifolds with various metrics (AIM, LEM, LCM), the Grassmann manifold, and the SPSD manifold. As summarized in Tab. 12, the WFM is globally unique for SPD manifolds under AIM, LEM, and LCM. For the Grassmann manifold, uniqueness holds when the radius r of the geodesic ball satisfies $r < \frac{\pi}{4\sqrt{2}}$ [2]. The same condition applies to the Grassmann component of the SPSD manifold under its canonical form. Following prior work [12, Sec. 2], we assume that the WFM is well-defined in our settings.

Computation. In terms of solution methods, the WFM can be computed via iterative optimization (e.g., Karcher flow) or, in some cases, closed-form expressions. For instance, SPD-LEM and SPD-LCM permit closed-form solutions [5, 45], while SPD-AIM and Grassmann manifolds typically require gradient-based optimization [56, 2, 39]. For SPSD, we follow the product formulation in [10], where the WFM is computed by combining the SPD and Grassmann components.

D.2.2 Weighted Fréchet Mean on SPD manifolds

Algorithm 1: Karcher Flow Algorithm on the SPD Manifold under AIM

Input : A set of SPD matrices $\mathbf{X}_{1\dots N} \in \mathcal{S}_d^{++}$
 A set of weights $w_{1\dots N} > 0$ with $\sum_i w_i = 1$
 Number of iterations K
Output : The WFM $\mathbf{G}_k \in \mathcal{S}_d^{++}$
 Initialize $\mathbf{G}_0 = \mathbf{I}$
for $k \leftarrow 1$ **to** K **do**
 | $\mathbf{G}_k \leftarrow \text{Exp}_{\mathbf{G}_{k-1}}^{ai} \left(\sum_{i=1}^N w_i \text{Log}_{\mathbf{G}_{k-1}}^{ai} (\mathbf{X}_i) \right)$
end

Affine-Invariant Metric. We begin by introducing the exponential and logarithmic maps under the affine-invariant metric (AIM), followed by the Karcher flow algorithm.

On the manifold \mathcal{S}_d^{++} endowed with AIM, the exponential map at a point $\mathbf{P} \in \mathcal{S}_d^{++}$ is given by [1]:

$$\text{Exp}_{\mathbf{P}}^{ai}(\mathbf{A}) = \mathbf{P}^{\frac{1}{2}} \exp \left(\mathbf{P}^{-\frac{1}{2}} \mathbf{A} \mathbf{P}^{-\frac{1}{2}} \right) \mathbf{P}^{\frac{1}{2}}, \quad (13)$$

where $\mathbf{A} \in T_{\mathbf{P}} \mathcal{S}_d^{++}$ is a tangent vector at \mathbf{P} . The logarithmic map, which is the inverse of the exponential map, is defined as

$$\text{Log}_{\mathbf{P}}^{ai}(\mathbf{Q}) = \mathbf{P}^{\frac{1}{2}} \log \left(\mathbf{P}^{-\frac{1}{2}} \mathbf{Q} \mathbf{P}^{-\frac{1}{2}} \right) \mathbf{P}^{\frac{1}{2}}, \quad (14)$$

for any $\mathbf{Q} \in \mathcal{S}_d^{++}$.

As shown in Alg. 1, 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 \mathbf{G}_{k-1} using the logarithmic map (Eq. (14)), a weighted average is calculated in this tangent space, and the result is mapped back to the manifold using the exponential map (Eq. (13)). This algorithm is guaranteed to converge on manifolds with non-positive curvatures, such as \mathcal{S}_d^{++} [38]. We initialize \mathbf{G}_0 as the identity matrix \mathbf{I} and set the number of iterations $K = 1$.

Log-Euclidean Metric. Under the log-Euclidean metric (LEM), the WFM has a closed-form expression provided by Chen et al. [15]:

$$\mathbf{G} = \exp \left(\sum_{i=1}^N w_i \text{logm}(\mathbf{X}_i) \right), \quad (15)$$

where $\mathbf{X}_{1\dots N} \in \mathcal{S}_d^{++}$, $w_{1\dots N} > 0$, and $\sum_i w_i = 1$.

Log-Cholesky Metric. Similarly, for the log-Cholesky metric (LCM), the WFM also admits a closed-form solution as shown by Chen et al. [15]:

$$\mathbf{G} = \mathcal{L}^{-1} \left(\sum_{i=1}^N w_i \lfloor \mathcal{L}(\mathbf{X}_i) \rfloor + \prod_{i=1}^N \mathbb{D}(\mathcal{L}(\mathbf{X}_i))^{w_i} \right), \quad (16)$$

where $\mathbf{X}_{1\dots N} \in \mathcal{S}_d^{++}$, $w_{1\dots N} > 0$, and $\sum_i w_i = 1$.

D.2.3 Weighted Fréchet Mean on Grassmannian manifolds

We now present the exponential and logarithmic maps, as well as the parallel translation under the ONB perspective, followed by the project perspective.

For the Grassmannian manifold $\tilde{\mathcal{G}}(q, d)$ in the ONB perspective, the exponential map at $\mathbf{X} \in \tilde{\mathcal{G}}(q, d)$ is defined as

$$\widetilde{\text{Exp}}_{\mathbf{X}}^{gr}(\mathbf{H}) = \mathbf{X}\mathbf{V} \cos \boldsymbol{\Sigma} + \mathbf{U} \sin \boldsymbol{\Sigma}, \quad (17)$$

where \mathbf{H} is a tangent vector at \mathbf{X} , and $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ is the thin singular value decomposition (SVD) of \mathbf{H} :

$$\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \text{thinSVD}(\mathbf{H}). \quad (18)$$

The logarithmic map, which is the inverse of the exponential map, is given by

$$\widetilde{\text{Log}}_{\mathbf{X}}^{gr}(\mathbf{Y}) = \mathbf{U} \tan^{-1} \boldsymbol{\Sigma} \mathbf{V}^\top, \quad (19)$$

where $\mathbf{X}, \mathbf{Y} \in \tilde{\mathcal{G}}(q, d)$, and

$$\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top = \text{thinSVD}((\mathbf{I} - \mathbf{X}\mathbf{X}^\top)\mathbf{Y}(\mathbf{X}^\top\mathbf{Y})^{-1}). \quad (20)$$

As stated in Edelman et al. [23, Theorem 2.4], let \mathbf{H} and Δ be tangent vectors at point \mathbf{Y} on the Grassmann manifold. The parallel transport of Δ along the geodesic in the direction $\dot{\mathbf{Y}}(0) = \mathbf{H}$ is given by

$$\tau\Delta(t) = \left((\mathbf{Y}\mathbf{V} - \mathbf{U}) \begin{pmatrix} -\sin(\boldsymbol{\Sigma}t) \\ \cos(\boldsymbol{\Sigma}t) \end{pmatrix} \mathbf{U}^\top + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top) \right) \Delta. \quad (21)$$

Shifting to the projector perspective for the Grassmannian manifold $\mathcal{G}(q, d)$, let $\mathbf{P} \in \mathcal{G}(q, d)$ and $\Delta \in T_{\mathbf{P}}\mathcal{G}(q, d)$. The exponential map is defined as [8]

$$\text{Exp}_{\mathbf{P}}^{gr}(\Delta) = \text{expm}([\Delta, \mathbf{P}])\mathbf{P} \text{expm}(-[\Delta, \mathbf{P}]). \quad (22)$$

As shown by Sakai [58], two points are in each other's cut locus if there exists more than one shortest geodesic connecting them. When the exponential map $\text{Exp}_{\mathbf{P}}^{gr}$ is restricted to the injectivity domain $\text{ID}_{\mathbf{P}}$, for any $\mathbf{F} \in \mathcal{G}(q, d) \setminus \text{Cut}_{\mathbf{P}}$, there exists a unique tangent vector $\Delta \in \text{ID}_{\mathbf{P}} \subset T_{\mathbf{P}}\mathcal{G}(q, d)$ such that $\text{Exp}_{\mathbf{P}}^{gr}(\Delta) = \mathbf{F}$. For such a point \mathbf{F} , the logarithmic map is given by

$$\text{Log}_{\mathbf{P}}^{gr}(\mathbf{Q}) = [\Omega, \mathbf{P}], \quad (23)$$

where $\mathbf{P}, \mathbf{Q} \in \text{Gr}_{n,p}$, and Ω is calculated as

$$\Omega = \frac{1}{2} \log ((\mathbf{I}_n - 2\mathbf{Q})(\mathbf{I}_n - 2\mathbf{P})). \quad (24)$$

As shown in Alg. 1, the Karcher flow algorithm computes the WFM on the Grassmannian manifold through an iterative process. We initialize \mathbf{G}_0 as the identity matrix \mathbf{X}_i and set the number of iterations $K = 1$.

D.2.4 Weighted Fréchet Mean on SPSD manifolds

As demonstrated by Bonnabel and Sepulchre [9], the WFM for a batch of points $\mathbf{X}_{1\dots N} \in \mathcal{S}_{d,q}^+$ can be expressed as $(\text{WFM}_{\text{gr}}(\mathbf{U}_X^i), \text{WFM}_{\text{spd}}^g(\mathbf{S}_X^i))$. Here, WFM_{gr} denotes the WFM on the Grassmannian manifold, while $\text{WFM}_{\text{spd}}^g(\cdot)$ represents the WFM on the SPD manifold under metric g . The matrices \mathbf{U}_X^i and \mathbf{S}_X^i correspond to the canonical representation of \mathbf{X}_i .

