RETHINKING SPATIAL INDUCTIVE BIAS IN EEG TRANSFORMERS: A GEODESIC PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

The generalization capacity of deep learning models depends heavily on embedding the correct inductive bias for the geometry of the underlying data. EEG electrodes are distributed over a near-spherical scalp surface whose natural metric is geodesic distance, yet existing spatial positional encoding schemes for EEG Transformers rely on Euclidean or coordinate-decomposed assumptions that do not match this geometry. We analyze this mismatch theoretically and prove that widely adopted encoding families either depend on the coordinate frame in which electrodes are recorded or distort the intrinsic scalp metric, each with a quantifiable bound. We then propose a geometric positional encoding framework with two components: one that constructs rotation encodings directly from geodesic distance and is exactly invariant to the choice of coordinate frame, and one that introduces an attention bias whose induced kernel we prove to be strictly positive definite on the sphere. On four EEG downstream tasks, the resulting model improves over existing baselines on 11 of 12 reported metrics. Code is available at `https://anonymous.4open.science/r/georope/`.

1 INTRODUCTION

Deep learning architectures generalize well when the symmetry group built into the model matches the symmetry inherent to the data: convolutional networks encode translation invariance through weight sharing, graph neural networks encode permutation invariance through message passing (Kipf & Welling, 2016; Veličković et al., 2017), and relative positional encodings such as RoPE (Su et al., 2024) and ALiBi (Press et al., 2021) encode translation invariance over one-dimensional sequence order by construction. Geometric deep learning formalizes this correspondence between data symmetry and model inductive bias as a general design principle (Bronstein et al., 2017; 2021).

Brain-computer interfaces built on electroencephalography (EEG) support communication and control for paralyzed patients (McFarland & Wolpaw, 2011), neurological rehabilitation (Daly & Wolpaw, 2008), and automated seizure detection (Jing et al., 2023), yet the deep networks underlying these systems have not established this correspondence for their most basic geometric input: electrode position. EEG electrodes placed according to the international 10-20 system occupy fixed anatomical sites on a surface well approximated by a sphere $\mathbb{S}^2$. The coordinates used to record them depend on a choice of head frame, and for positions normalized onto a common sphere a change of frame acts by $\mathrm{SO}(3)$, a group distinct from the translation and permutation groups addressed by prior inductive-bias designs. Recent EEG foundation models incorporate electrode position through Euclidean coordinates or per-axis embeddings (Jiang et al., 2024; Wang et al., 2025; Fang et al., 2026; Döner et al., 2026), and these constructions either depend on the coordinate frame or measure proximity through the head rather than along the scalp.

We analyze three encoding families in current use (Section 4). Replacing the RoPE angle with the Euclidean chord length is frame-invariant but distorts the metric, with an error of order ψ^3 relative to the geodesic distance $d_{\mathbb{S}^2}$ that compresses long-range separations. Encoding the azimuthal difference is invariant only under rotations about the polar axis and has an $O(1)$ discontinuity at the poles that no branch convention removes. Per-axis Cartesian encodings, whether absolute as in

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

NeurIPT (Fang et al., 2026) or in relative rotary form, depend on the frame; a $\pi/4$ rotation changes them by an amount bounded away from zero.

To address this, we adopt the geodesic distance $d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$ as the invariant scalar underlying spatial encoding (Section 3), a quantity available in closed form and frame-invariant by construction, unlike any per-axis coordinate representation. Building on this scalar, we introduce GeoRoPE, which substitutes the geodesic distance for the RoPE rotation angle and is therefore exactly frame-invariant, and GDA, an additive attention bias built from the same distance whose induced kernel we prove to be strictly positive definite on $\mathbb{S}^2$ through an explicit Hilbert-space embedding of the sphere. Both components thus enter the attention logit, one through the rotary term and one through an additive bias, without any dependence on the coordinate frame.

This paper contributes:

1. A geometric analysis of three EEG spatial encoding families: the Euclidean surrogate is frame-invariant but metrically distorted, azimuthal encoding is discontinuous at the poles, and per-axis Cartesian encodings are frame-dependent, each with an explicit bound (Section 4).
2. GeoRoPE and GDA, geodesic encodings that are exactly frame-invariant and Lipschitz-stable under electrode displacement, with GDA inducing a strictly positive-definite kernel on $\mathbb{S}^2$ (Sections 3 and 4).

We evaluate this framework in Section 5 on four EEG downstream tasks covering cognitive load, event-related potentials, sleep staging, and motor imagery.

2 Related Work

2.1 EEG Foundation Models

Self-supervised EEG representation learning follows three main lines: modeling EEG as a multivariate time series with transformer or contrastive objectives (Kostas et al., 2021); tokenizing continuous signals for language-style pretraining (Jiang et al., 2024); and architectures that separate spatial and temporal attention explicitly (Wang et al., 2025). BIOT (Yang et al., 2023) and EEGPT (Wang et al., 2024) establish that such pretrained representations transfer across datasets and tasks. NeurIPT (Fang et al., 2026) incorporates electrode geometry by encoding each of the three spatial coordinates independently and concatenating the result, combined with amplitude-aware masked pretraining, and reports state-of-the-art accuracy on several downstream tasks. LUNA (Döner et al., 2026) instead compresses a variable number of electrodes into a fixed-size latent space via learned queries and cross-attention, avoiding explicit geometric encoding altogether; Yi et al. (2023) pursue a related geometry-aware strategy for cross-topology transfer. None of these models establishes a theoretical invariance guarantee for its spatial encoding; NeurIPT's per-axis embedding is frame-dependent, as Proposition 4 shows.

2.2 Positional Encoding and Geometric Priors

RoPE (Su et al., 2024) and ALiBi (Press et al., 2021) are inductive biases for the one-dimensional translation group; porting either onto a sphere is not automatic. Spherical CNNs (Cohen et al., 2018) achieve exact rotation equivariance for dense spherical signals such as panoramic images, but the construction requires a regular sampling grid, which EEG electrode sets, sparse and irregularly placed, do not provide. Geometric deep learning surveys (Bronstein et al., 2017; 2021) establish the general correspondence between symmetry group and inductive bias that motivates our analysis; we specialize this correspondence to identify which pairwise scalar yields a frame-invariant and metrically faithful electrode-pair encoding, and by how much current practice deviates from it. Geodesic distances have also been used to build self-attention on learned manifolds of 3D point clouds (Li et al., 2022), and attention has been formulated directly on spherical domains for dense spherical signals (Bonev et al., 2026); EEG montages are instead sparse, fixed sensor sets, on which we analyze pairwise encodings.

