A HOMOMORPHISM AND BUSEMANN APPROACH TO BUILDING HYPERBOLIC TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic space provides compact representations for hierarchical and tree-like data, yet a principled construction of attention that faithfully respects this geometry remains challenging. To remedy this gap, we decompose Euclidean attention into three modules: linear map, similarity, and aggregation. We redesign the linear map and similarity intrinsically in the Lorentz model by leveraging its rich geometric and algebraic structure, which leads to Homomorphism and Busemann transformers (HBformer). We use gyro-homomorphisms for query and value updates, generalizing Euclidean linear layers in a way that respects the Lorentz gyrovector structure. As the Busemann function naturally extends the Euclidean inner product, we derive curvature-aware attention scores via the Lorentz Busemann function. Since Busemann similarity requires keys in the Euclidean tangent space, we introduce a *Lorentz-to-Euclidean fully connected layer* (L2EFC). Experiments on graph benchmarks of different scales, large-scale node property prediction, and semi-supervised image and text classification demonstrate consistent improvements over strong Euclidean and hyperbolic baselines. The code will be open-sourced once accepted.

1 INTRODUCTION

Hyperbolic representations have shown remarkable success on hierarchical and tree-like data (Nickel & Kiela, 2017; Ganea et al., 2018; Chami et al., 2019). As a central neural architecture, the transformer has therefore been adapted to hyperbolic geometry by recent studies (Bose & Das, 2023; Yang et al., 2024; Park et al., 2024; Yang et al., 2025). The self-attention layer within the transformer boils down to three operations: transformation for Q-K-V updates, similarity, and aggregation. The central question is how to realize these primitives intrinsically in hyperbolic space, so that the model preserves geometry, trains stably, and scales to realistic problem sizes.

Existing hyperbolic attention mechanisms still fall short of being fully intrinsic, leaving two core primitives of attention, transformation and similarity, only weakly aligned with hyperbolic geometry. For **transformation**, several Lorentz-based designs (Chen et al., 2022; Yang et al., 2024) perform Euclidean linear maps in the *flat ambient Minkowski space* and then reproject the results onto the Lorentz model. Although such updates preserve the manifold constraint, they primarily operate on extrinsic coordinates rather than leveraging the rich *geometric* and *algebraic (gyrovector) structure* of hyperbolic space. For **similarity**, many variants replace the Euclidean dot-product score with geodesic distance (Gulcehre et al., 2019; Park et al., 2024). While distance reflects proximity, it discards directional (angular) alignment. As shown in Fig. 1, two keys can lie on the same distance shell around a query yet exhibit different angular configurations, resulting in identical distance-based scores despite different dot-product similarities. Other approaches attempt to recover dot-product attention by mapping features to a tangent space via logarithmic/exponential maps (Park et al., 2024), but such local linearization is only locally faithful and can introduce geometric distortion when representations move away from the tangent point.

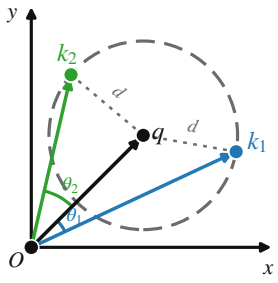


Figure 1: Two keys k_1 and k_2 lie on the same distance shell around query q, so distance-based scores are identical, but dot-product scores are different

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.

On the other hand, hyperbolic geometry admits a vector-space-like calculus through *gyrovector spaces*, where gyroaddition and scalar gyromultiplication generalize Euclidean vector addition and scalar multiplication. This gyrovector formalism provides a powerful tool for building network layers on hyperbolic space (Ganea et al., 2018; Shimizu et al., 2021; Chen et al., 2025b; Wang et al., 2025). Meanwhile, the *Busemann function* offers an intrinsic generalization of Euclidean inner product (Burago et al., 2001). Accordingly, a lot of work has lifted inner-product-based Euclidean methods to hyperbolic geometry, including prototype learning (Atigh et al., 2021), SVM (Fan et al., 2023), PCA (Chami et al., 2021), and Sliced Wasserstein distances (Bonet et al., 2025).

Motivated by these observations, we propose Homomorphism and Busemann Transformers (HBformer), an intrinsic approach to hyperbolic attention in the Lorentz model that respects the vibrant gyrovector algebra and geometric structure. On the **transformation** side, queries and values are updated via gyro homomorphisms, which serve as structure-preserving analogues of Euclidean linear layers by respecting gyroaddition and scalar gyromultiplication. Moreover, these maps act as isometries on the Lorentz model and therefore preserve the metric geometry. On the **similarity** side, we compute curvature-aware attention scores through the Lorentz Busemann function, which intrinsically generalizes Euclidean inner-product similarity. Since Busemann similarity requires key directions in the tangent space at the origin, we introduce a point-to-hyperplane (Shi et al., 2026) *Lorentz-to-Euclidean fully connected layer* (L2EFC). For **aggregation**, we adopt the standard Lorentz midpoint operator to map weighted averages back to the Lorentz model (Nickel & Kiela, 2018; Ungar, 2008). Furthermore, we develop a Busemann-based linear attention variant, termed *Busemann Linear Attention (BLA)*, which preserves linear computational complexity. Experiments on graph benchmarks at multiple scales and on semi-supervised image and text classification demonstrate consistent improvements over strong Euclidean and hyperbolic baselines, achieving state-of-the-art performance (Bose & Das, 2023; Yang et al., 2024; Park et al., 2024; Yang et al., 2025). Our contributions are summarized as follows:

- We develop intrinsic query/value transformations in the Lorentz model via gyro-homomorphisms induced by Lorentz isometries, ensuring geometry and algebra preservation.
- We generalize Euclidean dot-product attention to hyperbolic space by deriving a curvature-aware Lorentz similarity via the Busemann function and develop Busemann Linear Attention.
- We introduce L2EFC, a point-to-hyperplane Lorentz-to-Euclidean FC layer that produces the Euclidean key directions.
- We provide comprehensive empirical validation on small and medium graph benchmarks, large-scale node property prediction, and semi-supervised image and text classification, achieving state-of-the-art results under a unified training protocol.

2 PRELIMINARIES

Lorentz model. The Lorentz model, which is also known as the hyperboloid model, realizes hyperbolic space as the upper sheet of a two-sheeted hyperboloid in $(n+1)$-dimensional Minkowski space. It is given by

$$\mathbb{L}_K^n = \left\{ x \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_{\mathcal{L}} = \tfrac{1}{K}, x_0 > 0 \right\}, \tag{1}$$

where $K < 0$ is the curvature and the Lorentz inner product is $\langle x, y \rangle_{\mathcal{L}} = -x_0 y_0 + \sum_{i=1}^{n} x_i y_i$. Following Ratcliffe (2006), we write $x = (x_0, x_s)$ with *time* coordinate x_0 and *space* coordinates $x_s \in \mathbb{R}^n$. The origin is $\overline{\mathbf{0}} = ((-K)^{-1/2}, 0, \ldots, 0)$. The closed-form geodesic distance is

$$d_{\mathcal{L}}(x, y) = \frac{1}{\sqrt{-K}} \operatorname{arcosh}\left(K \langle x, y \rangle_{\mathcal{L}}\right), \tag{2}$$

Further Riemannian operators, including exponential and logarithmic maps and parallel transport, are summarized in the Appendix.

Busemann functions. Let (X, d) be a proper geodesic metric space, and let $\gamma : [0, \infty) \to X$ be a unit-speed geodesic ray. The Busemann function associated with γ is

$$B^{\gamma}(x) = \lim_{t \to \infty} \left(d\left(x, \gamma(t)\right) - t \right), \tag{3}$$

where $x \in X$.