Algorithm 2: Karcher Flow Algorithm on the Grassmannian Manifold under ONB Perspective

Input: A set of Grassmannian points $\mathbf{X}_{1\dots N} \in \tilde{\mathcal{G}}(q, d)$
A set of weights $w_{1\dots N} > 0$ with $\sum_i w_i = 1$
Number of iterations K
Output: The WFM $\mathbf{G} \in \tilde{\mathcal{G}}(q, d)$
Initialize $\mathbf{G}_0 = \mathbf{X}_1$
for $k \leftarrow 1$ **to** K **do**
 $\mathbf{G}_k \leftarrow \tilde{\text{Exp}}_{\mathbf{G}_{k-1}}^{gr} \left(\sum_{i=1}^N w_i \tilde{\text{Log}}_{\mathbf{G}_{k-1}}^{gr}(\mathbf{X}_i) \right)$
end

Algorithm 3: Computation of Canonical Representation in SPSD manifold

Input: A batch of SPSD matrices $\mathbf{X}_{1\dots N} \in \mathcal{S}_{n,q}^+$
A constant $\gamma \in [0, 1]$
Output: A batch of Canonical Representation $(\mathbf{U}_X^i, \mathbf{S}_X^i)_{i=1,\dots,N}$ of SPSD manifold
 $\mathbf{U}^m \leftarrow \tilde{\mathbf{I}}_{n,q};$
 $(\mathbf{U}_i, \Sigma_i, \mathbf{V}_i)_{i=1,\dots,N} \leftarrow \text{SVD}((\mathbf{X}_i)_{i=1,\dots,N})$
 $(\mathbf{U}_i)_{i=1,\dots,N} \leftarrow (\mathbf{U}_i[:, :q])_{i=1,\dots,N};$
if training **then**
 $\mathbf{U} \leftarrow \text{GrMean}((\mathbf{U}_i)_{i=1,\dots,N})$
 $\mathbf{U}^m \leftarrow \text{GrGeodesic}(\mathbf{U}^m, \mathbf{U}, \gamma)$
end
for $i \leftarrow 1$ **to** N **do**
 $(\mathbf{U}_i)^\top \mathbf{U}^m = \mathbf{Y}_i (\cos \Sigma_i) \mathbf{V}_i^\top$
 $(\mathbf{U}_X^i, \mathbf{S}_X^i) \leftarrow (\mathbf{U}_i \mathbf{Y}_i, \mathbf{V}_i \mathbf{Y}_i^\top \mathbf{U}_i^\top \Sigma_i \mathbf{U}_i \mathbf{Y}_i \mathbf{V}_i^\top)$
end

D.3 Canonical representation in SPSD manifolds

Nguyen et al. [53] introduced a canonical representation of \mathbf{P} in the structure space $\tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}$. As shown in Alg. 3, we follow this approach to derive the canonical representation of each point in $\mathcal{S}_{d,q}^+$. Canonical Representation of SPSD matrices is obtained in three steps. This first is to impose a decomposition on \mathbf{X}_i , *i.e.*, $\mathbf{X}_i \simeq \mathbf{U}_i \Sigma_i \mathbf{U}_i^\top$, where $\mathbf{U}_i \in \tilde{\mathcal{G}}(q, d)$ and $\Sigma_i \in \mathcal{S}_q^{++}$. Then we use the mean of $\mathbf{U}_i)_{i=1,\dots,N}$ as the common subspace, and rotated (\mathbf{U}^i, Σ^i) to the identified common subspace, denoted as $(\mathbf{U}_X^i, \mathbf{S}_X^i)$. Here, $\text{GrMean}((\mathbf{U}_i)_{i=1,\dots,N})$ computes the Fréchet mean of its arguments, as described in Alg. 2, with weights set to $w_{1,\dots,N} = \frac{1}{N}$. $\text{GrGeodesic}(\mathbf{U}^m, \mathbf{U}, \gamma)$ computes a point on a geodesic (Eq. (21)) from \mathbf{U}^m to \mathbf{U} at step γ ($\gamma = 0.1$ in our experiments).

E Implementation details and additional experiments

E.1 Datasets and EEG preprocessing

MAMEM-SSVEP-II. This dataset includes EEG recordings from 11 subjects performing SSVEP tasks. Participants focused on one of five visual stimuli flickering at different frequencies for five seconds. Each subject completed five sessions, with five trials per stimulation frequency in each session. EEG signals were captured with 256 channels at a sampling rate of 250 Hz.

BCI-ERN. This dataset involves 26 subjects in a P300-based spelling task to measure ERN. EEG data were recorded from 56 electrodes following the extended 10-20 system at a sampling rate of 600 Hz. Each subject underwent five sessions: the first four with 60 trials each and the fifth with 100 trials. We used data from 16 subjects available in the initial competition release.

BNCI2014001. This dataset comprises EEG recordings from 9 subjects performing four motor imagery tasks: imagining movements of the left hand, right hand, both feet, and tongue. Each subject participated in two sessions on different days, each containing six runs. Each run included 48 trials—12 per class—totaling 288 trials per session.

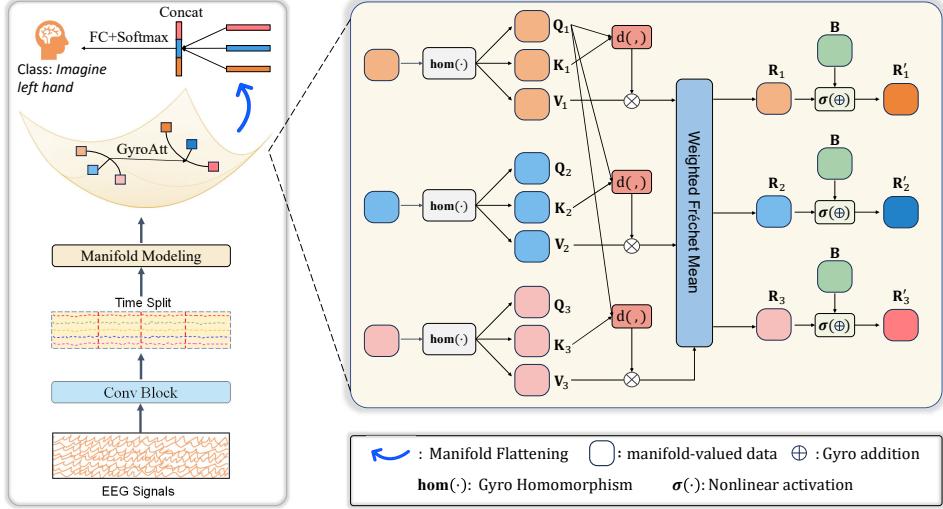


Figure 4: The GyroAtt network architecture comprises three components: a feature extraction module that converts EEG signals into manifold-valued data, a Gyro Attention module that explicitly captures long-range dependencies among features, and a classification module that flattens manifold data for a fully connected layer and softmax function.

BNCI2015001. EEG signals were recorded from electrodes centered around positions C3, Cz, and C4, according to the International 10-20 System. Data were collected using a g.GAMMASys active electrode system with a g.USBamp amplifier, sampled at 512 Hz with a bandpass filter between 0.5 and 100 Hz and a notch filter at 50 Hz.

EEG Preprocessing. For the MAMEM-SSVEP-II dataset, we adhered to the preprocessing protocol of Pan et al. [55]. The steps included: (1) band-pass filtering between 1–50 Hz; (2) selecting eight channels (PO7, PO3, PO, PO4, PO8, O1, Oz, and O2) located in the occipital area corresponding to the visual cortex; and (3) segmenting each trial into four 1-second segments from 1s to 5s after cue onset. This resulted in 500 trials of 1-second, 8-channel SSVEP signals per subject, with each input EEG segment comprising 125 time points.

For the BCI-ERN dataset, we followed the preprocessing procedure outlined by Pan et al. [55]. The steps involved: (1) downsampling the signals from 600 Hz to 128 Hz; (2) applying a band-pass filter between 1–40 Hz. After preprocessing, each trial consisted of 56 channels with 160 time points.

For the BNCI2014001 and BNCI2015001 datasets, we followed the preprocessing steps described by Kobler et al. [41]. Using the Python packages moabb and mne, we resampled the EEG signals to 250/256 Hz, applied temporal filters to extract oscillatory activity in the 4–36 Hz range, and extracted short segments (≤ 3 seconds) associated with class labels.

Table 13: GyroAtt-SPSD architectures across four datasets. The q is the rank of the SPSD matrices.

Block	MAMEM-SSVEP-II	BCI-ERN	BNCI2014001	BNCI2015001	Operation
Input data	$1 \times 8 \times 125$	$1 \times 56 \times 160$	$1 \times 22 \times 750$	$1 \times 13 \times 768$	
TempConv	$125 \times 1 \times 125$	$22 \times 1 \times 160$	$4 \times 22 \times 750$	$5 \times 13 \times 768$	Convolution
SpatConv	$21 \times 1 \times 126$	$57 \times 1 \times 161$	$43 \times 1 \times 750$	$44 \times 1 \times 768$	Convolution
Split & CovPool	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Split + Covariance
SPDDSMBN	w/o	w/o	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Domain Alignment Kobler et al. [41]
SPSDCom	$2 \times (21 \times 9, 9 \times 9)$	$3 \times (19 \times q, q \times q)$	$6 \times (43 \times 18, 18 \times 18)$	$3 \times (44 \times 18, 18 \times 18)$	Alg. 3
GyroAtt-SPSD	$2 \times (21 \times 9, 9 \times 9)$	$3 \times (19 \times q, q \times q)$	$6 \times (43 \times 18, 18 \times 18)$	$3 \times (44 \times 18, 18 \times 18)$	Alg. 4
R2E	$2 \times (21 \times 21, 9 \times 9)$	$3 \times (19 \times 19, q \times q)$	$6 \times (43 \times 43, 18 \times 18)$	$3 \times (44 \times 44, 18 \times 18)$	$(\Phi(\cdot), \psi(\cdot))$
Flat	(882, 162)	$(1083, q^2)$	(11094, 1944)	(5547, 972)	Vectorization
Classifier	5	2	4	2	FC + Softmax

Table 14: GyroAtt-SPD architectures across four datasets.