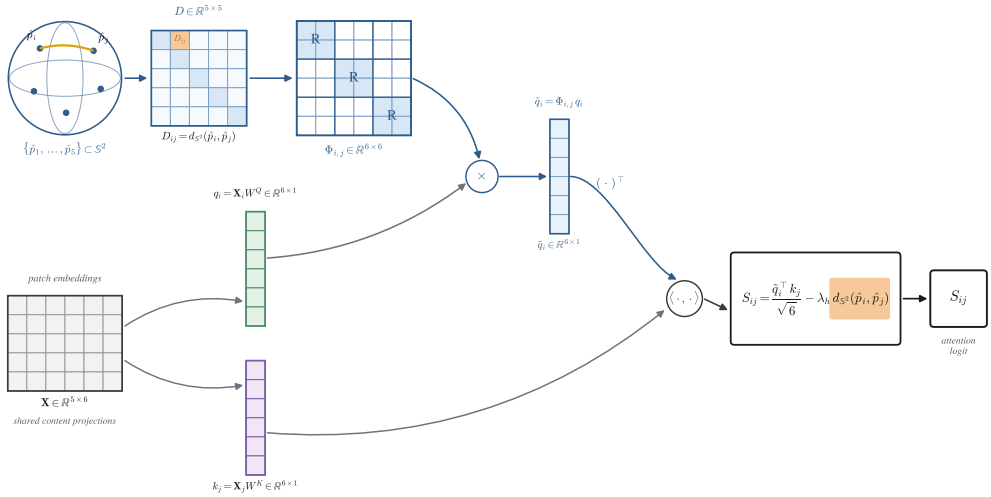

Figure 1: Overview of GeoRoPE and GDA acting jointly on a toy example with five electrodes ($\hat{p}_1, \ldots, \hat{p}_5 \in \mathbb{S}^2$, left) and embedding dimension $D = 6$. The geodesic arc between an electrode pair, e.g., $\hat{p}_i$ and $\hat{p}_j$ (highlighted), populates the pairwise distance matrix $D_{ij} = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j) \in \mathbb{R}^{5\times5}$. Each entry D_{ij} parameterizes a block-diagonal rotation matrix $\Phi_{i,j} \in \mathbb{R}^{6\times6}$, consisting of three 2×2 rotation blocks R, one per frequency pair; GeoRoPE applies $\Phi_{i,j}$ to the query vector $q_i = \mathbf{X}_i W^Q$, obtained from the shared patch embeddings $\mathbf{X} \in \mathbb{R}^{5\times6}$, to produce the rotated query $\tilde{q}_i = \Phi_{i,j} q_i$. The attention logit S_{ij} combines the rotated inner product $\tilde{q}_i^\top k_j/\sqrt{6}$ with the key vector $k_j = \mathbf{X}_j W^K$ and the GDA distance bias $-\lambda_h\, d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$ (highlighted), so that geodesic geometry enters the score through both the rotation and the additive term.

3 Methodology

We propose a unified geometric positional encoding framework for electrodes placed on the approximately spherical scalp. Figure 1 previews how the two components introduced in this section, GeoRoPE and GDA, act jointly within a single attention logit.

3.1 Preliminaries

EEG electrodes are placed at fixed anatomical landmarks (e.g., 10-20 system), well approximated by a sphere of radius $r \approx 9\,\mathrm{cm}$. Let the measurement position of electrode i be $p_i \in \mathbb{R}^3$, with normalized coordinate $\hat{p}_i = p_i/\|p_i\|_2 \in \mathbb{S}^2$. Adjacent electrodes share overlapping cortical sources along the scalp, so the relevant measure of proximity is the geodesic distance on the sphere,

$$d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j) = \arccos(\hat{p}_i \cdot \hat{p}_j) \in [0, \pi], \tag{1}$$

rather than the Euclidean distance through the skull. The geodesic distance satisfies $d_{\mathbb{S}^2}(Q\hat{p}_i, Q\hat{p}_j) = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$ for every $Q \in \mathrm{SO}(3)$, so any encoding expressed solely through $d_{\mathbb{S}^2}$ is independent of the coordinate frame in which the montage is recorded; Section 4 makes the scope of this invariance precise.

3.2 Geodesic Rotary Positional Encoding (GeoRoPE)

We define the 2D rotation matrix as:

$$R(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \tag{2}$$

Standard RoPE (Su et al., 2024) assigns block-diagonal matrices $\Phi_m = \bigoplus_{t=1}^{D/2} R(\omega_t m)$ to scalar positions, where $\omega_t = b^{-2(t-1)/D}$ and $b = 10000$. We take the geodesic distance as the sole scalar

parameter to construct GeoRoPE. For a given head h and embedding dimension D, the geometric attention formulation is:

$$A_{ij}^{\mathrm{GeoRoPE}} = \sum_{t=1}^{D/2} \langle q_i^{(t)}, R(\omega_t d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)) k_j^{(t)} \rangle. \tag{3}$$

Rotating the query instead of the key, $\langle R(\alpha)q, k\rangle = \langle q, R(-\alpha)k\rangle$, only reverses the sign of the angle, and every result in Section 4 holds for either convention. By construction, $A_{ij}^{\mathrm{GeoRoPE}}$ is frame-invariant and continuous on $\mathbb{S}^2 \times \mathbb{S}^2$. Because $d_{\mathbb{S}^2}$ is a pairwise quantity rather than a difference of per-token positions, the rotation cannot be factorized into per-token rotations as in RoPE. Writing $R(\alpha) = \cos\alpha\, I_2 + \sin\alpha\, J$ with $J = R(\pi/2)$, the logits require the two block Gram matrices $\langle q_i^{(t)}, k_j^{(t)}\rangle$ and $\langle q_i^{(t)}, Jk_j^{(t)}\rangle$, weighted entrywise by $\cos(\omega_t d_{\mathbb{S}^2})$ and $\sin(\omega_t d_{\mathbb{S}^2})$, a constant-factor overhead over standard attention.

3.3 Geodesic Distance Attention (GDA)

We define a complementary additive scheme, Geodesic Distance Attention (GDA), which extends ALiBi (Press et al., 2021) to the sphere. With a learnable per-head slope $\lambda_h > 0$, the attention logit is

$$A_{ij}^{\mathrm{GDA}} = \frac{A_{ij}^{\mathrm{GeoRoPE}}}{\sqrt{D}} - \lambda_h\, d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j), \tag{4}$$

with $A_{ij}^{\mathrm{GeoRoPE}}$ from Eq. (3); the full model uses this combined logit. Spatial structure thus enters the attention through both the rotation and the additive distance bias.

4 Theoretical Properties of Geometric Encodings

This section separates two properties of a pairwise spatial encoding: independence from the coordinate frame in which electrode positions are expressed, and faithfulness to the intrinsic metric of the scalp. The Euclidean chord surrogate has the first property but distorts the second; azimuthal and per-axis Cartesian encodings lose the first. We then show that the GDA bias induces a strictly positive-definite kernel on $\mathbb{S}^2$.