In Euclidean space, for $\gamma(t) = tv$ with unit direction $v \in \mathbb{S}^{n-1}$ as shown in Fig. 2, we have

$$B^v(x) = -\langle x, v \rangle. \tag{4}$$

Here, $S^{n-1} = \{v \in \mathbb{R}^n \mid \|v\| = 1\}$ denotes the Euclidean unit sphere. Hence, up to sign, the Busemann function recovers the directional inner-product coordinate. As the Busemann function is independent of the starting point of the geodesic (up to an additive constant) (Bridson & Haefliger, 2013, II. Cor. 8.20), we focus on the one starting from the origin with a unit direction v. With $x \in \mathbb{L}_K^n$ and $v \in S^{n-1}$[1], the Busemann function in the Lorentz model (Bonet et al., 2025) is

$$B^v(x) = \frac{1}{\sqrt{-K}} \log\left(-\sqrt{-K}\left\langle x, \sqrt{-K}\overline{\mathbf{0}} + v\right\rangle_{\mathcal{L}}\right), \tag{5}$$

Gyrovector structures. Intuitively, gyrovector spaces play the same role for non-Euclidean spaces as vector spaces for Euclidean geometry, with $\oplus$ and $\otimes$ generalizing vector addition and scalar multiplication (Ungar, 2022; Chen et al., 2025b). A gyrogroup $(\mathcal{G}, \oplus)$ is a non-empty set endowed with a binary operation $\oplus$ with gyroassociative law characterized by gyrations $\mathrm{gyr}[x, y]$. Like groups, each gyrogroup admits a unique neutral element and a unique inverse for each element. Equipping a gyrogroup with a scalar multiplication $\otimes\colon \mathbb{R} \times \mathcal{G} \to \mathcal{G}$ that is compatible with $\oplus$ yields a gyrovector space $(\mathcal{G}, \oplus, \otimes)$.

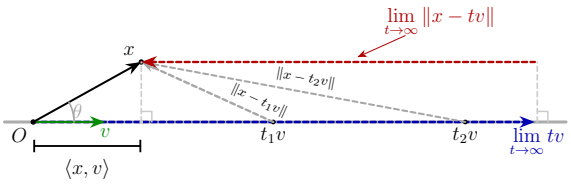


Figure 2: Assuming the red dashed segment represents $\lim_{t\to\infty} tv$, the blue is $\lim_{t\to\infty} \|x - tv\|$. According to the definition of Busemann function, subtracting the two yields the inner product.

Lorentz gyrovector. The Lorentz model $\mathbb{L}_K^n$ admits gyrovector operations $x \oplus_K y$ and $t \otimes_K x$ (Chen et al., 2025b). These operations serve as the hyperbolic analogues of Euclidean addition and scalar multiplication. A map $\Phi\colon \mathbb{L}_K^n \to \mathbb{L}_K^n$ is called a gyrovector homomorphism (Wang et al., 2025) if it preserves the gyrovector structure:

$$\Phi(x \oplus_K y) = \Phi(x) \oplus_K \Phi(y), \quad \Phi(t \otimes_K x) = t \otimes_K \Phi(x). \tag{6}$$

for all $x, y \in \mathbb{L}_K^n$ and $t \in \mathbb{R}$. Complete definitions of gyrogroups to gyrovector spaces, and explicit formulas for $x \oplus_K y$ and $t \otimes_K x$ in the Lorentz model are provided in the Appendix.

3 Method

In this section, we present HBformer. We first construct *Query and value* (Q and V) updates via homomorphism-constrained mappings on the Lorentz model, derived from the positive special Lorentz group and restricted to the isotropy subgroup at the origin, so the transformations are isometric, manifold-preserving, and rotation-equivariant. We then reformulate Euclidean dot-product attention through the Busemann function to obtain a similarity computed intrinsically on the Lorentz model. Finally, because Busemann scoring requires *key* (K) directions in the origin tangent space, we introduce a point-to-hyperplane transformation that both provides this mapping and converts Lorentz points to tangent-space keys via closed-form signed distances. This intrinsic pipeline avoids chart transitions and aligns the algebraic modules with hyperboloid geometry.

3.1 Lorentz homomorphism transformation

In Euclidean attention, Q and V updates are linear homomorphisms that preserve addition and scalar multiplication. We generalize them to the Lorentz model by requiring the Q and V update to be a gyrovector space homomorphism that preserves Lorentz gyroaddition and scalar gyromultiplication $(\oplus_K, \otimes_K)$, enabling an intrinsic attention pipeline. In contrast, prior Lorentz Transformers (e.g., FNN (Chen et al., 2022), Hypformer (Yang et al., 2024)) apply linear layers in the ambient

[1]The tangent space at $\overline{\mathbf{0}}$ is $T_{\overline{\mathbf{0}}}\mathbb{L}_K^n = \{(0, v) \mid v \in \mathbb{R}^n\} \cong \mathbb{R}^n$. Therefore, the unit tangent vectors lie in S^{n-1}.

Minkowski space with subsequent re-projection, which are not intrinsic to the Lorentz model and fail to respect the geometry.

We work on the Lorentz hyperboloid for any negative curvature $K < 0$, $\mathbb{L}_K^n$, equipped with the Lorentz inner product $\langle u, v\rangle_{\mathcal{L}} = -u_0 v_0 + \langle \mathbf{u}, \mathbf{v}\rangle$. Let $O(1,n) = \{A \in \mathbb{R}^{(n+1)\times(n+1)} : A^\top J A = J\}$ be the Lorentz group with $J = \mathrm{diag}(-1, I_n)$, and let $SO^+(1,n) = \{A \in O(1,n) : \det(A) = 1, A_{00} > 0\}$ be the positive special Lorentz group.

Classical facts. The action $A : x \mapsto Ax$ with $A \in SO^+(1,n)$ preserves the Lorentz product (Fan et al., 2022), hence defines an isometry of $(\mathbb{L}_{-1}^n, g_{-1})$. Moreover, the action is transitive and the stabilizer at the canonical origin $\overline{\mathbf{0}} = (1, 0, \ldots, 0)^\top$ is isomorphic to $SO(n)$:

$$\mathrm{Iso}_{\overline{\mathbf{0}}}(\mathbb{L}_{-1}^n) = \{\mathrm{diag}(1, R) : R \in SO(n)\} \cong SO(n). \tag{7}$$

These statements are standard in Lorentzian geometry, and the nested hyperbolic spaces literature (Fan et al., 2022) provides a detailed treatment in the case $K = -1$. We can next extend the transitivity statement to arbitrary negative curvature $K < 0$ (Gel'fand et al., 2018) as follows:

Proposition 1 (Transitivity for general $K < 0$)**.** *For every $x, y \in \mathbb{L}_K^n$ with $K < 0$, there exists $A \in SO^+(1,n)$ such that $Ax = y$.*

The proof follows the standard construction of Lorentz boosts and coincides with the $K = -1$ case in Fan et al. (2022) up to a curvature-dependent rescaling. For completeness, we provide the explicit argument in the Appendix.

Homogeneous space view and design principle. By Proposition 1 and the isotropy description,

$$(\mathbb{L}_K^n, g_K) \cong SO^+(1,n)/SO(n), \tag{8}$$

every Lorentz point can be represented by a coset in $SO^+(1,n)/SO(n)$. This observation informs our design, but we do not apply arbitrary group actions as network updates. Instead, we parameterize query and value updates as *gyrogroup homomorphisms* on $(\mathbb{L}_K^n, \oplus_K, \otimes_K)$ that preserve the Lorentz gyrostructure Eq. (6). Concretely, we restrict to maps induced by the isotropy subgroup at the origin. For $R \in SO(n)$ let $Q = \mathrm{diag}(1, R)$ and define $\Phi_R(x) = Qx$. Then Φ_R is an isometry on $\mathbb{L}_K^n$ and satisfies Eq. (6), which follows from the rotation equivariance of the stereographic interface and its gyro–isomorphism with the Lorentz model. Thus Φ_R acts as a homomorphism-constrained analogue of a linear layer on the manifold, guaranteeing manifold preservation and rotation equivariance. In the flat limit $K \to 0^-$, Φ_R reduces to a bias-free *orthogonal* linear map with weight R, i.e., a restricted special case of the unconstrained Euclidean projection matrices used in standard attention.