Block	MAMEM-SSVEP-II	BCI-ERN	BNCI2014001	BNCI2015001	Operation
Input data	$1 \times 8 \times 125$	$1 \times 56 \times 160$	$1 \times 22 \times 750$	$1 \times 13 \times 768$	
TempConv	$125 \times 1 \times 125$	$22 \times 1 \times 160$	$4 \times 22 \times 750$	$5 \times 13 \times 768$	Convolution
SpatConv	$21 \times 1 \times 126$	$57 \times 1 \times 161$	$43 \times 1 \times 750$	$44 \times 1 \times 768$	Convolution
Split & CovPool	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Split + Covariance
SPDDSMBN	w/o	w/o	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Domain Alignment
GyroAtt-SPD	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Alg. 4
R2E	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	$\psi(\cdot)$
Flat	882	1083	11094	5547	Vectorization
Classifier	5	2	4	2	FC + Softmax

Table 15: GyroAtt-Gr Architectures across four datasets. The q is the dimension of the linear subspaces.

Block	MAMEM-SSVEP-II	BCI-ERN	BNCI2014001	BNCI2015001	Operation
Input data	$1 \times 8 \times 125$	$1 \times 56 \times 160$	$1 \times 22 \times 750$	$1 \times 13 \times 768$	
TempConv	$125 \times 1 \times 125$	$22 \times 1 \times 160$	$4 \times 22 \times 750$	$5 \times 13 \times 768$	Convolution
SpatConv	$21 \times 1 \times 126$	$57 \times 1 \times 161$	$43 \times 1 \times 750$	$43 \times 1 \times 768$	Convolution
Split & CovPool	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Split + Covariance
SPDDSMBN	w/o	w/o	$6 \times 43 \times 43$	$3 \times 44 \times 44$	Domain Alignment
GrCom	$2 \times 21 \times 9$	$3 \times 19 \times q$	$6 \times 43 \times 18$	$3 \times 44 \times 18$	Alg. 3
GyroAtt-Gr	$2 \times 21 \times 9$	$3 \times 19 \times q$	$6 \times 43 \times 18$	$3 \times 44 \times 18$	Alg. 4
R2E	$2 \times 21 \times 21$	$3 \times 19 \times 19$	$6 \times 43 \times 43$	$3 \times 44 \times 44$	$\Phi(\cdot)$
Flat	882	1083	11094	5547	Vectorization
Classifier	5	2	4	2	FC + Softmax

E.2 Additional implementation details

E.2.1 Network architectures.

Fig. 4 and Tab. 13 provide a summary of the specific network architectures of GyroAtt-SPSD across the four datasets, and we summarize the forward pass of GyroAtt in Alg. 4. The network structures for GyroAtt-Gr (Tab. 15) and GyroAtt-SPD (Tab. 14) are identical to that of GyroAtt-SPSD. We just introduced GyroAtt-SPSD as an example.

For the MAMEM-SSVEP-II and BCI-ERN datasets, the initial convolutional block consists of a convolutional layer, followed by batch normalization and an ELU activation function. The subsequent convolutional block performs depthwise spatial convolution. A pointwise convolution, batch normalization, and another ELU activation follow this. In the MAMEM-SSVEP-II dataset, features are split into two non-overlapping segments, followed by covariance pooling. For the BCI-ERN dataset, the second convolutional block is repeated in two additional blocks. The outputs from these blocks are concatenated along the channel dimension. The data is then split along the channel dimension, and covariance pooling is applied, resulting in three covariance matrices.

For BNCI2014001 and BNCI2015001 datasets, the initial convolutional layer employs 4 or 5 filters with a kernel size of $(1, 25)$, performing temporal convolution while maintaining the same size through padding. The second convolutional layer applies spatial convolution with a kernel size of $(22, 1)$ to integrate information from different channels. The output sequences undergo temporal pyramid partitioning, dividing each sequence into i equal segments at the i -th level (with levels set to 3 and 2, respectively). To address distribution shifts across subjects and runs, we incorporate subject- and run-specific batch normalization layers [41].

The attention module designed in the gyrovector spaces is constituted by five operation layers, which are the Gyro homomorphism layer (f_{hom}) used to generate \mathbf{Q}_i , \mathbf{K}_i , and \mathbf{V}_i for each input data, the similarity measurement layer (f_{sim}) for computing the correlation between \mathbf{Q}_i and \mathbf{K}_j , the Softmax layer (f_{smix}) used to normalize the obtained attention matrix along the row direction, the weighted Fréchet Mean layer (f_{wFM}) for the implementation of weighted aggregation, and the power-based nonlinear activation layer (f_{pac}) used to improve the representational capacity of GyroAtt module by introducing nonlinearity to the underlying metric space.

For classification, our GyroAtt-SPD model employs matrix power normalization following Wang et al. [71] and Chen et al. [16]. Specifically, we apply the transformation $\psi_\theta(\mathbf{P}) = \frac{1}{\theta} \mathbf{P}^\theta$ to the i -th

Algorithm 4: Gyro Attention over gyrovector spaces

Input: A set of manifold-valued features $\{\mathbf{X}_{1\dots N}\}$
Output: A set of manifold-valued features $\{\mathbf{R}'_{1\dots N}\}$

```

for  $i \leftarrow 1$  to  $N$  do
  | Queries:  $\mathbf{Q}_i = \text{hom}(\mathbf{X}_i)$ ,
  | Keys:  $\mathbf{K}_i = \text{hom}(\mathbf{X}_i)$ ,
  | Values:  $\mathbf{V}_i = \text{hom}(\mathbf{X}_i)$ 
end
for  $i \leftarrow 1$  to  $N$  do
  | for  $j \leftarrow 1$  to  $N$  do
  |   | Similarity:  $\mathcal{S}_{ij} = (1 + \log(1 + d(\mathbf{Q}_i, \mathbf{K}_j)))^{-1}$ 
  | end
  | Attention calculation:  $\mathcal{A}_{ij} = \text{Softmax}(\mathcal{S}_{ij})$ 
  | Aggregation:  $\mathbf{R}_i = \text{WFM}(\{\mathcal{A}_{ij}\}_{j=1}^N, \{\mathbf{V}_j\}_{j=1}^N)$ 
  | Bias and nonlinearity:  $\mathbf{R}'_i = \sigma(\mathbf{R}_i \oplus \mathbf{B})$ 
end

```

output matrix $\mathbf{P} \in \mathcal{S}_d^{++}$, where $\theta > 0$. The coefficient $\frac{1}{\theta}$ stabilizes the gradient flow during training and facilitates convergence. In GyroAtt-Gr, we transform elements $\mathbf{Y}_i \in \mathcal{G}(q, d)$ by applying a projection operator $\Phi(\mathbf{Y}_i) = \mathbf{Y}_i \mathbf{Y}_i^\top$ to map them into the corresponding flat space. In contrast, for GyroAtt-SPSD, we project $(\mathbf{U}_X^i, \mathbf{S}_X^i) \in \tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}$ onto their respective manifolds. In all three GyroAtt, the transformed matrices are vectorized, concatenated, and fed into a fully connected layer followed by a Softmax function. All experiments were conducted on an i9-14900 CPU with 64GB RAM and two NVIDIA RTX4080 Super GPUs.

E.2.2 Optimization

We address the optimization of parameters that are SPD matrices by modeling them within the space of symmetric matrices and applying the exponential map to the identity matrix.

For any parameter $\mathbf{P} \in \tilde{\mathcal{G}}(d, q)$, we parameterize it using a matrix $\mathbf{B} \in \mathbb{R}^{q, d-q}$ such that

$$\begin{bmatrix} 0 & \mathbf{B} \\ -\mathbf{B}^\top & 0 \end{bmatrix} = [\text{Log}_{\mathbf{I}_{n,p}}^{gr}(\mathbf{P}\mathbf{P}^\top), \mathbf{I}_{n,p}]. \quad (25)$$

With this parameterization, the parameter \mathbf{P} can be computed as

$$\mathbf{P} = \exp\left(\begin{bmatrix} 0 & \mathbf{B} \\ -\mathbf{B}^\top & 0 \end{bmatrix}\right) \tilde{\mathbf{I}}_{n,p}.$$

To optimize parameters $\mathbf{O} \in SO(n)$, we start by generating parameter $\mathbf{A} \in \mathbb{R}^{n \times n}$, then compute its skew-symmetric matrix $\mathbf{S} = \mathbf{A} - \mathbf{A}^\top$. With this parameterization, the parameter \mathbf{P} can be computed as

$$\mathbf{O} = (\mathbf{I} - \mathbf{S})(\mathbf{I} + \mathbf{S})^{-1}, \quad (26)$$

This approach enables us to optimize all parameters within Euclidean spaces, eliminating the need to employ optimization techniques specific to Riemannian manifolds.

E.3 Statistical Significance Analysis

To assess the robustness of the reported improvements, we conducted paired *t*-tests comparing GyroAtt with the baseline MAtt across all evaluation folds. Tab. 16 reports the average performance gains and corresponding significance levels, where * denotes p -value < 0.05 and *n.s.* indicates “not significant”.

Although some confidence intervals overlap, the observed improvements are consistent across benchmarks. In particular, the +14.8% gain under the inter-subject condition—known to be the most challenging EEG setting—demonstrates the practical effectiveness and robustness of GyroAtt.

Table 16: Paired t -test results comparing GyroAtt with MAtt. Asterisks (*) denote statistical significance ($p < 0.05$).

	BNCI2014001		BNCI2015001		BCI-CHA	MAMEM
	Inter-session	Inter-subject	Inter-session	Inter-subject		
Gains (%)	7.8	8.9	5.4	14.8	3.4	3.2
Significance	*	n.s.	n.s.	*	*	*

E.4 Graph node classification and radar recognition tasks

While our primary evaluation focuses on EEG decoding, we further assess the generalization capability of GyroAtt on two additional domains: graph node classification and drone recognition.

Table 17: Accuracy (%) on graph node classification benchmarks.