Scope of the invariance. Throughout, $\mathrm{SO}(3)$ invariance refers to a *passive* change of frame, in which every position $\hat{p}_i$ is replaced by $Q\hat{p}_i$ with one common $Q \in \mathrm{SO}(3)$; a pairwise encoding E is *frame-invariant* if $E(Q\hat{p}_i, Q\hat{p}_j) = E(\hat{p}_i, \hat{p}_j)$ for all $Q \in \mathrm{SO}(3)$ and all i, j. This does not assume that the brain or the EEG signal is rotationally symmetric, nor does it make the model invariant to physically moving a scalp pattern, since electrodes keep their identities and signals. It guarantees that the spatial encoding depends on the montage itself rather than on the recording convention (head-coordinate axes, template versus digitized positions after rigid alignment), which varies across datasets and devices. When all subjects of a dataset share one template montage, no such rotation occurs, so within such a dataset frame invariance is a well-posedness property of the encoding, and what distinguishes geodesic encodings there is the proximity structure they impose, analyzed in Section 4.1. Because $d_{\mathbb{S}^2}$ is preserved by every orthogonal map, encodings built from $d_{\mathbb{S}^2}$ alone are invariant under $\mathrm{O}(3)$, reflections included, so left–right lateralized information reaches the network through channel identity and signal content rather than through the spatial encoding.

We rely on the following standing assumption. It holds, with constants depending on the trained weights, whenever queries and keys are linear projections of LayerNorm outputs, since a LayerNorm of width d with gain γ and bias β satisfies $\|\mathrm{LN}(x)\|_2 \le \sqrt{d}\,\|\gamma\|_\infty + \|\beta\|_2$.

Assumption 1 (Bounded representations)**.** *There exist constants $B_q, B_k > 0$ such that $\|q_i^{(t)}\|_2 \le B_q$ and $\|k_j^{(t)}\|_2 \le B_k$ for all i, j, t.*

Since q_i and k_j are produced by the network, upper bounds below hold for every representation admitted by Assumption 1, while lower bounds are attained by some admitted representation, i.e., they bound the worst case over the admissible class.

4.1 Metric Distortion of the Euclidean Surrogate

Replacing $d_{\mathbb{S}^2}$ in Eq. (3) by the chord length $\|\hat{p}_i - \hat{p}_j\|_2$ yields an attention term A_{ij}^{Eucl}. The chord length is itself frame-invariant, so this surrogate does not violate $\mathrm{SO}(3)$ invariance; it differs from $A_{ij}^{\mathrm{GeoRoPE}}$ because it measures proximity through the head rather than along the scalp, and since $\|\hat{p}_i - \hat{p}_j\|_2 = 2\sin(\psi/2) \le \psi$ it compresses long-range separations. The following bound quantifies the discrepancy.

Theorem 1 ($O(\psi^3)$ metric distortion)**.** *Let $\psi = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j) \in [0, \pi]$ and $\Omega = \sum_{t=1}^{D/2} \omega_t$. Under Assumption 1,*

$$\left|A_{ij}^{\mathrm{Eucl}} - A_{ij}^{\mathrm{GeoRoPE}}\right| \le \Omega B_q B_k \big(\psi - 2\sin(\psi/2)\big) \le \frac{\Omega B_q B_k}{24}\psi^3. \tag{5}$$

Proof sketch. By the triangle and Cauchy–Schwarz inequalities over the $D/2$ frequency blocks, $|A_{ij}^{\mathrm{Eucl}} - A_{ij}^{\mathrm{GeoRoPE}}| \le B_q B_k \sum_t \|R(\omega_t\|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t\psi)\|_{\mathrm{op}}$, and $\|R(\alpha) - R(\beta)\|_{\mathrm{op}} = 2|\sin((\alpha-\beta)/2)| \le |\alpha - \beta|$. The law of cosines gives $\|\hat{p}_i - \hat{p}_j\|_2 = 2\sin(\psi/2)$, and a third-order Taylor expansion gives $0 \le \psi - 2\sin(\psi/2) \le \psi^3/24$. Appendix A gives the full proof. □

Remark 2. *For $D = 512$ and $b = 10000$, $\Omega \approx 28.3$; at $\psi = \pi$ the first bound in Eq. (5) equals $\Omega(\pi - 2)B_qB_k \approx 32.3\, B_qB_k$ on the unscaled logit, i.e., before division by $\sqrt{D}$. This is a worst-case bound; the discrepancy realized by a trained network is an empirical quantity.*

4.2 Frame Dependence and Discontinuity of Coordinate-Based Encodings

Azimuthal encoding. Naive spherical coordinate encoding (NSCE) writes $\hat{p}_i = (\sin\theta_i\cos\phi_i, \sin\theta_i\sin\phi_i, \cos\theta_i)$ and replaces $d_{\mathbb{S}^2}$ in Eq. (3) by an azimuth difference $\Delta\phi_{ij} \equiv \phi_i - \phi_j \pmod{2\pi}$, under any fixed choice of representative, giving $A_{ij}^{\mathrm{NSCE}} = \sum_t \langle q_i^{(t)}, R(\omega_t \Delta\phi_{ij}) k_j^{(t)} \rangle$. It is invariant only under rotations about the polar axis, and it is discontinuous at the poles.

Proposition 3 (Irreducible pole discontinuity)**.** *There exist representations satisfying Assumption 1 and sequences $\hat{p}_i^{(n)}, \hat{p}_j^{(n)} \in \mathbb{S}^2$ converging to the North pole, with $d_{\mathbb{S}^2}(\hat{p}_i^{(n)}, \hat{p}_j^{(n)}) \to 0$, such that $\lim_{n\to\infty} |A_{ij}^{\mathrm{NSCE}} - A_{ij}^{\mathrm{GeoRoPE}}| = 2B_qB_k$ for every choice of representative of $\Delta\phi_{ij}$. Consequently, A^{NSCE} admits no continuous extension to $\mathbb{S}^2 \times \mathbb{S}^2$.*

Proof sketch. Place $\hat{p}_i^{(n)}$ on the meridian $\phi = 0$ and $\hat{p}_j^{(n)}$ on $\phi = \pi$ at polar angle $\theta_n \to 0$, so that $d_{\mathbb{S}^2} = 2\theta_n \to 0$ while $\Delta\phi_{ij} \in \pi + 2\pi\mathbb{Z}$. Since $\omega_1 = 1$, the first frequency block of NSCE equals $R(\pi) = -I_2$ for every representative, whereas that of GeoRoPE tends to I_2; aligned q_i, k_j supported on this block give $2B_qB_k$. Two points on a common meridian give the NSCE limit $+B_qB_k$ at the same limiting configuration. Appendix A gives the full proof. □

Per-axis Cartesian encoding. NeurIPT (Fang et al., 2026) embeds each coordinate separately with an absolute sinusoidal code, $\mathrm{PE}(\hat{p}) = \mathrm{Concat}(\mathrm{PE}_x(x), \mathrm{PE}_y(y), \mathrm{PE}_z(z))$ with d_{ax} channels per axis and frequencies $\omega_t^{\mathrm{ax}} = b^{-2(t-1)/d_{\mathrm{ax}}}$, $t = 1, \ldots, d_{\mathrm{ax}}/2$. Its relative (rotary) analogue, the per-axis counterpart of GeoRoPE, is $A_{ij}^{\mathrm{3D}} = \sum_{c\in\{x,y,z\}} \sum_t \langle q_i^{(t,c)}, R(\omega_t^{\mathrm{ax}} \Delta c_{ij}) k_j^{(t,c)} \rangle$ with $\Delta c_{ij} = c_i - c_j$, where each head splits its channels equally across the three axes. Both depend on the frame.