Updates restricted to the isotropy subgroup. Motivated by the homogeneous space view, we realize query and value updates by maps taken from the isotropy subgroup at $\overline{\mathbf{0}}$. For $R \in SO(n)$ set $Q = \mathrm{diag}(1, R)$ and define

$$\Phi_R : \mathbb{L}_K^n \to \mathbb{L}_K^n, \qquad \Phi_R(x) = Qx. \tag{9}$$

The following result summarizes the properties that are relevant to our attention design.

Theorem 1 (Isotropy homomorphisms)**.** *Let $K < 0$ and let $\oplus_K$ and $\otimes_K$ denote the Lorentz gyroaddition and scalar multiplication on $\mathbb{L}_K^n$. For every $R \in SO(n)$ the map Φ_R in Eq.* (9) *satisfies:*

1. *Φ_R is an isometry of $(\mathbb{L}_K^n, g_K)$ which preserves the hyperboloid;*
2. *Φ_R is a gyrogroup homomorphism, that is,*

$$\Phi_R(x \oplus_K y) = \Phi_R(x) \oplus_K \Phi_R(y), \forall x, y \in \mathbb{L}_K^n; \tag{10}$$

3. *Φ_R is compatible with scalar multiplication,*

$$\Phi_R(t \otimes_K x) = t \otimes_K \Phi_R(x), \forall t \in \mathbb{R}, x \in \mathbb{L}_K^n. \tag{11}$$

The proof follows from the fact that elements of $SO^+(1,n)$ act as isometries on $\mathbb{L}_K^n$, and that the gyrovector operations $\oplus_K$ and $\otimes_K$ are constructed to be equivariant with respect to the action of the isotropy subgroup at the origin. A complete derivation is provided in the Appendix. Eq. (10) and Eq. (11) justify interpreting Φ_R as a homomorphism-constrained analogue of a linear layer on the Lorentz model, which we use as the basic transformation for queries and values in HBformer.

Consequences for attention. The family $\{\Phi_R : R \in SO(n)\}$ provides intrinsic updates for queries and values that are geometry-preserving and bias-free. They act entirely on $\mathbb{L}_K^n$ for any $K < 0$, maintain rotation equivariance through the isotropy subgroup, and avoid chart transitions. The above homomorphism implements the transformation from the input x to the query and value representations, which is intrinsic to the Lorentz model.

3.2 BUSEMANN ATTENTION

We design an intrinsic similarity for attention by viewing the Euclidean dot-product similarity through the lens of the Busemann potential, and then extending it to the Lorentz model. In Euclidean space, this can be written as (Bonet et al., 2025), with the proof in Appendix:

$$\mathrm{sim}(x, y) = \langle x, y \rangle = -\|y\| B_{y/\|y\|}(x), \tag{12}$$

which motivates replacing the inner product by the Busemann score on the Lorentz model. Let $K < 0$ and let $\mathbb{L}_K^d$ be the Lorentz model. For a unit direction $v \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^d \cap S^{d-1}$ and any $x \in \mathbb{L}_K^d$, the Lorentz Busemann function (Bonet et al., 2025) is defined in Eq. (5). The argument of the logarithm is positive on the forward sheet, which makes Eq. (5) well-defined.

Similarity in the Lorentz model. Let $q \in \mathbb{L}_K^d$ be a query and let $m : \mathbb{L}_K^d \to T_{\overline{\mathbf{0}}}\mathbb{L}_K^d \setminus \{0\}$ denote the tangent–key mapping, realized either by the logarithmic map at the origin or by the point–to–hyperplane construction. The Busemann similarity score of query q and key k is given in closed form by

$$s(q, k) = -\frac{\|k\|}{\sqrt{-K}} \log\left(-\sqrt{-K}\left\langle q, \sqrt{-K}\overline{\mathbf{0}} + \frac{k}{\|k\|}\right\rangle_{\mathcal{L}}\right), \tag{13}$$

where $\overline{\mathbf{0}} = ((-K)^{-1/2}, 0, \ldots, 0)^\top$. Eq. (13) is computed entirely on the manifold for any $K < 0$. In the flat–space limit $K \to 0^-$, the Busemann function linearizes and $s(q, k)$ reduces to a scaled dot–product, recovering the Euclidean attention score, which can be written as (Chen et al., 2026):

$$B^v(x) \xrightarrow[K\to 0^-]{} -\langle v, x_s \rangle. \tag{14}$$

Attention weights and aggregation. Given keys $\{k_j\}_{j=1}^N$ in $T_{\overline{\mathbf{0}}}\mathbb{L}_K^d$ and values $\{\mathbf{V}_j\}_{j=1}^N$ on $\mathbb{L}_K^d$, the attention weights for a query q_i are

$$\alpha_{ij} = \mathrm{softmax}_j\left(\frac{s(q_i, k_j)}{\sigma} + b\right), \qquad \sum\nolimits_{j=1}^N \alpha_{ij} = 1, \tag{15}$$

where $\sigma > 0$ is a scale and b is a learnable bias within the head.

Aggregation uses the closed-form Lorentzian midpoint (Law et al., 2019), which is given by

$$\mathbf{Att}_i \odot^K \mathbf{V}_j := \frac{\sum_{j=1}^N \alpha_{ij}\mathbf{V}_j}{\sqrt{|K|\left|\|\sum_{k=1}^N \alpha_{ik}\mathbf{V}_k\|_{\mathcal{L}}^2\right|}}. \tag{16}$$

In the flat limit $K \to 0^-$, if each value is viewed as the curvature-dependent lift of a fixed Euclidean vector, then the space-like component of Eq. (16) converges to the Euclidean weighted average $\sum_{j=1}^N \alpha_{ij} v_j$. A proof is given in the Appendix. Here, the absolute value in the denominator ensures a positive normalization factor. Eq. (16) realizes the Lorentzian midpoint explicitly.

Practical computation. Implementation follows directly from Eq. (13). For each head and each key k_j, compute its spatial Euclidean norm $\|k_j\|$ and the associated unit direction $u_j = k_j/\|k_j\|$. Calculate the pairwise Lorentzian inner products $\langle q_i, \overline{\mathbf{0}} + u_j \rangle_{\mathcal{L}}$ for all queries q_i, and the Busemann scores are then obtained from Eq. (13) by applying the logarithm with a small positive lower bound on its argument for numerical stability, followed by division by $\sqrt{-K}$ and multiplication by $\|k_j\|$. In practice, we replace the final multiplication $\|k_j\|$ by a learnable scalar τ, which is equivalent to a learnable temperature in classical attention, and observe improved performance. A standard softmax as in Eq. (15) produces the attention weights, and the Lorentz midpoint in Eq. (16) yields the outputs on $\mathbb{L}_K^d$. The entire computation is intrinsic for every $K < 0$ and does not require any extrinsic calculations.

Table 1: Comparison of attention components in Euclidean space and on the Lorentz hyperboloid. α is calculated from the similarity by softmax, and we set $K = -1$.