Method	Pubmed	Cora
H-GAT	77.5 ± 1.6	78.1 ± 1.1
SPD-GAT	77.8 ± 0.6	79.4 ± 0.6
SPD-GAT-MAtt	77.5 ± 0.6	80.1 ± 1.1
SPD-GAT-GyroAtt	78.5 ± 0.4	81.6 ± 1.0

Graph node classification. We conduct experiments on two standard graph node classification datasets: Pubmed [49] and Cora [60]. We adopt SPD-GAT [75] as the backbone and substitute its attention module with GyroAtt-SPD (denoted SPD-GAT-GyroAtt) and MAtt (SPD-GAT-MAtt). We also include H-GAT [75] for comparison. As shown in Tab. 17, GyroAtt-SPD consistently outperforms all baselines on both datasets.

Table 18: Accuracy (%) on Radar dataset for drone recognition.

Method	Radar
MAtt	96.8 ± 1.1
GDLNet	94.7 ± 0.9
GyroAtt-SPD	98.5 ± 0.8
GyroAtt-Gr	96.2 ± 1.2
GyroAtt-SPSD	96.9 ± 0.9

Drone recognition. We evaluate GyroAtt on the Radar dataset [11], following the GDLNet [72] architecture. The dataset comprises 3,000 synthetic radar signals, evenly distributed across three classes. Tab. 18 shows that GyroAtt-SPD attains the highest accuracy, further validating the adaptability of our method across different modalities.

E.5 Additional ablations studies

E.5.1 Ablations on the Riemannian metrics and matrix power-based nonlinear activation

Tab. 19 illustrates the impact of the different metrics and power parameter p (as defined in Tab. 3) on the performance of GyroAtt based on two Riemannian matrix manifolds. The candidate values of metrics are AIM, LEM, and LCM, with p values set to $\{0.25, 0.50, 0.75\}$. As shown in this table, for SPD-based architectures, GyroAtt under the SPD-AIM geometry with $p = 0.5$ achieves the highest accuracy on both the BNCI2014001 and BNCI2015001 datasets, while the SPD-LCM geometry with $p = 0.75$ records the second-highest inter-session accuracy (86.0%) on the BNCI2015001 dataset. For SPSD-based settings, GyroAtt under the SPSD-LCM geometry with $p = 0.75$ reaches the highest accuracy (68.7%) on the MAMEM-SSVEP-II dataset. Furthermore, it is evident that GyroAtt is generally robust to variations in p across all experimental scenarios. These findings emphasize the importance of selecting the metric space of the underlying feature manifold and demonstrate that the proposed matrix power activation enhances model performance by introducing nonlinearity into the metric space.

Table 19: Ablations of GyroAtt on Riemannian metrics and matrix power activation p . The best result under each geometry is highlighted in **bold**.

Geometry	p	BNCI2014001		BNCI2015001		MAMEM-SSVEP-II
		Inter-session	Inter-subject	Inter-session	Inter-subject	
SPD-AIM	w/o	74.8 \pm 6.7	51.2 \pm 15.7	85.5 \pm 5.0	75.4 \pm 12.9	64.3 \pm 2.4
	0.25	75.2 \pm 6.9	51.4 \pm 14.3	85.8 \pm 6.8	77.1 \pm 12.8	61.9 \pm 2.5
	0.50	75.4 \pm 7.1	53.1 \pm 14.8	86.2 \pm 4.5	77.9 \pm 13.0	66.1 \pm 2.6
	0.75	75.0 \pm 8.1	51.0 \pm 13.8	85.9 \pm 6.6	77.4 \pm 12.6	66.3 \pm 2.2
SPD-LEM	w/o	74.9 \pm 7.3	51.7 \pm 15.8	85.2 \pm 5.2	75.3 \pm 12.3	65.6 \pm 2.3
	0.25	74.7 \pm 6.7	52.3 \pm 14.1	85.6 \pm 6.7	75.6 \pm 13.0	63.7 \pm 2.5
	0.50	75.3 \pm 6.5	51.4 \pm 14.1	85.7 \pm 5.5	76.6 \pm 13.7	65.3 \pm 2.7
	0.75	75.1 \pm 7.3	52.3 \pm 15.0	85.4 \pm 7.0	75.5 \pm 12.8	66.2 \pm 2.5
SPD-LCM	w/o	73.2 \pm 6.7	51.9 \pm 14.8	85.3 \pm 7.2	76.2 \pm 13.3	64.0 \pm 2.8
	0.25	73.4 \pm 7.5	52.4 \pm 13.4	85.6 \pm 7.5	75.4 \pm 14.0	64.3 \pm 2.5
	0.50	74.0 \pm 8.2	52.7 \pm 13.6	85.9 \pm 6.7	77.4 \pm 13.2	64.1 \pm 3.2
	0.75	74.2 \pm 7.8	51.7 \pm 14.6	86.0 \pm 6.8	76.3 \pm 13.2	65.1 \pm 2.5
SPSD-AIM	w/o	72.2 \pm 7.2	49.2 \pm 13.7	84.1 \pm 7.2	73.6 \pm 14.3	65.8 \pm 2.6
	0.25	72.4 \pm 6.8	50.7 \pm 14.8	84.0 \pm 6.8	75.5 \pm 13.8	66.4 \pm 3.0
	0.50	72.9 \pm 7.1	52.4 \pm 15.6	84.7 \pm 6.6	74.2 \pm 14.2	66.5 \pm 2.9
	0.75	72.5 \pm 6.7	51.0 \pm 15.3	84.0 \pm 4.9	74.5 \pm 13.6	65.7 \pm 2.7
SPSD-LEM	w/o	72.1 \pm 6.7	49.8 \pm 12.9	83.9 \pm 5.1	74.2 \pm 14.4	66.2 \pm 1.9
	0.25	72.8 \pm 6.9	49.9 \pm 14.0	85.3 \pm 5.3	76.0 \pm 14.1	66.5 \pm 2.3
	0.50	72.5 \pm 6.6	50.5 \pm 14.2	84.5 \pm 5.8	75.4 \pm 14.5	66.4 \pm 2.5
	0.75	72.7 \pm 7.6	50.5 \pm 13.2	85.2 \pm 4.8	75.0 \pm 14.2	66.2 \pm 1.7
SPSD-LCM	w/o	72.3 \pm 7.3	49.5 \pm 12.0	84.8 \pm 6.1	75.4 \pm 13.2	66.5 \pm 2.4
	0.25	72.2 \pm 7.5	50.6 \pm 13.9	85.1 \pm 4.8	74.9 \pm 12.6	67.7 \pm 2.3
	0.50	72.9 \pm 6.7	48.4 \pm 13.3	84.9 \pm 6.1	74.5 \pm 13.6	66.2 \pm 3.6
	0.75	72.8 \pm 6.3	51.7 \pm 13.1	85.1 \pm 5.8	73.9 \pm 15.4	68.7 \pm 1.5

Table 20: Ablations of GyroAtt on matrix power normalization θ used in classification and Riemannian metrics. The best result under each geometry is highlighted in **bold**.

Geometry	θ	BNCI2014001		BNCI2015001		MAMEM-SSVEP-II
		Inter-session	Inter-subject	Inter-session	Inter-subject	
SPD-AIM	0.25	74.9 \pm 6.9	51.2 \pm 13.6	86.1 \pm 7.3	76.2 \pm 12.8	61.9 \pm 2.5
	0.50	75.4 \pm 7.1	53.1 \pm 14.8	86.2 \pm 4.5	77.9 \pm 13.0	66.2 \pm 2.8
	0.75	75.0 \pm 8.1	51.7 \pm 14.5	86.0 \pm 6.5	77.1 \pm 14.3	66.3 \pm 2.2
SPD-LEM	0.25	75.2 \pm 6.7	52.7 \pm 12.9	85.1 \pm 5.7	76.9 \pm 14.5	60.7 \pm 2.4
	0.50	75.3 \pm 6.5	51.4 \pm 14.1	85.7 \pm 5.5	76.6 \pm 13.7	66.1 \pm 2.8
	0.75	75.1 \pm 7.3	52.3 \pm 13.3	85.8 \pm 6.3	76.4 \pm 13.1	66.2 \pm 2.5
SPD-LCM	0.25	74.2 \pm 7.5	52.1 \pm 14.5	85.6 \pm 5.9	77.3 \pm 13.4	64.5 \pm 2.9
	0.50	74.0 \pm 8.2	52.7 \pm 13.6	85.9 \pm 6.7	77.4 \pm 13.2	64.3 \pm 2.8
	0.75	74.1 \pm 7.8	52.0 \pm 14.7	86.0 \pm 5.3	75.8 \pm 13.8	65.1 \pm 2.5
SPSD-AIM	0.25	72.7 \pm 7.0	51.2 \pm 15.8	84.0 \pm 6.8	75.5 \pm 13.8	66.3 \pm 2.9
	0.50	72.9 \pm 6.2	52.4 \pm 15.6	84.5 \pm 6.6	74.2 \pm 15.2	66.3 \pm 2.4
	0.75	72.7 \pm 6.7	50.0 \pm 15.2	84.4 \pm 4.9	75.3 \pm 13.5	65.7 \pm 2.7
SPSD-LEM	0.25	72.8 \pm 7.1	50.7 \pm 13.9	85.3 \pm 5.3	76.0 \pm 14.1	66.6 \pm 2.6
	0.50	72.5 \pm 6.6	50.6 \pm 14.2	84.5 \pm 5.8	75.1 \pm 12.9	66.5 \pm 1.9
	0.75	72.7 \pm 7.4	49.5 \pm 12.9	84.3 \pm 4.8	74.7 \pm 14.3	66.2 \pm 1.7
SPSD-LCM	0.25	72.1 \pm 7.4	49.9 \pm 13.1	85.1 \pm 4.8	74.9 \pm 12.6	67.6 \pm 2.1
	0.50	72.9 \pm 6.7	48.4 \pm 13.3	84.1 \pm 5.6	74.4 \pm 13.7	68.1 \pm 1.6
	0.75	71.6 \pm 6.1	50.1 \pm 12.8	84.1 \pm 5.7	75.0 \pm 12.9	68.7 \pm 1.5

E.5.2 Ablations on the matrix power normalization

We conduct ablation experiments to assess the impact of the power normalization parameter θ on the performance of the proposed GyroAtt, as summarized in Tab. 20. For each gyro structure, we let the parameter θ vary within the set $\{0.25, 0.50, 0.75\}$. Among the SPD-based configurations, our GyroAttNet under SPD-AIM geometry achieves the highest inter-session accuracy on the BNCI2014001 dataset and the best inter-subject accuracy on the BNCI2015001 dataset at $p = 0.5$. For the SPSD-based settings, SPSD-LEM geometry consistently performs well across multiple metrics, especially for the inter-session scenario in BNCI2015001, where it achieves a top accuracy of 85.3%. It also

can be noted that smaller or larger values of p (e.g., 0.25 or 0.75) tend to yield lower accuracy in most cases. In contrast, a moderate value of $p = 0.5$ appears to be more suitable for both SPD and SPSD geometries, as it could maintain a good normalization power. Besides, GyroAtt tends to be less sensitive to changes in θ across all experimental scenarios. In short, these results confirm the effectiveness of the introduced matrix power normalization in classification.