Proposition 4 (Frame dependence of per-axis encodings)**.** *Let $\hat{p}_i = (1,0,0)$, $\hat{p}_j = (0,1,0)$, and let $Q \in \mathrm{SO}(3)$ be the rotation by $\pi/4$ about the z-axis. Then $d_{\mathbb{S}^2}(Q\hat{p}_i, Q\hat{p}_j) = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$, whereas (a) for some representations satisfying Assumption 1, $|A_{ij}^{\mathrm{3D}}(Q\hat{p}_i, Q\hat{p}_j) - A_{ij}^{\mathrm{3D}}(\hat{p}_i, \hat{p}_j)| = 2B_qB_k \sin(1/2) \approx 0.96\, B_qB_k$; and (b) for every $d_{\mathrm{ax}} \ge 2$, the similarity $\kappa(\hat{p}, \hat{p}') = \langle \mathrm{PE}(\hat{p}), \mathrm{PE}(\hat{p}') \rangle$ induced by the absolute embedding satisfies $\kappa(Q\hat{p}_i, Q\hat{p}_j) - \kappa(\hat{p}_i, \hat{p}_j) \ge 1 + \cos\sqrt{2} - 2\cos 1 \approx 0.075$.*

Proof sketch. The coordinate differences change from $(1,-1,0)$ to $(\sqrt{2},0,0)$. For (a), concentrating q_i, k_j on the first y-block gives the discrepancy $\|I_2 - R(-1)\|_{\mathrm{op}} B_qB_k = 2\sin(1/2)B_qB_k$. For (b), $\kappa = \sum_c \sum_t \cos(\omega_t^{\mathrm{ax}} \Delta c)$, so the change is $-\sum_t f(\omega_t^{\mathrm{ax}})$ with $f(\omega) = 2\cos\omega - \cos(\sqrt{2}\omega) - 1$,

and Taylor bounds give $f(\omega) \le -\omega^4(1/12 - \omega^2/90) < 0$ on $(0, 1]$. Appendix A gives the full proof. □

An absolute embedding changes with the frame, so any invariance of the downstream network must be learned; part (b) shows that even the parameter-free similarity it induces is frame-dependent. Beyond rigid changes of frame, electrode positions also vary non-rigidly with cap placement and head shape, and geodesic encodings change continuously under such displacements.

Corollary 5 (Stability under electrode displacement)**.** *Under Assumption 1, for all $\hat{p}_i, \hat{p}_j, \hat{p}'_i, \hat{p}'_j \in \mathbb{S}^2$, with $\varepsilon = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}'_i) + d_{\mathbb{S}^2}(\hat{p}_j, \hat{p}'_j)$, $|A_{ij}^{\mathrm{GeoRoPE}}(\hat{p}'_i, \hat{p}'_j) - A_{ij}^{\mathrm{GeoRoPE}}(\hat{p}_i, \hat{p}_j)| \le \Omega B_q B_k \, \varepsilon$, and the GDA bias changes by at most $\lambda_h \varepsilon$.*

By Proposition 3, no bound of this form holds for NSCE.

4.3 Positive Definiteness of the GDA Kernel

Under softmax, the GDA bias multiplies the unnormalized attention weights by $K_{ij} = \exp(-\lambda_h d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j))$. We show that K is strictly positive definite for every slope, so GDA imposes a valid spatial similarity on any montage.

Theorem 6 (Strict positive definiteness of the GDA kernel)**.** *Let $\hat{p}_1, \dots, \hat{p}_n \in \mathbb{S}^2$ be distinct and $\lambda_h > 0$. The matrix $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = \exp(-\lambda_h d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j))$ is strictly positive definite.*

Proof sketch. For u uniform on $\mathbb{S}^2$, $\Pr[\mathrm{sgn}(u \cdot x) \ne \mathrm{sgn}(u \cdot y)] = d_{\mathbb{S}^2}(x, y)/\pi$, so $\varphi(x) = \frac{\sqrt{\pi}}{2} \mathrm{sgn}(u \cdot x) \in L^2(\mathbb{S}^2)$ satisfies $d_{\mathbb{S}^2}(x, y) = \|\varphi(x) - \varphi(y)\|_{L^2}^2$. Hence $K_{ij} = \exp(-\lambda_h \|\varphi(\hat{p}_i) - \varphi(\hat{p}_j)\|^2)$ is the Gaussian kernel evaluated at $\varphi(\hat{p}_1), \dots, \varphi(\hat{p}_n)$. These points are distinct, as $d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j) > 0$ for $i \ne j$, and span a finite-dimensional subspace, on which the Gaussian kernel is strictly positive definite (Wendland, 2005). Appendix A gives the full proof. □

Remark 7. *(i) The squared form $\exp(-\lambda d_{\mathbb{S}^2}^2)$ is not positive definite for any $\lambda \in (0, 0.247)$, already on the four points $(\pm 1, 0, 0), (0, \pm 1, 0)$ (Appendix A), in line with the fact that geodesic Gaussian kernels are positive definite for all bandwidths only on flat spaces (Feragen et al., 2015). (ii) Positive definiteness does not by itself favor the geodesic over the chord: $\exp(-\lambda \|\hat{p}_i - \hat{p}_j\|_2)$ is also strictly positive definite, as a restriction of the Laplace kernel on $\mathbb{R}^3$; the preference for $d_{\mathbb{S}^2}$ rests on the metric argument of Theorem 1. (iii) Theorem 6 concerns the spatial factor K, not the row-normalized and generally non-symmetric content-dependent attention operator.*

5 Experiments

We describe the datasets, baselines, and training settings, then report quantitative results and qualitative visualizations.