Geometry	$\mathbb{R}^n$	$\mathbb{L}_K^n$	$\mathbb{L}_K^n$
Transformation	$\mathrm{Linear}(x_i)$ Linear homomorphism	$\pi_{\mathbb{R}\to\mathbb{L}}(\mathrm{Linear}(x_i))$ Ambient Minkowski	$\Phi_R(x_i)$ gyro-homomorphism
Similarity	$\langle q_i, k_j\rangle = -\|k_j\| B_{k_j/\|k_j\|}(q_i)$ Dot product	$\operatorname{arcosh}(-\langle q_i, k_j\rangle_{\mathcal{L}})$ Geodesic distance	$-\tau B_{k_j/\|k_j\|}(q_i)$ Busemann similarity
Aggregation	$\sum_j \alpha_{ij} v_j$	$\frac{\sum_j \alpha_{ij} v_j}{\sqrt{\lvert\lVert\sum_k \alpha_{ik} v_k\rVert_{\mathcal{L}}^2\rvert}}$	$\frac{\sum_j \alpha_{ij} v_j}{\sqrt{\lvert\lVert\sum_k \alpha_{ik} v_k\rVert_{\mathcal{L}}^2\rvert}}$
References	Vaswani et al. (2017)	Gulcehre et al. (2019)	Ours

3.3 Point-to-hyperplane transformation for keys

The Busemann similarity in Section 3.2 requires the homomorphism-constrained updates produce query and value points on the Lorentz model, while keys are represented as directions that lie in the tangent space at the origin $T_{\mathbf{0}}\mathbb{L}_K^m$. Mapping every key to $T_{\mathbf{0}}\mathbb{L}_K^m$ by the logarithmic map at the origin introduces chart transitions, which can introduce geometric distortion when representations move away from the tangent point. We therefore construct a direct, intrinsic transformation from Lorentz points to the origin tangent space using signed distances to learned Lorentz hyperplanes.

The Euclidean FC layer is defined as $\boldsymbol{y} = W\boldsymbol{x} + b$ with $W \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. It can be expressed element-wise as $y_k = \langle w_k, \boldsymbol{x}\rangle - b_k$ with $w_k \in \mathbb{R}^n$ and $b_k \in \mathbb{R}$. Shimizu et al. (2021, Sec. 3.2) and Chen et al. (2025c, Sec. 3.1) interpreted the Euclidean FC layer as an operation that transforms the input $\boldsymbol{x}$ via $\langle w_k, \boldsymbol{x}\rangle + b_k$, treating the output y_k as the signed distance from the hyperplane, $H_{e^{(k)},\mathbf{0}} = \{\, \boldsymbol{y} \in \mathbb{R}^m \mid \langle e^{(k)}, \boldsymbol{y} - \mathbf{0}\rangle = 0 \,\}$, passing through the origin and orthogonal to the k-th axis of the output space $\mathbb{R}^m$, which can be written as:

$$\operatorname{sign}\big(\langle e^{(k)}, \boldsymbol{y} - \mathbf{0}\rangle\big)\, d\big(\boldsymbol{y}, H_{e^{(k)},\mathbf{0}}\big) = v_k(\boldsymbol{x}), \qquad k = 1, \ldots, m, \tag{17}$$

where $\mathbf{0} \in \mathbb{R}^m$ is zero vector and $v_k(\boldsymbol{x}) = \langle w_k, \boldsymbol{x}\rangle - b_k$. Here, $e^{(k)}$ denotes the vector whose k-th entry is 1 and all other entries are 0. The left-hand side of (17) equals y_k.

While keeping the left-hand side and replacing the right-hand side with the Lorentz counterpart, we can extend the above formulation into the FC layer (L2EFC) that maps an input $\boldsymbol{x} \in \mathbb{L}_K^n$ to an output $\boldsymbol{y} \in \mathbb{R}^m \cong T_{\mathbf{0}}\mathbb{L}_K^m$. Let $\{(\boldsymbol{z}_k, a_k)\}_{k=1}^m$ be learnable FC parameters and define $v_k(\boldsymbol{x}) = v_{\boldsymbol{z}_k, a_k}(\boldsymbol{x})$ as in Bdeir et al. (2024, Eq. 12), which is reviewed in the Appendix. Once matching each Lorentz version of $v_k(\boldsymbol{x})$ to the Euclidean signed distance from $\boldsymbol{y}$ to the k-th coordinate hyperplane, we have the following theorem.

Theorem 2 (L2EFC)**.** *Let $\boldsymbol{x} \in \mathbb{L}_K^n$, $Z = \{\boldsymbol{z}_k\}_{k=1}^m \subset \mathbb{R}^n$ and $a = \{a_k\}_{k=1}^m \subset \mathbb{R}$. The Lorentz-to-Euclidean fully connected layer* $\mathrm{L2EFC} : \mathbb{L}_K^n \to T_{\mathbf{0}}\mathbb{L}_K^m$ *is*

$$\boldsymbol{y} = \Big[0,\ \big(v_{\boldsymbol{z}_k, a_k}(\boldsymbol{x})\big)_{k=1}^m\Big] = \Big[0,\ \big(v_{\boldsymbol{z}_1, a_1}(\boldsymbol{x}), \ldots, v_{\boldsymbol{z}_m, a_m}(\boldsymbol{x})\big)^\top\Big]. \tag{18}$$

where $v_{\boldsymbol{z}_k, a_k}(\boldsymbol{x})$ is defined in the Appendix and $\{(\boldsymbol{z}_k \in \mathbb{R}^n, a_k \in \mathbb{R})\}_{k=1}^m$ are FC parameters.

3.4 Busemann linear attention

For a linear-attention variant with linear computational complexity, let $Q_s, K_s, V_s \in \mathbb{R}^{N\times d}$ collect the spatial components of the Lorentz query, key, and value points produced by gyro-homomorphic transformations. Hypformer (Yang et al., 2024) computes $Z_s = Q_s\left(K_s^\top V_s\right) / Q_s\left(K_s^\top \mathbf{1}_N\right)$. We use the Busemann function to replace the ambient inner product of Hypformer and the resulting *Busemann Linear Attention (BLA)* is

$$\left[Z_s^B\right]_{ir} = \frac{s\left(q_i, \left[K_s^\top V_s\right]_{:r}\right)}{s\left(q_i, K_s^\top \mathbf{1}_N\right)}. \tag{19}$$

Here, $i = 1, \ldots, N$ indexes queries and $r = 1, \ldots, d$ indexes output features. The overall complexity is $O(N)$ and can be scalable to large graph datasets.

Table 2: Testing results on large-scale node property prediction benchmarks and small and medium-sized graph benchmarks. ROC-AUC is used for ogbn-proteins, and Accuracy for other datasets. OOM denotes out-of-memory during training or testing. The best and second-best results are highlighted in red bold and underlined, respectively.