E.5.3 Ablation on the number of GyroAtt blocks

We perform an ablation study to assess the impact of stacking multiple GyroAtt blocks. Tab. 21 reports results on BNCI2014001 and BNCI2015001 under inter-session settings, with the number of GyroAtt blocks varying from 1 to 3.

For both SPD and SPSD variants, using two blocks slightly improves performance over one, but adding a third block provides no further gain and even causes mild degradation. These results suggest that a single GyroAtt block is sufficient to model EEG signals effectively, likely due to the low dimensionality and limited temporal complexity of the data. Extended results are included in the supplementary material.

Table 21: Ablation on the number of GyroAtt blocks. Performance (%) on BNCI2014001 and BNCI2015001 under inter-session settings.

Method	Number of Blocks	BNCI2014001	BNCI2015001
GyroAtt-SPD	1	75.4 ± 7.1	86.2 ± 6.5
GyroAtt-SPD	2	76.5 ± 6.7	86.1 ± 6.2
GyroAtt-SPD	3	76.0 ± 6.3	86.0 ± 6.1
GyroAtt-SPSD	1	72.9 ± 6.7	84.7 ± 6.6
GyroAtt-SPSD	2	73.1 ± 6.2	84.9 ± 5.2
GyroAtt-SPSD	3	72.9 ± 5.1	84.8 ± 5.3

E.6 Implementation details of gyro inner product.

In Euclidean space, attention mechanisms commonly use the inner product as a similarity measure. Nguyen and Yang [52], Nguyen et al. [53] extends this concept by defining the inner product on SPD, SPSD, and Grassmannian manifolds. The specific formulations are detailed as follows:

For $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, the SPD inner product is given by [52]:

$$\langle \mathbf{P}, \mathbf{Q} \rangle^g = \langle \text{Log}_{\mathbf{I}_d}^g(\mathbf{P}), \text{Log}_{\mathbf{I}_d}^g(\mathbf{Q}) \rangle_{\mathbf{I}_d}^g, \quad (27)$$

For $\mathbf{U}, \mathbf{V} \in \tilde{\mathcal{G}}(q, d)$, the inner product is given by:

$$\langle \mathbf{U}, \mathbf{V} \rangle^{gr} = \langle \widetilde{\text{Log}}_{\mathbf{I}_{d,q}}^{gr}(\mathbf{U}), \widetilde{\text{Log}}_{\mathbf{I}_{d,q}}^{gr}(\mathbf{V}) \rangle_{\mathbf{I}_{d,q}}, \quad (28)$$

For $(\mathbf{U}_P, \mathbf{S}_P), (\mathbf{U}_Q, \mathbf{S}_Q) \in \tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}$, the inner product is defined as:

$$\langle (\mathbf{U}_P, \mathbf{S}_P), (\mathbf{U}_Q, \mathbf{S}_Q) \rangle^{psd,g} = \lambda \langle \mathbf{U}_P \mathbf{U}_P^\top, \mathbf{U}_Q \mathbf{U}_Q^\top \rangle_{\mathbf{I}_{d,q}}^{gr} + \langle \mathbf{S}_P, \mathbf{S}_Q \rangle_{\mathbf{I}_q}^g, \quad (29)$$

We replaced the distance-based similarity computation in Eq. (5) with the inner product defined in follow and conducted ablation experiments on the MAMEM, BNCI2014001, and BNCI2015001 datasets under inter-session settings.

E.7 Implementation details of replacing geometric components in GyroAtt

We conducted an ablation study to evaluate the contributions of the Gyro Homomorphism and nonlinear activation in GyroAtt. Specifically, we replaced these components in GyroAtt-SPD and GyroAtt-SPSD with equivalent layers from SPDNet and GrNet, such as Bimap, Frmap, and ReEig, to assess their impact on performance.

Implementation of component replacement on GyroAtt. We replaced components in GyroAtt with their equivalents from MAtt and GDLNet to assess their contributions. Specifically, in GyroAtt-SPD, we replaced the Gyro Homomorphism $\text{hom}(\cdot)$ with the Bimap layer and the matrix power-based nonlinear activation $\sigma(\cdot)$ with the ReEig layer. In GyroAtt-SPSD, we replaced $\text{hom}(\cdot)$ with the Frmap layer and $\sigma(\cdot)$ with the ReEig layer. That is, we substituted $\text{hom}_{gr}(\mathbf{U}_P)$ in $\text{hom}_{psd,g}(\cdot)$ with the Frmap layer and replaced $(\mathbf{S}_{R_i})^p$ in $(\mathbf{U}_{R_i}, (\mathbf{S}_{R_i})^p)$ with the ReEig layer.

The BiMap (bilinear transformation) layer is defined as:

$$\mathbf{X}^{(l)} = \mathbf{W}^{(l)} \mathbf{X}^{(l-1)} \mathbf{W}^{(l)\top}, \quad (30)$$

where $\mathbf{X}^{(l)} \in \mathcal{S}_{d_2}^{++}$, $\mathbf{X}^{(l-1)} \in \mathcal{S}_{d_1}^{++}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{d_2 \times d_1}$ with $d_1 > d_2$ is a semi-orthogonal matrix. For the parameter $\mathbf{W}^{(l)}$, we use the geoopt [43] package to optimize. The FrMap layer is defined as:

$$\mathbf{X}^{(l)} = \mathbf{W}^{(l)\top} \mathbf{X}^{(l-1)}, \quad (31)$$

where $\mathbf{X}^{(l)} \in \mathcal{G}(d_2, q)$, $\mathbf{X}^{(l-1)} \in \mathcal{G}(d_1, q)$, and $\mathbf{W}^{(l)} \in \mathbb{R}^{d_2 \times d_1}$ is a semi-orthogonal matrix with $d_1 > d_2$. We optimized $\mathbf{W}^{(l)}$ using Geoopt.

The ReEig (rectified eigenvalues activation) layer is defined as:

$$\mathbf{X}^l = \mathbf{U}^{(l)} \max(\Sigma^{(l)}, \epsilon \mathbf{I}_d) \mathbf{U}^{(l)\top}, \quad (32)$$

with $\mathbf{X}^{l-1} = \mathbf{U}^{(l)} \Sigma^{(l)} \mathbf{U}^{(l)\top}$, where $\Sigma^{(l)}$ contains the eigenvalues of \mathbf{X}^{l-1} , and $\epsilon \mathbf{I}_d$ is used to ensure numerical stability and set by 1e-4. Here, we set the dimensions of the Bimap layer to 21×18 , 43×20 , and 44×20 and the frmap layer to 21×18 , 43×30 , and 44×30 for the MAMEM-SSVEP-II, BNCI2014001, and BNCI2015001 datasets, respectively.

As shown in Tab. 8 and Tab. 22, replacing $\text{hom}(\cdot)$ with the Bimap layer or $\sigma(\cdot)$ with the ReEig layer leads to significant performance degradation across the datasets. Similarly, for GyroAtt-SPSD, replacing $\text{hom}(\cdot)$ with Frmap or $\sigma(\cdot)$ with ReEig degrades performance. This occurs because $\text{hom}(\cdot)$ and $\sigma(\cdot)$ respect the gyro algebraic structure and underlying Riemannian geometry. The $\text{hom}(\cdot)$ function, as a Gyro homomorphism, preserves the Gyro algebraic structure of \oplus and \otimes , serving as a natural generalization of linear transformations in Euclidean spaces. In contrast, Bimap lacks these properties. Similarly, $\sigma(\cdot)$ introduces nonlinearity to SPD matrices and, more importantly, acts as an activation and deformation mechanism for the Riemannian metric, as discussed in Chen et al. [17]. On the other hand, to some extent, ReEig is primarily a numerical activation method, ensuring only $\mathcal{S}_d^{++} \rightarrow \mathcal{S}_d^{++}$ without addressing these deeper structural and geometric considerations.

Table 22: Ablations of GyroAtt-SPSD, Replacing Gyro Homomorphisms and Power Activations with SPDNet or GrNet methods (The Frmap and ReEig layers).

Transformation	Activation	BNCI2014001		BNCI2015001	
		Inter-session	Inter-subject	Inter-session	Inter-subject
Frmap	Power	68.9 ± 6.9	51.2 ± 12.9	82.3 ± 6.2	65.8 ± 13.1
Homomorphisms	ReEig	72.3 ± 6.9	49.6 ± 13.3	84.9 ± 6.2	74.1 ± 12.3
Frmap	ReEig	68.8 ± 7.2	50.7 ± 13.8	81.6 ± 6.1	72.9 ± 13.3
Homomorphisms	Power	72.9 ± 7.1	52.4 ± 15.6	85.3 ± 5.3	76.0 ± 14.1

E.8 Computational complexity analysis and comparison

The computational complexity of GyroAtt, MAtt, and GDLNet primarily depends on operations such as WFM, gyro homomorphisms, and similarity measurements. The primary computational overhead arises from the types and quantities of matrix functions used in manifold-valued computations.