5.1 Experimental Setup

We use four EEG datasets. Mental Arithmetic (Zyma et al., 2019) is a binary task, resting state versus mental calculation, with 36 subjects. PhysioP300 (Citi et al., 2010) is an event-related potential task, target versus non-target visual stimuli, with 21 subjects. Sleep-EDFx (Kemp et al., 2000) contains whole-night polysomnography from 82 subjects, labeled with five sleep stages. BCIC-IV-2A (Tangermann et al., 2012) is a motor imagery dataset with nine subjects and four classes: left hand, right hand, feet, and tongue. Table 1 summarizes the key statistics of these datasets.

Table 1: Summary of evaluated EEG datasets.

Dataset	Subjects	Classes	Task Type
Mental Arithmetic	36	2	Cognitive Load
PhysioP300	21	2	Event Related Potential
Sleep-EDFx	82	5	Sleep Staging
BCIC-IV-2A	9	4	Motor Imagery

Table 2: Cross-subject results on four datasets. Baseline results are taken from Fang et al. (2026); our results are averaged over five seeds under the same data splits and preprocessing. Best results are in bold.

Mental Arithmetic			
Method	Balanced Accuracy	AUC-PR	AUROC
BIOT	68.75	60.04	75.36
LaBraM	69.09	59.99	77.21
CBraMod	72.56	62.67	79.05
NeurIPT	86.46	78.27	91.11
GeoRoPE (Ours)	**90.21**	**95.56**	**97.92**

Sleep EDFx			
Method	Balanced Accuracy	Cohen Kappa	Weighted F1
BENDR	66.55	66.59	75.07
BIOT	66.22	64.61	74.15
LaBraM	67.71	67.10	75.92
EEGPT	69.17	68.57	76.54
NeurIPT	**70.47**	77.57	87.39
GeoRoPE (Ours)	68.43	**79.09**	**88.39**

PhysioP300			
Method	Balanced Accuracy	Cohen Kappa	AUROC
BENDR	61.14	22.27	65.88
BIOT	54.85	9.68	53.08
LaBraM	64.77	29.35	70.68
EEGPT	65.02	29.99	71.68
NeurIPT	67.31	34.26	76.83
GeoRoPE (Ours)	**76.12**	**49.31**	**80.63**

BCIC-IV-2A			
Method	Balanced Accuracy	Cohen Kappa	Weighted F1
BIOT	47.48	29.97	46.07
LaBraM	48.69	31.59	47.58
CBraMod	51.38	35.18	49.84
NeurIPT	55.04	40.04	53.76
GeoRoPE (Ours)	**61.98**	**49.31**	**59.98**

The baselines are BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), CBraMod (Wang et al., 2025), EEGPT (Wang et al., 2024), and NeurIPT (Fang et al., 2026); NeurIPT uses the per-axis Cartesian embedding analyzed in Proposition 4.

All experiments are conducted on a single NVIDIA RTX 4090 GPU. We train for 50 epochs with batch size 64, the AdamW optimizer, and a cosine annealing learning-rate schedule. All datasets use the cross-subject splits of Fang et al. (2026), so test subjects are never seen during training. We report balanced accuracy on every dataset, together with the dataset-specific metrics in Table 2: AUC-PR, AUROC, Cohen's kappa, and weighted F1. Results are averaged over five random seeds.

5.2 Main Results

Table 2 reports cross-subject results on the four datasets. Our model obtains the best value on 11 of the 12 reported metrics.

On Mental Arithmetic, our model reaches a balanced accuracy of 90.21%, compared with 86.46% for NeurIPT, the strongest baseline. The margin is larger for the threshold-free metrics: 95.56% versus 78.27% AUC-PR and 97.92% versus 91.11% AUROC.

On PhysioP300 and BCIC-IV-2A, our model is best on all three metrics. On BCIC-IV-2A, where decoding depends on localizing sensorimotor rhythms over the scalp, it improves on NeurIPT by 6.94 points in balanced accuracy and 9.27 points in Cohen's kappa.

On Sleep-EDFx, our model has the highest Cohen's kappa (79.09%) and weighted F1 (88.39%), while NeurIPT has a higher balanced accuracy (70.47% versus 68.43%). Sleep staging is strongly class-imbalanced, and the two groups of metrics weight the minority stages differently.

Taken together, these results are consistent with a scalp-intrinsic proximity prior benefiting cross-subject decoding.

5.3 Qualitative Analysis and Visualizations

We visualize, on Mental Arithmetic, the spatial distribution of the model's attributions and the geometry of its learned features.

The top panel of Figure 2 shows raw EEG energy topographies averaged across subjects for the resting and calculation states. The task-related changes are concentrated over central and parietal regions. The bottom panel of Figure 2 shows gradient-based attribution maps from the final layer, which resemble the raw energy distribution. This qualitative resemblance suggests that the model attends to task-relevant regions; it does not by itself establish a causal role for the geodesic bias.

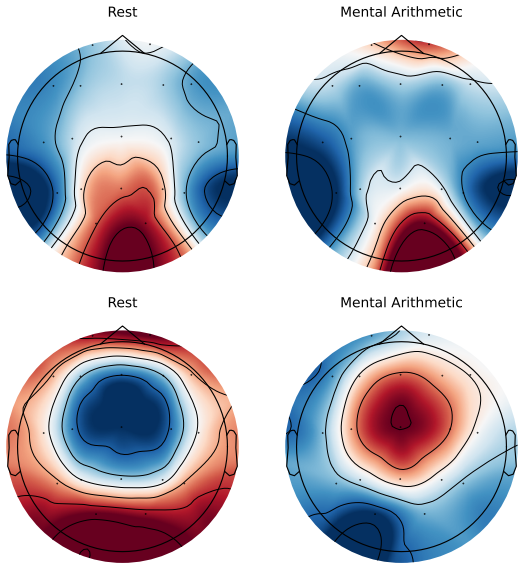

Figure 2: Raw EEG energy topographies (top) and gradient-based attention activation topographies generated by our trained model (bottom), for resting and mental arithmetic states.

Figure 3 projects raw EEG segments and learned features to two dimensions with UMAP. The two classes overlap heavily in the raw data and form separate clusters in the learned features. Such separation shows that the learned features are discriminative, which is expected of a well-trained classifier and is not evidence for a specific geometric mechanism.

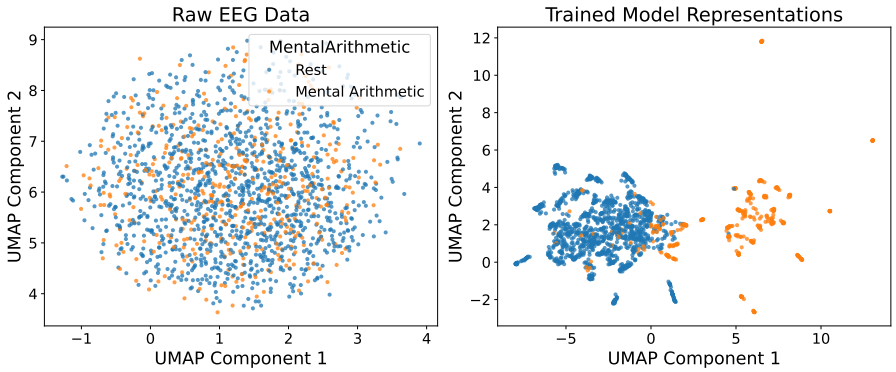

Figure 3: Two dimensional feature manifold projection comparing raw EEG data and trained model representations.