	Small-Medium				Large		
Method	AIRPORT	CORA	CITESEER	PUBMED	ogbn-proteins	Amazon2M	ogbn-arxiv
#Nodes	2,665	2,708	3,327	19,717	132,534	2,449,029	169,343
#Edges	2,664	5,429	4,732	88,651	39,561,252	61,859,140	1,166,243
MLP	–	–	–	–	72.0 ± 0.5	63.5 ± 0.1	55.5 ± 0.2
GCN (Kipf & Welling, 2017)	81.4 ± 0.6	81.3 ± 0.3	71.6 ± 0.4	78.1 ± 0.2	72.5 ± 0.4	83.9 ± 0.1	71.7 ± 0.3
GAT (Velickovic et al., 2018)	81.5 ± 0.3	83.0 ± 0.7	72.5 ± 1.1	79.0 ± 0.3	–	–	–
SGC (Wu et al., 2019)	82.1 ± 0.5	80.1 ± 0.2	71.9 ± 0.1	78.7 ± 0.1	70.3 ± 0.2	81.2 ± 0.1	67.8 ± 0.3
GCN-NSampler	–	–	–	–	73.5 ± 1.3	83.8 ± 0.4	68.5 ± 0.2
GAT-NSampler	–	–	–	–	74.6 ± 1.2	85.2 ± 0.3	67.6 ± 0.2
SIGN (Frasca et al., 2020)	–	–	–	–	71.2 ± 0.5	81.0 ± 0.3	70.3 ± 0.3
Graphormer (Ying et al., 2021)	88.1 ± 1.2	60.0 ± 0.5	61.4 ± 0.6	73.3 ± 0.7	OOM	OOM	OOM
GraphTrans (Wu et al., 2021)	94.3 ± 0.6	77.6 ± 0.8	65.1 ± 1.4	77.5 ± 0.7	OOM	OOM	OOM
GraphGPS (Rampášek et al., 2022)	94.5 ± 0.9	73.0 ± 1.4	62.0 ± 1.5	72.8 ± 1.4	OOM	OOM	OOM
HGNN (Liu et al., 2019)	84.7 ± 1.0	77.1 ± 0.8	70.0 ± 1.0	78.3 ± 1.2	–	–	–
HGCN (Chami et al., 2019)	89.3 ± 1.2	76.5 ± 0.6	68.0 ± 0.6	78.0 ± 1.0	–	–	–
HGAT (Chami et al., 2019)	89.6 ± 1.0	77.4 ± 0.7	68.6 ± 0.3	78.3 ± 1.4	–	–	–
FPS-T (Cho et al., 2023)	**96.0 ± 0.6**	82.3 ± 0.7	70.0 ± 0.7	78.5 ± 0.6	–	–	–
HAN (Gulcehre et al., 2019)	92.9 ± 0.6	83.1 ± 0.5	72.4 ± 0.5	79.0 ± 0.6	OOM	OOM	OOM
HNN++ (Shimizu et al., 2021)	92.3 ± 0.3	82.8 ± 0.6	71.5 ± 1.3	79.9 ± 0.4	OOM	OOM	OOM
F-HNN (Chen et al., 2022)	93.0 ± 0.7	81.0 ± 0.7	71.2 ± 0.4	77.5 ± 0.8	OOM	OOM	OOM
NodeFormer (Wu et al., 2022)	80.2 ± 0.6	82.2 ± 0.9	72.5 ± 1.1	79.9 ± 1.0	77.5 ± 1.2	87.9 ± 0.2	59.9 ± 0.4
SGFormer (Wu et al., 2023)	92.9 ± 0.5	83.2 ± 0.9	72.2 ± 0.3	80.0 ± 0.8	79.5 ± 0.3	89.1 ± 0.1	72.4 ± 0.3
Hypformer (Yang et al., 2024)	95.0 ± 0.5	**85.0** ± 0.3	**73.3** ± 0.4	**81.3** ± 0.3	**80.4 ± 0.5**	**89.4 ± 0.3**	**73.2 ± 0.2**
HBformer	**96.72 ± 0.3**	**86.00 ± 0.3**	**73.72 ± 0.3**	**83.30 ± 0.23**	**81.25 ± 0.1**	**91.55 ± 0.10**	**74.10 ± 0.1**

3.5 GENERALIZATION

Table 1 highlights that each component of HBformer admits a well-defined Euclidean limit, in contrast to prior Lorentz-based designs. In Euclidean space, $\mathrm{Linear}(\cdot) : \mathbb{R}^n \to \mathbb{R}^m$ is a vector-space homomorphism, and gyrovector homomorphisms are its natural generalization to gyrovector spaces (Wang et al., 2025). The Lorentz Busemann similarity $-\|k\| B_{k/\|k\|}(q)$ converges to the standard inner product between q and k, which recovers the usual scaled dot-product score. The Lorentz midpoint aggregation reduces to the arithmetic mean, yielding the standard weighted sum of values. Therefore, when the Lorentzian space degenerates into Euclidean space, the HBformer block becomes almost identical to the Euclidean transformer attention block in the flat-space limit, while existing hyperbolic transformers based on ambient Minkowski fully connected layers and geodesic distance similarities do not recover the exact Euclidean attention formulation when the curvature parameter approaches zero.

4 EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed method on graph benchmarks and classical image and natural language classification settings. Graphs often exhibit complex topological and hierarchical relations, which makes them an ideal testbed for evaluating the effectiveness of our approach. We therefore begin with a comprehensive study on small, medium, and large graph benchmarks. Here we emphasize that our method is a general-purpose attention module for Transformer architectures, rather than a specialized solution for graph datasets. We then evaluate classical image and natural language classification settings where input graphs are constructed by a k-nearest-neighbor procedure, in order to assess transfer beyond purely graph-structured domains. To ensure a fair comparison, our architectural and training choices follow Hypformer. The only differences lie in the attention core: we replace the attention block with ours on the Lorentz model.

4.1 MAIN RESULTS

Small, medium, and large graphs. We evaluate HBformer on seven benchmarks spanning small and medium graphs (AIRPORT, CORA, CITESEER, and PUBMED) and large-scale node property prediction (ogbn-proteins, Amazon2M, and ogbn-arxiv), using accuracy for all datasets except ROC-AUC for ogbn-proteins. As shown in Table 2, HBformer achieves the best mean performance on

Table 3: Results on semi-supervised classification on Mini-ImageNet and 20News-Groups, where we use k-NN (with different values of k) to construct input graphs. The best and second-best results are highlighted in red bold and underlined, respectively.

Method	Mini-ImageNet				20News-Groups			
	$k=5$	$k=10$	$k=15$	$k=20$	$k=5$	$k=10$	$k=15$	$k=20$
GCN (Kipf & Welling, 2017)	84.86 ± 0.42	85.61 ± 0.40	85.93 ± 0.59	85.96 ± 0.66	65.98 ± 0.68	64.13 ± 0.88	62.95 ± 0.70	62.59 ± 0.62
GAT (Velickovic et al., 2018)	84.70 ± 0.48	85.24 ± 0.42	85.41 ± 0.43	85.37 ± 0.51	64.06 ± 0.44	62.51 ± 0.71	61.38 ± 0.88	60.80 ± 0.59
DropEdge (Rong et al., 2020)	83.91 ± 0.24	85.35 ± 0.44	85.25 ± 0.63	85.81 ± 0.65	64.46 ± 0.43	64.01 ± 0.42	62.46 ± 0.51	62.68 ± 0.71
IDGL (Chen et al., 2020)	83.63 ± 0.32	84.41 ± 0.35	85.50 ± 0.24	85.66 ± 0.42	65.09 ± 1.23	63.41 ± 1.26	61.57 ± 0.52	62.21 ± 0.79
LDS (Franceschi et al., 2019)	OOM	OOM	OOM	OOM	66.15 ± 0.36	64.70 ± 1.07	63.51 ± 0.64	63.51 ± 1.75
NodeFormer (Wu et al., 2022)	86.77 ± 0.45	86.74 ± 0.23	86.87 ± 0.41	86.64 ± 0.42	66.01 ± 1.18	65.21 ± 1.14	64.69 ± 1.31	64.55 ± 0.97
NodeFormer w/o graph (Wu et al., 2023)	87.46 ± 0.36				64.71 ± 1.33			
SGFormer (Wu et al., 2023)	86.21 ± 0.66	86.46 ± 0.61	86.73 ± 0.84	86.76 ± 0.72	$\underline{68.55 \pm 0.54}$	67.96 ± 0.68	66.44 ± 0.87	65.46 ± 0.59
SGFormer w/o graph (Wu et al., 2023)	87.25 ± 0.38				67.53 ± 0.43			
Hypformer (Yang et al., 2024)	$\underline{87.36 \pm 0.73}$	$\underline{87.30 \pm 0.65}$	$\underline{87.41 \pm 0.59}$	$\underline{87.48 \pm 0.61}$	68.21 ± 0.78	$\underline{68.01 \pm 0.34}$	$\underline{66.87 \pm 0.30}$	$\underline{66.74 \pm 0.19}$
Hypformer w/o graph (Yang et al., 2024)	$\underline{87.73 \pm 0.63}$				$\underline{67.73 \pm 0.23}$			
HBformer	**88.04 ± 0.22**	**88.06 ± 0.19**	**88.12 ± 0.27**	**88.11 ± 0.14**	**70.11 ± 0.53**	**69.90 ± 0.57**	**69.53 ± 0.50**	**69.32 ± 0.54**
HBformer w/o graph	**88.39 ± 0.17**				**69.20 ± 0.32**			

Table 5: Curvature sensitivity. Span denotes the maximum-minus-minimum accuracy gap.