E.8.1 Number of matrix functions in manifold attention mechanisms

To analyze computational complexity, we examine the matrix functions employed in the attention mechanisms of MAtt, GDLNet, and our GyroAtt framework. These methods primarily use Singular Value Decomposition (SVD) and Cholesky decomposition, each with a computational complexity of $\mathcal{O}(d^3)$ for matrices of dimension d . In GyroAtt-Gr and GyroAtt-SPSD, SVD is applied to matrices sized $d \times q$, resulting in a complexity of $\mathcal{O}(dq^2)$. Tab. 23 outlines the number of required matrix functions for each method, where C is the number of inputs to the attention module (e.g., queries or keys), and q denotes the subspace dimension.

Table 23: Complexity comparison of different manifold attention models, where Chol denotes the Cholesky decomposition.

Methods	Metrics	SVD ($d \times d$) $\mathcal{O}(d^3)$	Chol ($d \times d$) $\mathcal{O}(d^3)$	SVD ($d \times q$) $\mathcal{O}(dq^2)$	SVD ($q \times q$) $\mathcal{O}(q^3)$	Chol ($q \times q$) $\mathcal{O}(q^3)$
MAtt	LEM	$3C$	0	0	0	0
GDLNet	ONB	C	0	0	0	0
GyroAtt-Gr	ONB	0	0	$C^2 + C$	$2C$	0
GyroAtt-SPD	AIM	$2C^2 + 3C$	0	0	0	0
GyroAtt-SPD	LEM	$4C$	0	0	0	0
GyroAtt-SPD	LCM	C	$3C$	0	0	0
GyroAtt-SPSD	AIM	0	0	$C^2 + C$	$3C^2 + 4C$	0
GyroAtt-SPSD	LEM	0	0	$C^2 + C$	$C^2 + 6C$	0
GyroAtt-SPSD	LCM	0	0	$C^2 + C$	$C^2 + 2C$	$3C$

Table 24: Training time (seconds/epoch) comparison of different manifold attention networks on the BNCI2014001 and BNCI2015001 datasets.

Methods	Metrics	BNCI2014001	BNCI2014001	BNCI2015001	BNCI2015001
		Inter-session	Inter-subject	Inter-session	Inter-subject
MAtt	LEM	4.86	89.12	2.74	56.78
GDLNet	ONB	4.55	88.59	1.71	47.66
GyroAtt-Gr	ONB	5.28	100.20	1.98	54.07
GyroAtt-SPD	AIM	10.36	149.14	4.44	81.73
GyroAtt-SPD	LEM	7.89	138.59	4.28	80.66
GyroAtt-SPD	LCM	4.11	87.44	2.42	49.86
GyroAtt-SPSD	AIM	6.78	123.44	2.37	65.30
GyroAtt-SPSD	LEM	6.54	116.49	2.34	64.23
GyroAtt-SPSD	LCM	5.71	103.57	2.16	58.50

E.8.2 Experimental comparison of time efficiency

For practical comparison, we measured the average training time per epoch (seconds) for our GyroAtt variants and baseline models MAtt and GDLNet on the BNCI2014001 and BNCI2015001 datasets, as shown in Tab. 24. The results indicate that GyroAtt-Gr runs slower than GDLNet, mainly because of more SVD operations in WFM, increasing computational complexity. As predicted by Tab. 23 and confirmed in Tab. 24, GyroAtt-SPSD is less computationally efficient than GDLNet and MAtt across all metrics. In contrast, GyroAtt-SPD with the LCM metric slightly outperforms MAtt in runtime. Additionally, Tab. 24 shows that any GyroAtt variant using the LCM metric runs faster than those using AIM and LEM metrics, primarily because the Cholesky decomposition (used in LCM) is generally more efficient than SVD due to smaller constant factors. Notably, AIM-based GyroAtt variants are slower than their LEM-based counterparts because AIM requires more eigenvalue computations than LEM.

Our computational complexity analysis and empirical runtime comparisons highlight key observations:

- **GyroAtt-Gr:** Runs slower than GDLNet due to more SVD operations.
- **GyroAtt-SPSD:** Less efficient than GDLNet and MAtt across all metrics, involving extra SVD computations on $d \times q$ and $q \times q$ matrices.
- **GyroAtt-SPD with LCM:** Achieves comparable or better runtime than MAtt, benefiting from the efficiency of the Cholesky decomposition.
- **Metric Impact:** The LCM metric leads to lower training times due to the efficiency of Cholesky decomposition over SVD. AIM-based GyroAtt variants are slower than LEM-based ones because AIM requires more eigenvalue operations.

Although some GyroAtt variants introduce additional computational overhead compared to baseline methods, especially under certain metrics, the performance improvements justify the increased

complexity. Selecting efficient metrics like LCM can mitigate computational costs, making GyroAtt practical for real-world applications.

E.9 Model compactness and parameter efficiency

Table 25: Learnable parameters (in millions) of different methods across four EEG datasets.

Method	MAMEM-SSVEP-II	BCI-ERN	BCI2014001	BCI2015001
MAtt	0.07M	0.03M	0.03M	0.01M
GDLNet	0.51M	0.05M	0.03M	0.01M
GyroAtt-SPD	0.11M	0.08M	0.04M	0.01M
GyroAtt-Gr	0.11M	0.07M	0.04M	0.01M
GyroAtt-SPSD	0.12M	0.09M	0.05M	0.01M

As shown in Tab. 25, all GyroAtt variants maintain a comparable parameter count to MAtt, and are substantially more compact than GDLNet on MAMEM-SSVEP-II. Despite this, GyroAtt achieves stronger performance across all datasets.

E.10 Long-range dependencies

A central advantage of self-attention is its ability to capture long-range dependencies via position-independent pairwise interactions, where each token attends to all others in a single step.

Our GyroAtt module inherits this property by extending self-attention to Riemannian manifolds. Specifically, it computes pairwise attention over manifold-valued features using geodesic distances, followed by aggregation via the weighted Fréchet mean. This allows any two positions to interact directly, independent of sequence length.

E.11 Visualizing learned representations via Rie-SNE

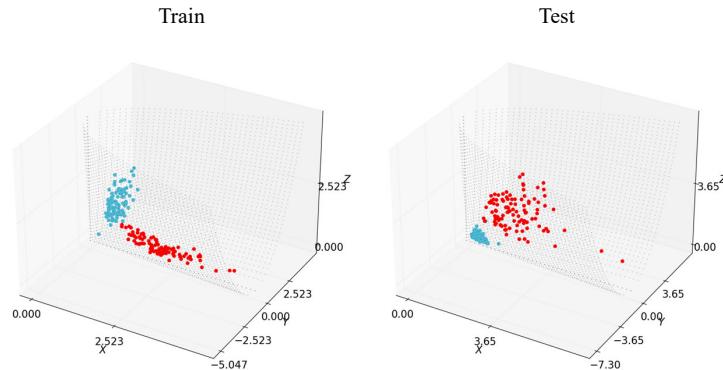


Figure 5: Rie-SNE visualization of SPD representations extracted by GyroAtt-SPD-LEM on BCI2015001. We plot the manifold-valued features from the training (left) and test (right) sets. Each point corresponds to SPD matrices used for final classification, with colors denoting different classes. Clear separation between classes and structural consistency across splits highlight the model’s capacity to learn discriminative geometric features.

We visualize the SPD matrices output from GyroAtt-SPD-LEM prior to classification using Rie-SNE [21], separately for the training and test sets on BCI2015001. As shown in Fig. 5, the learned representations exhibit compact intra-class clustering and clear inter-class separation across both splits.

F Proofs of the Theorems in the Main Paper

F.1 Proof of the Thm. 3.2

Proof of Thm. 3.2. This theorem can be induced from Nguyen and Yang [52][Lem. 2.1-2.2]. Let the Riemannian metric tensor be g . We have

$$f : (\mathcal{M}, f^*g) \rightarrow (\mathcal{M}, g), \text{ with } f(e) = e,$$

where f^*g is the pullback metric under the diffeomorphism f . As f is an isometry, we have $f^*g = g$. Applying Lem. 2.1-2.2 in Nguyen and Yang [52], one can immediately prove $f(\mathbf{P} \oplus \mathbf{Q}) = f(\mathbf{P}) \oplus f(\mathbf{Q})$ and $f(t \otimes \mathbf{Q}) = t \otimes f(\mathbf{Q})$ for any $t \in \mathbb{R}$ and $\mathbf{P}, \mathbf{Q} \in M$. \square

F.2 Proof of the Thm. 4.1

Proof of Thm. 4.1. For each $g \in \{ai, le, lc\}$, let $(\mathcal{S}_d^{++}, \oplus_g, \otimes_g)$ be the gyrovector space specified in Tab. 1. We must show that the map

$$\text{hom}_g : \mathcal{S}_d^{++} \rightarrow \mathcal{S}_d^{++}, \mathbf{P} \mapsto \begin{cases} \mathbf{O}\mathbf{P}\mathbf{O}^\top, & g = ai, \mathbf{O} \in O(d), \\ \text{expm}(\mathbf{M} \logm(\mathbf{P})\mathbf{M}^\top), & g = le, \mathbf{M} \in \mathbb{R}^{d \times d}, \\ \mathcal{L}^{-1}(\lfloor L(\mathbf{P}) \rfloor + \exp(D(L(\mathbf{P})))), & g = lc, \mathbf{M} \in \mathbb{R}^{d \times d}, \end{cases} \quad (33)$$

with

$$L(\mathbf{P}) = \mathbf{M} (\lfloor \mathcal{L}(\mathbf{P}) \rfloor + \lfloor \mathcal{L}(\mathbf{P}) \rfloor^\top + D(\mathcal{L}(\mathbf{P}))) \mathbf{M}^\top, \quad (34)$$

is a gyro homomorphism, *i.e.*,

$$\text{hom}_g(\mathbf{P}) \oplus_g \text{hom}_g(\mathbf{Q}) = \text{hom}_g(\mathbf{P} \oplus_g \mathbf{Q}), \quad (35)$$

$$t \otimes_g \text{hom}_g(\mathbf{P}) = \text{hom}_g(t \otimes_g \mathbf{P}) \quad \forall \mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}, t > 0. \quad (36)$$

AIM case ($g = ai$).