6 Conclusion

This paper studies a geometric mismatch in the spatial positional encodings of EEG Transformers. Electrodes lie on a near-spherical surface whose coordinate description is defined only up to a change of frame, yet existing encoding schemes rely on Euclidean assumptions or Cartesian decompositions that either depend on that frame or distort the scalp metric. We quantified the metric distortion, frame

dependence, and pole discontinuity of these encodings. GeoRoPE, which uses the geodesic distance as its rotary angle, is exactly frame-invariant and Lipschitz-stable under electrode displacement, and GDA adds a geodesic bias whose spatial kernel is strictly positive definite on the sphere. On four downstream tasks, the resulting model improves over existing baselines on 11 of 12 metrics. Frame invariance makes the encoding independent of how a montage is recorded, a prerequisite for pooling EEG across datasets and devices.

AI Use Statement

In this work, we used generative AI tools solely to polish the language of the manuscript, including improving grammar, wording, and clarity of sentences drafted by the authors. We did not use generative AI tools for research ideation, experimental design, the development of the proposed method, theoretical derivations or proofs, code implementation, data analysis, or the generation of any results, figures, or tables reported in this paper; all such tasks were carried out entirely by the authors. All AI-assisted language edits were reviewed by the authors to confirm that they preserved the intended scientific meaning and introduced no factual, technical, or numerical changes. We take responsibility for the final content of this work, including all text, claims, and results.

Reproducibility Statement

Theorem 1, Propositions 3 and 4, Corollary 5, and Theorem 6 are stated with all assumptions explicit (Assumption 1) in Section 4, with proof sketches in the main text and complete, unabridged proofs in Appendix A. The four evaluation datasets are publicly available benchmarks (cited in Section 5), and our results are averaged over five random seeds under a cross-subject split. An anonymized implementation of GeoRoPE and GDA, together with the training and evaluation scripts for our results in Table 2, is available at `https://anonymous.4open.science/r/georope/`.

References

Boris Bonev, Max Rietmann, Andrea Paris, Alberto Carpentieri, and Thorsten Kurth. Attention on the sphere. *Advances in Neural Information Processing Systems*, 38:61662–61688, 2026.

Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Luca Citi, Riccardo Poli, and Caterina Cinel. Documenting, modelling and exploiting p300 amplitude changes due to variable target delays in donchin's speller. *Journal of Neural Engineering*, 7 (5):056006, 2010.

Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

Janis J Daly and Jonathan R Wolpaw. Brain–computer interfaces in neurological rehabilitation. *The Lancet Neurology*, 7(11):1032–1043, 2008.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. Luna: efficient and topology-agnostic foundation model for eeg signal analysis. *Advances in Neural Information Processing Systems*, 38:70682–70708, 2026.

Zitao Fang, Chenxuan Li, Hongting Zhou, Shuyang Yu, Guodong Du, Ashwaq Qasem, Yang Lu, Jing Li, Junsong Zhang, and Sim Kuan Goh. Neuript: Foundation model for neural interfaces. *Advances in Neural Information Processing Systems*, 38:150819–150861, 2026.

Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F Struck, Aline Herlopian, Ioannis Karakis, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during eeg interpretation. *Neurology*, 100(17): e1750–e1762, 2023.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Zhengyu Li, Xuan Tang, Zihao Xu, Xihao Wang, Hui Yu, Mingsong Chen, et al. Geodesic self-attention for 3d point clouds. *Advances in Neural Information Processing Systems*, 35:6190–6203, 2022.

Dennis J McFarland and Jonathan R Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60–66, 2011.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J Miller, Gernot R Müller-Putz, et al. Review of the bci competition iv. *Frontiers in neuroscience*, 6:55, 2012.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press Cambridge, 2005.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36:53875–53891, 2023.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A FULL PROOFS

This appendix contains complete proofs of the results in Section 4.

Proof of Theorem 1. Let $\Delta A_{ij} = A_{ij}^{\mathrm{Eucl}} - A_{ij}^{\mathrm{GeoRoPE}}$. By the definitions of the two attention terms and the triangle inequality,

$$\begin{aligned} &|\Delta A_{ij}| \\ &= \left| \sum_{t=1}^{D/2} \langle q_i^{(t)}, \Big[R(\omega_t \|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t \psi) \Big] k_j^{(t)} \rangle \right| \\ &\le \sum_{t=1}^{D/2} \left| \langle q_i^{(t)}, \Big[R(\omega_t \|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t \psi) \Big] k_j^{(t)} \rangle \right|. \end{aligned} \tag{6}$$

For each t, the Cauchy–Schwarz inequality gives

$$\begin{aligned} &\left| \langle q_i^{(t)}, \Big[R(\omega_t \|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t \psi) \Big] k_j^{(t)} \rangle \right| \\ &\le \|q_i^{(t)}\|_2 \cdot \|R(\omega_t \|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t \psi)\|_{\mathrm{op}} \cdot \|k_j^{(t)}\|_2 \\ &\le B_q B_k \|R(\omega_t \|\hat{p}_i - \hat{p}_j\|_2) - R(\omega_t \psi)\|_{\mathrm{op}}, \end{aligned} \tag{7}$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the spectral norm (operator norm induced by the ℓ_2-norm). For any α, β, $R(\alpha) - R(\beta) = R(\beta)\big(R(\alpha - \beta) - I_2\big)$, and $R(\beta)$ is orthogonal. The two columns of $R(\gamma) - I_2$ are orthogonal and both have length $\sqrt{(\cos\gamma - 1)^2 + \sin^2\gamma}$, so this length is its spectral norm, giving

$$\begin{aligned} &\|R(\alpha) - R(\beta)\|_{\mathrm{op}} \\ =&\|R(\alpha - \beta) - I_2\|_{\mathrm{op}} \\ =&\sqrt{(\cos(\alpha - \beta) - 1)^2 + \sin^2(\alpha - \beta)} \\ =&\sqrt{2 - 2\cos(\alpha - \beta)} = 2\Big|\sin\Big(\frac{\alpha - \beta}{2}\Big)\Big|. \end{aligned} \tag{8}$$

Since $|\sin x| \le |x|$ for all $x \in \mathbb{R}$,

$$\|R(\alpha) - R(\beta)\|_{\mathrm{op}} \le |\alpha - \beta|. \tag{9}$$

Let $\delta(\psi) = \psi - \|\hat{p}_i - \hat{p}_j\|_2$. Substituting this bound into the sum gives

$$|\Delta A_{ij}| \le B_q B_k \sum_{t=1}^{D/2} \omega_t |\delta(\psi)| = \Omega B_q B_k |\delta(\psi)|. \tag{10}$$

It remains to bound $\delta(\psi)$. By the law of cosines in $\mathbb{R}^3$, the chord length relates to the geodesic angle ψ via $\|\hat{p}_i - \hat{p}_j\|_2^2 = 2 - 2\cos\psi = 4\sin^2(\psi/2)$, giving $\|\hat{p}_i - \hat{p}_j\|_2 = 2\sin(\psi/2)$ for $\psi \in [0, \pi]$. We define the error function $f(\psi) = \psi - 2\sin(\psi/2)$.