Dataset / k	0.5	1	2	3	4	20	Span
Cora	$85.32_{\pm 0.49}$	$85.50_{\pm 0.55}$	$85.38_{\pm 0.47}$	**$86.00_{\pm 0.30}$**	$85.62_{\pm 0.60}$	$85.42_{\pm 0.72}$	0.68
PubMed	$82.90_{\pm 1.18}$	$82.80_{\pm 1.29}$	**$83.30_{\pm 0.23}$**	$82.60_{\pm 0.53}$	$82.72_{\pm 0.48}$	$82.62_{\pm 0.82}$	0.70
Mini-ImageNet	$87.94_{\pm 0.45}$	$87.99_{\pm 0.44}$	$87.82_{\pm 0.45}$	$88.01_{\pm 0.38}$	$88.05_{\pm 0.40}$	**$88.11_{\pm 0.14}$**	0.29
20News-Groups	$68.76_{\pm 0.55}$	$68.82_{\pm 0.51}$	$68.84_{\pm 0.78}$	$68.54_{\pm 0.45}$	$69.09_{\pm 0.55}$	**$69.32_{\pm 0.54}$**	0.78

every benchmark, improving the strongest prior result on AIRPORT from 96.0 ± 0.6 to 96.72 ± 0.3. It also surpasses Hypformer on all three citation networks and all three large-scale benchmarks, including an improvement from 89.4 ± 0.3 to 91.55 ± 0.10 on Amazon2M, while maintaining standard deviations that are smaller than or comparable to those of the baselines. Dataset protocols, configurations, and the BLA implementation used for large graphs are detailed in the Appendix.

Text and vision results. We further evaluate HBformer on semi-supervised classification using Mini-ImageNet and 20News-Groups, where we follow graph Transformer protocols (Wu et al., 2022; 2023; Yang et al., 2024) to construct k-nearest-neighbor graphs for $k \in \{5, 10, 15, 20\}$ and also report graph-free variants. As shown in Table 3, HBformer outperforms all baselines for every value of k and without an explicit graph. Compared with the strongest prior results, it improves Mini-ImageNet accuracy by approximately 0.6 to 0.8 percentage points across the graph-based settings and improves 20News-Groups accuracy by approximately 1.5 to 2.7 percentage points, with similarly consistent gains in the graph-free setting. Detailed feature construction, data splits, baseline configurations, and evaluation protocols are provided in the Appendix.

Conventional vision classification. We further assess whether our attention module transfers to a standard Vision Transformer by replacing self-attention in every LViT block with our hyperbolic attention, following He et al. (2025), while keeping the remaining architecture and training pipeline unchanged. As shown in Table 4, compared with the LViT baseline, our attention improves test accuracy from 85.02 to 86.18 on CIFAR-10 and from 69.11 to 70.36 on CIFAR-100.

Table 4: Test accuracy (ACC, %) on CIFAR-10 and CIFAR-100.

Method	CIFAR-10	CIFAR-100
HVT	61.44	42.77
LViT	85.02	69.11
LViT + ours	**86.18**	**70.36**

4.2 ABLATION STUDY

Curvature sensitivity. We evaluate sensitivity to curvature by sweeping the Lorentz parameter $k \in \{0.5, 1, 2, 3, 4, 20\}$, corresponding to curvature $-1/k$, on Cora, PubMed, Mini-ImageNet, and 20News-Groups. As shown in Table 5, the maximum-minus-minimum accuracy spans across the six curvature choices are only 0.68, 0.70, 0.29, and 0.78 points on the four datasets, respectively. These small variations across a wide range of curvature values indicate that HBformer is not sensitive to the curvature choice and does not rely on a narrowly tuned value of k.

Similarity ablation. To isolate the contribution of the similarity function, we replace only Busemann attention with the linear attention used by Hypformer (Yang et al., 2024), full atten-

Table 6: Ablation on similarity.

Dataset	Busemann (ours)	Linear attention	Full attention	Geodesic attention
Cora	**86.00**$_{\pm 0.30}$	85.52$_{\pm 0.71}$	85.52$_{\pm 0.96}$	85.22$_{\pm 0.65}$
PubMed	**83.30**$_{\pm 0.23}$	82.26$_{\pm 0.33}$	81.94$_{\pm 0.42}$	82.04$_{\pm 0.66}$
Mini	**88.11**$_{\pm 0.14}$	87.43$_{\pm 0.44}$	87.25$_{\pm 0.70}$	87.61$_{\pm 0.49}$
20News	**69.32**$_{\pm 0.54}$	68.29$_{\pm 0.71}$	68.29$_{\pm 0.37}$	68.41$_{\pm 0.48}$

Table 7: Ablation on transformation.

Dataset	homomorphism (ours)	HTC	KAK Lorentz	Lorentz FC
Cora	**86.00**$_{\pm 0.30}$	85.04$_{\pm 0.36}$	85.36$_{\pm 0.43}$	85.22$_{\pm 0.26}$
PubMed	**83.30**$_{\pm 0.23}$	81.86$_{\pm 0.46}$	82.18$_{\pm 0.73}$	81.88$_{\pm 0.54}$
Mini	**88.11**$_{\pm 0.14}$	87.54$_{\pm 0.54}$	87.84$_{\pm 0.64}$	87.58$_{\pm 0.57}$
20News	**69.32**$_{\pm 0.54}$	68.35$_{\pm 0.86}$	69.14$_{\pm 0.66}$	68.91$_{\pm 0.52}$

tion (Vaswani et al., 2017), or geodesic attention (Gulcehre et al., 2019). As shown in Table 6, Busemann attention achieves the highest accuracy on all datasets. The consistent gains across graph, image, and text benchmarks show that the intrinsic Busemann similarity provides a more effective measure of query-key compatibility in hyperbolic space than the three alternative similarity choices.

Transformation ablation. To examine the role of the homomorphism-constrained transformation, we replace only the proposed homomorphism query/value mapping with HTC (Yang et al., 2024), KAK Lorentz (Bonet et al., 2025), or Lorentz FC (Chen et al., 2022). Here, KAK Lorentz represents a general Lorentz transformation that combines rotations and boosts, while HTC and Lorentz FC cover representative transformations used by prior hyperbolic networks. As shown in Table 7, the proposed mapping performs best on every dataset. These results support that explicitly preserving Lorentz gyroaddition and scalar gyromultiplication in the query/value updates yields more effective representations than unconstrained or coordinate-based Lorentz transformations. Furthermore, KAK Lorentz requires about 2 times the runtime of our homomorphism transformation.