The \oplus_{ai} and \otimes_{ai} are defined by:

$$\mathbf{P} \oplus_{ai} \mathbf{Q} = \mathbf{P}^{\frac{1}{2}} \mathbf{Q} \mathbf{P}^{\frac{1}{2}}. \quad (37)$$

$$t \otimes_{ai} \mathbf{P} = \mathbf{P}^t \quad (38)$$

Let the isometry $\text{hom}_{ai}(\mathbf{P}) = \mathbf{O}\mathbf{P}\mathbf{O}^\top$ with $\mathbf{O} \in O(d)$ (orthogonal group). Since the identity element is $e = \mathbf{I}$ and $\text{hom}_{ai}(e) = \mathbf{O}\mathbf{I}\mathbf{O}^\top = \mathbf{I} = e$, the sufficient condition in Theorem 3.2 applies: an isometry fixing e is a gyro homomorphism.

For completeness, we verify both operations explicitly.

We begin by showing that $\text{hom}_{ai}(\cdot)$ satisfies Eq. (35). let $\text{hom}_{ai}(\mathbf{P}) = \mathbf{O}\mathbf{P}\mathbf{O}^\top$, with any $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, then we have

$$\begin{aligned} \text{hom}_{ai}(\mathbf{P}) \oplus_{ai} \text{hom}_{ai}(\mathbf{Q}) &\stackrel{(1)}{=} (\mathbf{O}\mathbf{P}\mathbf{O}^\top)^{\frac{1}{2}} \mathbf{O}\mathbf{Q}\mathbf{O}^\top (\mathbf{O}\mathbf{P}\mathbf{O}^\top)^{\frac{1}{2}} \\ &\stackrel{(2)}{=} \mathbf{O}\mathbf{P}^{\frac{1}{2}} \mathbf{O}^\top \mathbf{O}\mathbf{Q}\mathbf{O}^\top \mathbf{O}\mathbf{P}^{\frac{1}{2}} \mathbf{O}^\top \\ &= \mathbf{O}\mathbf{P}^{\frac{1}{2}} \mathbf{Q} \mathbf{P}^{\frac{1}{2}} \mathbf{O}^\top \\ &= \text{hom}_{ai}(\mathbf{P} \oplus_{ai} \mathbf{Q}). \end{aligned} \quad (39)$$

The derivation of Eq. (39) follows.

(1) follow from Eqs. (8) and (37).

(2) follows from the fact that \mathbf{P} is an SPD matrix and \mathbf{O} is an orthogonal matrix.

Now, we proof that $\text{hom}_{ai}(\cdot)$ satisfies Eq. (36). For the \otimes_{ai} , we have

$$\begin{aligned} t \otimes_{ai} \text{hom}_{ai}(\mathbf{P}) &\stackrel{(1)}{=} (\mathbf{O}\mathbf{P}\mathbf{O}^\top)^t \\ &\stackrel{(2)}{=} \mathbf{O}\mathbf{P}^t \mathbf{O}^\top \\ &= \text{hom}_{ai}(t \otimes_{ai} \mathbf{P}). \end{aligned} \quad (40)$$

The derivation of Eq. (40) follows.

(1) follow from Eqs. (8) and (38).

(2) follows from the fact that \mathbf{P} is an SPD matrix and \mathbf{O} is an orthogonal matrix.

LEM case ($g = le$).

The \oplus_{le} and \otimes_{le} are defined by:

$$\mathbf{P} \oplus_{le} \mathbf{Q} = \exp(\logm(\mathbf{P}) + \logm(\mathbf{Q})), \quad (41)$$

$$t \otimes_{le} \mathbf{P} = \mathbf{P}^t \quad (42)$$

We begin by showing that $\hom_{le}(\cdot)$ satisfies Eq. (35). For the \oplus_{le} , with any $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, we have

$$\begin{aligned} \hom_{le}(\mathbf{P}) \oplus_{le} \hom_{le}(\mathbf{Q}) &\stackrel{(1)}{=} \exp(\mathbf{M} \logm(\mathbf{P}) \mathbf{M}^\top + \mathbf{M} \logm(\mathbf{Q}) \mathbf{M}^\top) \\ &= \exp(\mathbf{M}(\logm(\mathbf{P}) + \logm(\mathbf{Q})) \mathbf{M}^\top) \\ &= \hom_{le}(\mathbf{P} \oplus_{le} \mathbf{Q}). \end{aligned} \quad (43)$$

The derivation of Eq. (43) follows.

(1) follow from Eqs. (8) and (41).

For \otimes_{le} , we have

$$\begin{aligned} t \otimes_{le} \hom_{le}(\mathbf{P}) &\stackrel{(1)}{=} (\exp(\mathbf{M} \logm(\mathbf{P}) \mathbf{M}^\top))^t \\ &\stackrel{(2)}{=} \exp(t \mathbf{M} \logm(\mathbf{P}) \mathbf{M}^\top) \\ &= \hom_{le}(t \otimes_{le} \mathbf{P}). \end{aligned} \quad (44)$$

Reduction to AIM for orthogonal $\mathbf{M} = \mathbf{O} \in \mathrm{O}(d)$.

For the \oplus_{le} , with any $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, $\mathbf{O} \in \mathrm{O}(d)$ we have

$$\begin{aligned} \hom_{le}(\mathbf{P}) \oplus_{le} \hom_{le}(\mathbf{Q}) &\stackrel{(1)}{=} \exp(\mathbf{O}(\logm(\mathbf{P}) + \logm(\mathbf{Q})) \mathbf{O}^\top) \\ &\stackrel{(2)}{=} \mathbf{O} \exp((\logm(\mathbf{P}) + \logm(\mathbf{Q}))) \mathbf{O}^\top \\ &= \hom_{le}(\mathbf{P} \oplus_{ai} \mathbf{Q}). \end{aligned} \quad (45)$$

The derivation of Eq. (45) follows.

(1) follow from Eqs. (41) and (43).

(2) follows from the fact that \mathbf{P} is an SPD matrix and \mathbf{O} is an orthogonal matrix.

For the \otimes_{le} , we have

$$\begin{aligned} t \otimes_{le} \hom_{le}(\mathbf{P}) &\stackrel{(1)}{=} \exp(t \mathbf{O} \logm(\mathbf{P}) \mathbf{O}^\top) \\ &\stackrel{(2)}{=} \mathbf{O} \exp(t \logm(\mathbf{P})) \mathbf{O}^\top \\ &= \hom_{le}(t \otimes_{le} \mathbf{P}). \end{aligned} \quad (46)$$

The derivation of Eq. (46) follows.

(1) follow from Eqs. (42) and (44).

(2) follows from the fact that \mathbf{P} is an SPD matrix and \mathbf{O} is an orthogonal matrix.

LCM case ($g = lc$).

The \oplus_{lc} and \otimes_{lc} are defined by:

$$t \otimes_{lc} \mathbf{P} = \mathcal{L}^{-1}(t \lfloor \mathcal{L}(\mathbf{P}) \rfloor + \mathbb{D}(\mathcal{L}(\mathbf{P})))^t, \quad (47)$$

$$\mathbf{P} \oplus_{lc} \mathbf{Q} = \mathcal{L}^{-1}(\lfloor \mathcal{L}(\mathbf{P}) \rfloor + \lfloor \mathcal{L}(\mathbf{Q}) \rfloor + \mathbb{D}(\mathcal{L}(\mathbf{P})) \mathbb{D}(\mathcal{L}(\mathbf{Q}))). \quad (48)$$

We begin by showing that $\hom_{lc}(\cdot)$ satisfies Eq. (35). With any $\mathbf{P}, \mathbf{Q} \in \mathcal{S}_d^{++}$, for \oplus_{lc} , we can rewrite \oplus_{lc} and \hom_{lc} as

$$\mathbf{P} \oplus_{lc} \mathbf{Q} = \mathcal{L}^{-1}(\exp \mathbb{D}(\log \mathbb{D}(\mathcal{L}(\mathbf{P}))) + \log \mathbb{D}(\mathcal{L}(\mathbf{Q}))), \quad (49)$$

$$\hom_{lc}(\mathbf{P}) = \mathcal{L}^{-1}(\exp \mathbb{D}(L(\mathbf{P}))), \quad (50)$$

where $L(\mathbf{P}) = \mathbf{M}(\lfloor \mathcal{L}(\mathbf{P}) \rfloor + \lfloor \mathcal{L}(\mathbf{P}) \rfloor^\top + \mathbb{D}(\mathcal{L}(\mathbf{P}))) \mathbf{M}^\top$, $\log \mathbb{D}(\mathbf{F})$ and $\exp \mathbb{D}(\mathbf{F})$ are given by

$$\log \mathbb{D}(\mathbf{F}) = \lfloor \mathbf{F} \rfloor + \logm(\mathbb{D}(\mathbf{F})), \quad (51)$$

$$\exp \mathbb{D}(\mathbf{F}) = \lfloor \mathbf{F} \rfloor + \expm(\mathbb{D}(\mathbf{F})), \quad (52)$$

Then we have

$$\begin{aligned}
\text{hom}_{lc}(\mathbf{P}) \oplus_{lc} \text{hom}_{lc}(\mathbf{Q}) &\stackrel{(1)}{=} \mathcal{L}^{-1}(\exp \mathbb{D}(L(\mathbf{P}) + L(\mathbf{Q}))) \\
&\stackrel{(2)}{=} \mathcal{L}^{-1}(\exp \mathbb{D}(L(\mathbf{P} + \mathbf{Q}))) \\
&= \text{hom}_{lc}(\mathbf{P} \oplus_{lc} \mathbf{Q})
\end{aligned} \tag{53}$$

The derivation of Eq. (53) follows.

- (1) follow from Eqs. (8) and (48).
- (2) follow from the properties of $L(\cdot)$.