Taylor's theorem with the Lagrange remainder at $\psi = 0$ gives

$$f(\psi) = f(0) + f'(0)\psi + \frac{f''(0)}{2!}\psi^2 + \frac{f'''(\xi)}{3!}\psi^3, \quad \xi \in (0, \psi).$$

where

$$\begin{aligned} f'(x) &= 1 - \cos(x/2) \implies f'(0) = 0, \\ f''(x) &= \frac{1}{2}\sin(x/2) \implies f''(0) = 0, \\ f'''(x) &= \frac{1}{4}\cos(x/2). \end{aligned}$$

Thus, $f(\psi) = \frac{1}{24}\cos(\xi/2)\psi^3$. Since $\cos(\xi/2) \in (0, 1]$ for $\xi \in (0, \pi)$,

$$0 \le \delta(\psi) = \psi - 2\sin(\psi/2) \le \frac{1}{24}\psi^3. \tag{11}$$

Since $\delta(\psi) \ge 0$, substituting into the inequality for $|\Delta A_{ij}|$ gives both bounds in Eq. (5). □

Proof of Proposition 3. Fix a unit vector $e \in \mathbb{R}^2$ and set $q_i^{(1)} = B_q e$, $k_j^{(1)} = B_k e$, and $q_i^{(t)} = k_j^{(t)} = 0$ for $t \geq 2$; Assumption 1 holds. Let $\theta_n \downarrow 0$, let $\hat{p}_i^{(n)}$ have polar angle θ_n and azimuth 0, and let $\hat{p}_j^{(n)}$ have polar angle θ_n and azimuth π. Then $\hat{p}_i^{(n)} \cdot \hat{p}_j^{(n)} = \cos^2\theta_n - \sin^2\theta_n = \cos 2\theta_n$, so $\psi_n = d_{\mathbb{S}^2}(\hat{p}_i^{(n)}, \hat{p}_j^{(n)}) = 2\theta_n \to 0$ and both sequences converge to the North pole. Since $\omega_1 = b^0 = 1$ and only the first block is non-zero, $A_{ij}^{\mathrm{GeoRoPE}} = B_q B_k\, e^\top R(\psi_n) e = B_q B_k \cos(2\theta_n)$. Every representative of the azimuth difference has the form $\Delta\phi_{ij} = \pi + 2\pi m$ with $m \in \mathbb{Z}$, possibly depending on n, and $R(\pi + 2\pi m) = -I_2$, so $A_{ij}^{\mathrm{NSCE}} = -B_q B_k$. Therefore $|A_{ij}^{\mathrm{NSCE}} - A_{ij}^{\mathrm{GeoRoPE}}| = B_q B_k(1 + \cos 2\theta_n) \to 2B_q B_k$.

For the second claim, let $\hat{p}_i'^{(n)}$ and $\hat{p}_j'^{(n)}$ have azimuth 0 and polar angles θ_n and $2\theta_n$. Both converge to the North pole, and every representative satisfies $\Delta\phi_{ij} \in 2\pi\mathbb{Z}$, so $R(\omega_1 \Delta\phi_{ij}) = I_2$ and $A_{ij}^{\mathrm{NSCE}} = B_q B_k$. The two sequences of pairs converge to the same point of $\mathbb{S}^2 \times \mathbb{S}^2$, while A^{NSCE} tends to $-B_q B_k$ along the first and to $B_q B_k$ along the second; as $B_q B_k > 0$, no continuous extension exists. □

Proof of Proposition 4. Since Q is orthogonal, $Q\hat{p}_i \cdot Q\hat{p}_j = \hat{p}_i \cdot \hat{p}_j$, so $d_{\mathbb{S}^2}(Q\hat{p}_i, Q\hat{p}_j) = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$. With $c = s = 1/\sqrt{2}$, $Q\hat{p}_i = (c, s, 0)$ and $Q\hat{p}_j = (-s, c, 0)$, so the coordinate differences $(\Delta x_{ij}, \Delta y_{ij}, \Delta z_{ij})$ change from $(1, -1, 0)$ to $(\sqrt{2}, 0, 0)$.

(a) Put $M = R(\omega_1^{\mathrm{ax}} \cdot 0) - R(\omega_1^{\mathrm{ax}} \cdot (-1)) = I_2 - R(-1)$, using $\omega_1^{\mathrm{ax}} = 1$. Since $R(a) - R(-a) = 2\sin a\, J$ with $J = R(\pi/2)$, we have $M = R(-\frac{1}{2})\big(R(\frac{1}{2}) - R(-\frac{1}{2})\big) = 2\sin(\frac{1}{2})\, R(\frac{\pi}{2} - \frac{1}{2})$. Let $q_i^{(1,y)} = B_q\, R(\frac{\pi}{2} - \frac{1}{2})e$ and $k_j^{(1,y)} = B_k e$ for a unit vector e, with all other blocks zero; Assumption 1 holds. All blocks other than $(1, y)$ contribute zero to both logits, so $|A_{ij}^{\mathrm{3D}}(Q\hat{p}_i, Q\hat{p}_j) - A_{ij}^{\mathrm{3D}}(\hat{p}_i, \hat{p}_j)| = |\langle q_i^{(1,y)}, M k_j^{(1,y)}\rangle| = 2\sin(\frac{1}{2}) B_q B_k$.