Computational efficiency. We compare BLA with Hypformer in runtime and attention-core training memory. As shown in Table 8, BLA achieves slightly lower training and inference times on both ogbn-arxiv and Amazon2M. Like Hypformer, BLA avoids an $N \times N$ attention matrix and has $O(Nd^2)$ computational complexity, linear in N for fixed d. Figure 3 shows approximately linear memory growth for both methods, whereas HBformer with classical attention exhibits much steeper growth and encounters OOM errors. At 200,000 nodes, BLA uses 1.15 GiB compared with Hypformer's 1.91 GiB, reducing memory by approximately 40%.

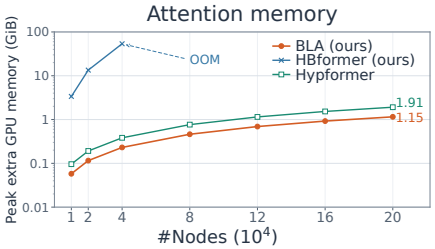


Figure 3: Attention-core training memory. OOM denotes out of memory.

Table 8: Runtime on large graphs. Bold indicates the lower time for each dataset.

Dataset	Method	Training (s/epoch)	Inference (s)
ogbn-arxiv	BLA	**0.519**	**0.193**
	Hypformer	0.539	0.199
Amazon2M	BLA	**6.090**	**2.645**
	Hypformer	6.739	2.718

5 Conclusion

In this paper, we proposed a homomorphism and Busemann approach to building hyperbolic Transformers and instantiated it as HBformer. By aligning Euclidean transformation, similarity, and aggregation with the homogeneous structure of the Lorentz model, we designed homomorphism-constrained Lorentz transformations for queries and values that act as intrinsic isometries, a curvature-aware Busemann similarity that generalizes Euclidean dot product attention to the Lorentz model, and a point-to-hyperplane transformation that generates tangent space keys through closed-form signed distances while avoiding chart transitions. Together with Lorentzian midpoint aggregation, these components yield an attention pipeline that remains entirely on the Lorentz model and preserves its geometry. Experiments on graph benchmarks of different scales and on semi-supervised image and text classification show consistent improvements over strong Euclidean and hyperbolic baselines and indicate enhanced robustness through reduced performance variance. As future work, we plan to further investigate more efficient and effective attention aggregation mechanisms in hyperbolic space and related non-Euclidean geometries.

AI USE STATEMENT

In this work, we used generative AI tools for [generating synthetic data sets, helping develop theoretical models or conceptual frameworks, designing research methods or experiments or providing feedback on them, implementing methods, assisting with translation, assisting in writing proofs, or formulating mathematical claims, providing key elements for proving mathematical claims]. We have not used generative AI tools for [supporting qualitative and thematic data analysis, interpreting results], and [generating or refining hypotheses, cleaning and reformatting data sets] are not applicable to this work. We have reviewed all AI-assisted work. In particular, all LLM-generated code was verified and tested for correctness by two authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Giorgio Angelotti. HyPE: Attention with hyperbolic biases for relative positional encoding. *arXiv preprint arXiv:2310.19676*, 2023.

Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. Hyperbolic busemann learning with ideal prototypes. In *NeurIPS*, 2021.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR*, 2024.

Alexander Bogatskiy, Brandon M. Anderson, Jan T. Offermann, Marwah Roussi, David W. Miller, and Risi Kondor. Lorentz group equivariant neural network for particle physics. In *ICML*, 2020.

Clément Bonet, Lucas Drumetz, and Nicolas Courty. Sliced-wasserstein distances and flows on cartan-hadamard manifolds. *JMLR*, 2025.

Kushal Bose and Swagatam Das. HyPE-GT: where graph transformers meet hyperbolic positional encodings. *arXiv preprint arXiv:2312.06576*, 2023.

Johann Brehmer, Víctor Bresó, Pim de Haan, Tilman Plehn, Huilin Qu, Jonas Spinner, and Jesse Thaler. A lorentz-equivariant transformer for all of the LHC. *SciPost Physics*, 2025.

Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *NeurIPS*, 2019.

Ines Chami, Albert Gu, Dat P Nguyen, and Christopher Ré. HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In *ICML*, 2021.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In *ACL*, 2022.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *NeurIPS*, 2020.

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

Ziheng Chen, Xiao-Jun Wu, Bernhard Schölkopf, and Nicu Sebe. Riemannian batch normalization: A gyro approach. *arXiv preprint arXiv:2509.07115*, 2025b.

Ziheng Chen, Xiaojun Wu, Bernhard Schölkopf, and Nicu Sebe. Building transformation layers for Riemannian neural networks, 2025c. URL https://openreview.net/forum?id=1tJVBCpVD0.

Ziheng Chen, Bernhard Schölkopf, and Nicu Sebe. Hyperbolic Busemann neural networks. In *CVPR*, 2026.

Sungjun Cho, Seunghyuk Cho, Sungwoo Park, Hankook Lee, Honglak Lee, and Moontae Lee. Curve your attention: Mixed-curvature transformers for graph representation learning. *arXiv preprint arXiv:2309.04082*, 2023.

Chang Dai, Hongyu Shan, Mingyang Song, and Di Liang. HoPE: Hyperbolic rotary positional encoding for stable long-range dependency modeling in large language models. *arXiv preprint arXiv:2509.05218*, 2025.

Xiran Fan, Chun-Hao Yang, and Baba C. Vemuri. Nested hyperbolic spaces for dimensionality reduction and hyperbolic NN design. In *CVPR*, 2022.

Xiran Fan, Chun-Hao Yang, and Baba C. Vemuri. Horospherical decision boundaries for large margin classification in hyperbolic space. In *NeurIPS*, 2023.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *ICML*, 2019.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Ben Chamberlain, Michael M. Bronstein, and Federico Monti. SIGN: Scalable inception graph neural networks. *arXiv preprint arXiv:2004.11198*, 2020.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, 2018.

I. M. Gel'fand, R. A. Minlos, and Z. Ya. Shapiro. *Representations of the Rotation and Lorentz Groups and Their Applications*. Dover Publications, 2018.

Shiqi Gong, Qi Meng, Jue Zhang, Huilin Qu, Congqiao Li, Sitian Qian, Weitao Du, Zhi-Ming Ma, and Tie-Yan Liu. An efficient lorentz equivariant graph neural network for jet tagging. *Journal of High Energy Physics*, 2022.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter W. Battaglia, Victor Bapst, David Raposo, Adam Santoro, and Nando de Freitas. Hyperbolic attention networks. In *ICLR*, 2019.

Neil He, Menglin Yang, and Rex Ying. Hypercore: The core framework for building hyperbolic foundation models with comprehensive modules. *arXiv preprint arXiv:2504.08912*, 2025.

Kyle Helfrich, Devin Willmott, and Qiang Ye. Orthogonal recurrent neural networks with scaled cayley transform. In *International conference on machine learning*, pp. 1969–1978. PMLR, 2018.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *NeurIPS*, 2020.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

Marc Teva Law, Renjie Liao, Jake Snell, and Richard S. Zemel. Lorentzian distance learning for hyperbolic representations. In *ICML*, 2019.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *Advances in neural information processing systems*, 32, 2019.

Galileo Mark Namata, Ben London, Lise Getoor, and Bert Huang. Query-driven active surveying for collective classification. In *MLG*, 2012.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, 2018.

Jongmin Park, Seunghoon Han, Soohwan Jeong, and Sungsu Lim. Hyperbolic heterogeneous graph attention networks. In *WWW Companion*, 2024.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *NeurIPS*, 2022.

John G. Ratcliffe. *Foundations of Hyperbolic Manifolds*. Springer, 2006.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. DropEdge: Towards deep graph convolutional networks on node classification. In *ICLR*, 2020.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 2008.