Hence hom_g is a gyro homomorphism for each $g \in \{ai, le, lc\}$, completing the proof. \square

E.3 Proof of the Thm. 4.2

Proof of Thm. 4.2. The $\tilde{\oplus}_{gr}$ and $\tilde{\otimes}_{gr}$ are defined by:

$$\mathbf{U} \tilde{\oplus}_{gr} \mathbf{V} = \text{expm}([\text{Log}_{\mathbf{I}_{d,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{d,q}])\mathbf{V}, \tag{54}$$

$$t \tilde{\otimes}_{gr} \mathbf{U} = \text{expm}\left(\left[t \text{Log}_{\mathbf{I}_{n,q}}^{gr}, \mathbf{I}_{d,q}\right]\right) \mathbf{I}_{d,q} \tag{55}$$

Let the isometry $\text{hom}_{gr}(\mathbf{U}) = \mathbf{O}\mathbf{U}$ with $\mathbf{O} \in \text{O}(d)$. The reference element $\mathbf{I}_{d,q}$ denotes the canonical subspace spanned by the first q basis vectors. Since $\mathbf{O}\mathbf{I}_{d,q}$ and $\mathbf{I}_{d,q}$ represent the same subspace on $\mathcal{G}(q, d)$, hom_{gr} fixes the identity element in the quotient sense and satisfies the condition of Thm. 3.2.

For completeness, we verify that it preserves both gyroaddition and gyromultiplication below.

we begin by showing that $\text{hom}_{gr}(\cdot)$ satisfies Eq. (35). For any $\mathbf{U}, \mathbf{V} \in \mathcal{G}(q, d)$, we have

$$\begin{aligned}
\text{hom}_{gr}(\mathbf{U}) \tilde{\oplus}_{gr} \text{hom}_{gr}(\mathbf{V}) &\stackrel{(1)}{=} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{O}\mathbf{U}\mathbf{U}^\top\mathbf{O}^\top), \mathbf{I}_{n,q}])\mathbf{O}\mathbf{V} \\
&\stackrel{(2)}{=} \text{expm}([\mathbf{O} \text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top)\mathbf{O}^\top, \mathbf{O}\mathbf{I}_{n,q}\mathbf{O}^\top])\mathbf{O}\mathbf{V} \\
&= \text{expm}(\mathbf{O}[\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}]\mathbf{O}^\top)\mathbf{O}\mathbf{V} \\
&\stackrel{(3)}{=} \mathbf{O} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{O}^\top\mathbf{O}\mathbf{V} \\
&= \mathbf{O} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{V} \\
&= \text{hom}_{gr}(\mathbf{U} \tilde{\oplus}_{gr} \mathbf{V}).
\end{aligned} \tag{56}$$

The derivation of Eq. (56) follows.

- (1) follow from Eqs. (9) and (54).
- (2) follows from the fact that $\text{Log}_{\mathbf{O}\mathbf{I}_{n,q}\mathbf{O}^\top}^{gr}(\mathbf{O}\mathbf{U}\mathbf{U}^\top\mathbf{O}^\top) = \mathbf{O} \text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top)\mathbf{O}^\top$, and for $\mathbf{O} = \begin{bmatrix} \mathbf{O}_q & 0 \\ 0 & \mathbf{O}_{d-q} \end{bmatrix}$, $\mathbf{O}\mathbf{I}_{n,q}\mathbf{O}^\top = \mathbf{I}_{n,q}$.
- (3) follows from the fact that \mathbf{O} is an orthogonal matrix.

Now, we proof that $\text{hom}_{gr}(\cdot)$ satisfies Eq. (36). The differential homomorphism $\Phi : \tilde{\mathcal{G}}(q, d) \rightarrow \mathcal{G}(q, d)$, $\mathbf{U} \rightarrow \mathbf{U}\mathbf{U}^\top$ exists between $\tilde{\mathcal{G}}(q, d)$ and $\mathcal{G}(q, d)$, and $\tilde{\otimes}_{gr}$ is derived from \otimes_{gr} via this differential homomorphism. Thus, to prove that $\tilde{\otimes}_{gr}$ satisfies Eq. (36), it suffices to show that \otimes_{gr} satisfies Eq. (36). The \otimes_{gr} is defined by:

$$t \otimes_{gr} \mathbf{U} = \text{expm}([t \bar{\mathbf{U}}, \mathbf{I}_{d,q}]) \mathbf{I}_{d,q} \text{expm}([-t \bar{\mathbf{U}}, \mathbf{I}_{d,q}]) \tag{57}$$

For \otimes_{gr} , we have

$$\begin{aligned}
t \otimes_{gr} \text{hom}_{gr}(\mathbf{U}\mathbf{U}^\top) &= (t \tilde{\otimes}_{gr} \text{hom}_{gr}(\mathbf{U}))(t \tilde{\otimes}_{gr} \text{hom}_{gr}(\mathbf{U}))^\top \\
&\stackrel{(1)}{=} \text{expm}(t[\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{O}\mathbf{U}\mathbf{U}^\top \mathbf{O}^\top), \mathbf{I}_{n,q}])\mathbf{I}_{n,q} \text{expm}(t[\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{O}\mathbf{U}\mathbf{U}^\top \mathbf{O}^\top), \mathbf{I}_{n,q}]) \\
&\stackrel{(2)}{=} \mathbf{O} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{O}^\top \mathbf{I}_{n,q} \mathbf{O} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{O}^\top \\
&= \mathbf{O} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{I}_{n,q} \text{expm}([\text{Log}_{\mathbf{I}_{n,q}}^{gr}(\mathbf{U}\mathbf{U}^\top), \mathbf{I}_{n,q}])\mathbf{O}^\top \\
&= \text{hom}_{gr}(t \otimes_{gr} \mathbf{U}\mathbf{U}^\top).
\end{aligned} \tag{58}$$

Since \otimes_{gr} satisfies Eq. (36), we can proof $\tilde{\otimes}_{gr}$ satisfies Eq. (36). \square

E.4 Proof of the Thm. 4.3

Proof of Thm. 4.3. The $\tilde{\oplus}_{psd,g}$ and $\otimes_{psd,g}$ are defined by:

$$(\mathbf{U}_P, \mathbf{S}_P) \oplus_{psd,g} (\mathbf{U}_Q, \mathbf{S}_Q) = (\mathbf{U}_P \tilde{\oplus}_{gr} \mathbf{U}_Q, \mathbf{S}_P \oplus_g \mathbf{S}_Q), \tag{59}$$

$$t \otimes_{psd,g} (\mathbf{U}_P, \mathbf{S}_P) = (t \tilde{\otimes}_{gr} \mathbf{U}_P, t \otimes_g \mathbf{S}_P) \tag{60}$$

we begin by showing that $\text{hom}_{psd,g}$ satisfies Eq. (35). As shown in Eq. (10) For any $(\mathbf{U}_P, \mathbf{S}_P), (\mathbf{U}_Q, \mathbf{S}_Q) \in \tilde{\mathcal{G}}(q, d) \times \mathcal{S}_q^{++}$, we have:

$$\begin{aligned}
\text{hom}_{psd,g}((\mathbf{U}_P, \mathbf{S}_P) \oplus_{psd,g} (\mathbf{U}_Q, \mathbf{S}_Q)) &\stackrel{(1)}{=} \text{hom}_{psd,g}(\mathbf{U}_P \tilde{\oplus}_{gr} \mathbf{U}_Q, \mathbf{S}_P \oplus_g \mathbf{S}_Q) \\
&= (\text{hom}_{gr}(\mathbf{U}_P \tilde{\oplus}_{gr} \mathbf{U}_Q), \text{hom}_g(\mathbf{S}_P \oplus_g \mathbf{S}_Q)) \\
&\stackrel{(2)}{=} (\text{hom}_{gr}(\mathbf{U}_P) \tilde{\oplus}_{gr} \text{hom}_{gr}(\mathbf{U}_Q), \text{hom}_g(\mathbf{S}_P) \oplus_g \text{hom}_g(\mathbf{S}_Q)) \\
&\stackrel{(3)}{=} (\text{hom}_{gr}(\mathbf{U}_P), \text{hom}_g(\mathbf{S}_P)) \oplus_{psd,g} (\text{hom}_{gr}(\mathbf{U}_Q), \text{hom}_g(\mathbf{S}_Q)) \\
&= \text{hom}_{psd,g}(\mathbf{U}_P, \mathbf{S}_P) \oplus_{psd,g} \text{hom}_{psd,g}(\mathbf{U}_Q, \mathbf{S}_Q).
\end{aligned} \tag{61}$$

The derivation of Eq. (61) follows.

(1) follow from Eqs. (10) and (59).

(2) and (3) follow from the fact that hom_{gr} and hom_g are gyro homomorphisms.

For scalar multiplication, we have:

$$\begin{aligned}
\text{hom}_{psd,g}(t \otimes_{psd,g} (\mathbf{U}_P, \mathbf{S}_P)) &\stackrel{(1)}{=} \text{hom}_{psd,g}(t \tilde{\otimes}_{gr} \mathbf{U}_P, t \otimes_g \mathbf{S}_P) \\
&= (\text{hom}_{gr}(t \tilde{\otimes}_{gr} \mathbf{U}_P), \text{hom}_g(t \otimes_g \mathbf{S}_P)) \\
&\stackrel{(2)}{=} (t \tilde{\otimes}_{gr} \text{hom}_{gr}(\mathbf{U}_P), t \otimes_g \text{hom}_g(\mathbf{S}_P)) \\
&\stackrel{(3)}{=} t \otimes_{psd,g} (\text{hom}_{gr}(\mathbf{U}_P), \text{hom}_g(\mathbf{S}_P)) \\
&= t \otimes_{psd,g} \text{hom}_{psd,g}(\mathbf{U}_P, \mathbf{S}_P).
\end{aligned} \tag{62}$$

The derivation of Eq. (62) follows.

(1) follow from Eqs. (10) and (60).

(2) and (3) follow from the fact that $\text{hom}_{gr}(\cdot)$ and $\text{hom}_g(\cdot)$ are gyro homomorphisms. \square