(b) Each axis code has entries $\sin(s\,\omega_t^{\mathrm{ax}})$ and $\cos(s\,\omega_t^{\mathrm{ax}})$ for $t = 1, \ldots, d_{\mathrm{ax}}/2$, so $\langle \mathrm{PE}_c(s), \mathrm{PE}_c(s')\rangle = \sum_t \cos(\omega_t^{\mathrm{ax}}(s - s'))$ and $\kappa(\hat{p}, \hat{p}') = \sum_c \sum_t \cos(\omega_t^{\mathrm{ax}} \Delta c)$. Substituting the two sets of differences,

$$\kappa(Q\hat{p}_i, Q\hat{p}_j) - \kappa(\hat{p}_i, \hat{p}_j) = \sum_t \big[\cos(\sqrt{2}\,\omega_t^{\mathrm{ax}}) + 2 - 2\cos\omega_t^{\mathrm{ax}} - 1\big] = -\sum_t f(\omega_t^{\mathrm{ax}}),$$

with $f(\omega) = 2\cos\omega - \cos(\sqrt{2}\omega) - 1$. The inequalities $\cos x \leq 1 - \frac{x^2}{2} + \frac{x^4}{24}$ and $\cos x \geq 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720}$, valid for all real x, give $f(\omega) \leq -\frac{\omega^4}{12} + \frac{\omega^6}{90} = -\omega^4\big(\frac{1}{12} - \frac{\omega^2}{90}\big)$, which is negative for $0 < \omega^2 < 7.5$. All frequencies lie in $(0, 1]$, so every summand $-f(\omega_t^{\mathrm{ax}})$ is positive, and the $t = 1$ summand, with $\omega_1^{\mathrm{ax}} = 1$, equals $1 + \cos\sqrt{2} - 2\cos 1$. □

Proof of Corollary 5. Let $\psi = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j)$ and $\psi' = d_{\mathbb{S}^2}(\hat{p}_i', \hat{p}_j')$. As in the proof of Theorem 1, the triangle and Cauchy–Schwarz inequalities and $\|R(\alpha) - R(\beta)\|_{\mathrm{op}} \leq |\alpha - \beta|$ give $|A_{ij}^{\mathrm{GeoRoPE}}(\hat{p}_i', \hat{p}_j') - A_{ij}^{\mathrm{GeoRoPE}}(\hat{p}_i, \hat{p}_j)| \leq B_q B_k \sum_t \omega_t |\psi' - \psi| = \Omega B_q B_k |\psi' - \psi|$. Since $d_{\mathbb{S}^2}$ is a metric on $\mathbb{S}^2$, the triangle inequality gives $\psi' \leq d_{\mathbb{S}^2}(\hat{p}_i', \hat{p}_i) + \psi + d_{\mathbb{S}^2}(\hat{p}_j, \hat{p}_j')$ and symmetrically $\psi \leq d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_i') + \psi' + d_{\mathbb{S}^2}(\hat{p}_j', \hat{p}_j)$, so $|\psi' - \psi| \leq \varepsilon$. The GDA bias changes by $\lambda_h |\psi' - \psi| \leq \lambda_h \varepsilon$. □

Proof of Theorem 6. Let σ be the uniform probability measure on $\mathbb{S}^2$ and fix $x, y \in \mathbb{S}^2$ with $d_{\mathbb{S}^2}(x, y) = \theta$. The set of u with $u \cdot x = 0$ or $u \cdot y = 0$ is σ-null. For the remaining u, the great circle orthogonal to u separates x from y, i.e., $\mathrm{sgn}(u \cdot x) \neq \mathrm{sgn}(u \cdot y)$, exactly when u lies in one of two antipodal lunes of dihedral angle θ. Each such lune has area 2θ and the sphere has area 4π, so $\sigma(\{u : \mathrm{sgn}(u \cdot x) \neq \mathrm{sgn}(u \cdot y)\}) = \theta/\pi$. Define $\varphi : \mathbb{S}^2 \to L^2(\sigma)$ by $\varphi(x)(u) = \frac{\sqrt{\pi}}{2}\,\mathrm{sgn}(u \cdot x)$. Since $(\mathrm{sgn}\, a - \mathrm{sgn}\, b)^2 = 4 \cdot \mathbf{1}[\mathrm{sgn}\, a \neq \mathrm{sgn}\, b]$ for non-zero a, b,

$$\|\varphi(x) - \varphi(y)\|_{L^2(\sigma)}^2 = \frac{\pi}{4} \int_{\mathbb{S}^2} \big(\mathrm{sgn}(u \cdot x) - \mathrm{sgn}(u \cdot y)\big)^2\, d\sigma(u) = \pi \cdot \frac{\theta}{\pi} = d_{\mathbb{S}^2}(x, y).$$

Let $z_i = \varphi(\hat{p}_i)$. For $i \neq j$, $\|z_i - z_j\|^2 = d_{\mathbb{S}^2}(\hat{p}_i, \hat{p}_j) > 0$, so $z_1, \ldots, z_n$ are distinct. They lie in the subspace $V = \mathrm{span}\{z_1, \ldots, z_n\}$ of dimension $m \leq n$, and an isometry $V \to \mathbb{R}^m$

maps them to distinct points $w_1, \dots, w_n \in \mathbb{R}^m$ with $\|w_i - w_j\|_2 = \|z_i - z_j\|$. Hence $K_{ij} = \exp(-\lambda_h \|w_i - w_j\|_2^2)$ is the Gaussian kernel matrix of n distinct points in $\mathbb{R}^m$. For $\lambda_h > 0$ the Gaussian kernel is strictly positive definite on $\mathbb{R}^m$ (Wendland, 2005, Chapter 6): for every non-zero $c \in \mathbb{R}^n$, $c^\top K c = \int_{\mathbb{R}^m} \left|\sum_i c_i e^{\mathrm{i}\xi^\top w_i}\right|^2 g(\xi)\, d\xi > 0$, where $g > 0$ is the Gaussian Fourier density, because the exponentials $e^{\mathrm{i}\xi^\top w_i}$ at distinct w_i are linearly independent functions of ξ. Therefore K is strictly positive definite.

□

Proof of Remark 7(i). Order the four points cyclically as $(1,0,0), (0,1,0), (-1,0,0), (0,-1,0)$. Consecutive points are at geodesic distance $\pi/2$ and opposite points at π, so any kernel $\exp(-g(d_{\mathbb{S}^2}))$ yields the circulant matrix with first row $(1, a, b, a)$, $a = e^{-g(\pi/2)}$, $b = e^{-g(\pi)}$, whose eigenvalues are $1 + 2a + b$, $1 - b$ (twice), and $1 - 2a + b$, the last with eigenvector $(1, -1, 1, -1)$. For $g(d) = \lambda d^2$, $a = x$ and $b = x^4$ with $x = e^{-\lambda\pi^2/4}$, and $1 - 2x + x^4 = (x-1)(x^3 + x^2 + x - 1)$ is negative exactly when $x^3 + x^2 + x > 1$, i.e., when $x > x^* \approx 0.5437$, equivalently $\lambda < -4\ln(x^*)/\pi^2 \approx 0.247$. For $g(d) = \lambda d$, $a = y$ and $b = y^2$ with $y = e^{-\lambda\pi/2}$, giving eigenvalues $(1+y)^2$, $1 - y^2$, and $(1-y)^2$, all positive. □