Xianglong Shi, Ziheng Chen, Yunhan Jiang, and Nicu Sebe. Intrinsic lorentz neural network, 2026.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *ICLR*, 2021.

Abraham A. Ungar. *Analytic Hyperbolic Geometry and Albert Einstein's Special Theory of Relativity*. World Scientific, 2008.

Abraham A. Ungar. *Analytic Hyperbolic Geometry in N Dimensions: An Introduction*. CRC Press, 2015.

Abraham Albert Ungar. *Analytic Hyperbolic Geometry and Albert Einstein's Special Theory of Relativity (Second Edition)*. World Scientific, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

Rui Wang, Chen Hu, Xiaoning Song, Xiaojun Wu, Nicu Sebe, and Ziheng Chen. Towards a general attention framework on gyrovector spaces for matrix manifolds. In *NeurIPS*, 2025.

Felix Wu, Amauri H. Souza Jr., Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. In *ICML*, 2019.

Qitian Wu, Wentao Zhao, Zenan Li, David P. Wipf, and Junchi Yan. NodeFormer: A scalable graph structure learning transformer for node classification. In *NeurIPS*, 2022.

Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan. SGFormer: Simplifying and empowering transformers for large-graph representations. In *NeurIPS*, 2023.

Zhanghao Wu, Paras Jain, Matthew A. Wright, Azalia Mirhoseini, Joseph E. Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. In *NeurIPS*, 2021.

Menglin Yang, Harshit Verma, Delvin Ce Zhang, Jiahong Liu, Irwin King, and Rex Ying. Hypformer: Exploring efficient hyperbolic transformer fully in hyperbolic space. In *KDD*, 2024.

Xin Yang, Xingrun Li, Heng Chang, Jinze Yang, Xihong Yang, Shengyu Tao, Maiko Shigeno, Ningkang Chang, Junfeng Wang, Dawei Yin, and Erxue Min. Hgformer: Hyperbolic graph transformer for collaborative filtering. In *ICML*, 2025.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform badly for graph representation? In *NeurIPS*, 2021.

A Notation **14**

B Related Work **14**

B.1 Hyperbolic neural networks 14

B.2 Transformers in hyperbolic space 14

C Preliminaries **15**

C.1 Operations in the Lorentz model 15

C.2 Hyperbolic gyrovector structures 16

C.3 Metric geometry 18

D Lorentz MLR's Output Logit $v_{z_k,a_k}(x)$ **19**

E Proofs **19**

E.1 Proof of Proposition 1 19

E.2 Proof of Theorem 1 20

E.3 Proof of Theorem 2 21

E.4 Euclidean inner product from the Busemann definition 22

E.5 Flat limit of Busemann Function 23

E.6 Flat limit of Lorentzian midpoint 23

F Discussion of Our Homomorphic Transformations **24**

F.1 Current maximality and the choice of $SO(n)$. 24

F.2 Expressiveness, dimension-changing cases, and future work. 24

G Additional Ablation Studies **25**

G.1 Benefits of Native Homomorphic Mappings and Dot-Product Attention 25

G.2 Ablation on L2EFC 25

H More Implementation Details **25**

H.1 Dataset Details 25

H.2 Additional Vision Classification Experiments 26

H.3 Graph Experiments 26

H.4 Text and Vision Experiments 27

A NOTATION

Table 9 summarizes the notation used throughout the main paper and this supplementary material. We use bold lowercase letters for Euclidean vectors and ambient Lorentz vectors when the context is clear. For a Lorentz point $x \in \mathbb{L}_K^n$, we write $x = (x_t, x_s)$ for its temporal and spatial components.

Table 9: Summary of notation used in the main paper and supplementary material.

Notation	Description
$\mathbb{R}^n$	n-dimensional Euclidean space.
$\langle \cdot, \cdot \rangle$ and $\| \cdot \|$	Euclidean inner product and norm.
$\mathbb{L}_K^n$ and $\mathbb{P}_K^n$	Lorentz model and Poincaré ball model with curvature $K < 0$.
$\overline{\mathbf{0}}$	Canonical origin in the hyperbolic model under discussion.
$\langle x, y \rangle_{\mathcal{L}}$ and $\|v\|_{\mathcal{L}}$	Lorentz inner product and the induced norm on tangent vectors.
$x = (x_t, x_s)$	Temporal and spatial decomposition of a Lorentz point or ambient vector.
$\mathcal{T}_x \mathbb{L}_K^n$	Tangent space at $x \in \mathbb{L}_K^n$.
$\exp_x^K(v)$, $\log_x^K(y)$	Exponential and logarithmic maps on $\mathbb{L}_K^n$.
$\mathrm{PT}_{x \to y}^K(v)$	Parallel transport of $v \in \mathcal{T}_x \mathbb{L}_K^n$ from x to y.
$d_{\mathcal{L}}(x, y)$	Geodesic distance on the Lorentz model.
$\oplus_K$ and $\otimes_K$	Lorentz gyroaddition and scalar gyromultiplication.
Φ_R	Isotropy-induced homomorphism with $R \in SO(n)$ used for query and value updates.
$B^v(x)$	Busemann function associated with the unit direction $v \in \mathbb{S}^{n-1}$.
H_τ^v	Horosphere defined by the level set $B^v(x) = \tau$.
q_i, k_j, and v_j	Query point, key direction, and value point in one attention head.
α_{ij}	Attention weight from the ith query to the jth key-value pair.
$v_{z,a}(x)$	Lorentz MLR logit induced by the hyperplane parameters (z, a).
$e^{(k)} \in \mathbb{R}^m$	kth standard basis vector in the output space.
$\mathrm{L2EFC}(x)$	Point-to-hyperplane transformation from $\mathbb{L}_K^n$ to $T_{\overline{\mathbf{0}}} \mathbb{L}_K^m$.

B RELATED WORK

B.1 HYPERBOLIC NEURAL NETWORKS

Early hyperbolic neural networks primarily operated in the Poincaré model, where the gyrovector formalism makes many neural primitives straightforward to define (Ganea et al., 2018; Shimizu et al., 2021). However, the unit-ball constraint and boundary saturation can cause numerical instabilities, which has motivated a shift toward Lorentz neural networks that often exhibit improved optimization stability and a cleaner isometry structure (Nickel & Kiela, 2018; Fan et al., 2022; Bdeir et al., 2024). Within this line of work, prior studies have implemented Lorentz-space counterparts of classical modules. Lorentz-space multinomial logistic regression has been derived and employed as a classification head within fully hyperbolic architectures (Chen et al., 2022; Bdeir et al., 2024). Fully connected mappings have been realized using Lorentz transformations and their relaxations, as well as through intrinsic constructions that preserve the Lorentz constraint (Ganea et al., 2018; Shimizu et al., 2021; Chen et al., 2022; Fan et al., 2022). Convolutional operators have been developed in the Lorentz model for images and graphs, demonstrating practical benefits on vision benchmarks and large networks (Bdeir et al., 2024; Chami et al., 2019; Gong et al., 2022; Bogatskiy et al., 2020). Attention mechanisms have been explored from early hyperbolic attention to fully hyperbolic Transformers and Lorentz-equivariant Transformers, which align the architecture with the geometry of the data (Gulcehre et al., 2019; Chen et al., 2022; Yang et al., 2024; Brehmer et al., 2025). Overall, the strategy for building hyperbolic neural networks is to identify principled replacements of Euclidean primitives. Due to the Euclidean inner-product similarity admitting a representation through the Busemann function, it can be naturally extended to the hyperbolic space.

B.2 TRANSFORMERS IN HYPERBOLIC SPACE

A lot of research adapts Transformer architectures to hyperbolic manifolds. Early work recast the soft attention mechanism with operations defined on hyperbolic models and showed that attention

can be formulated directly on hyperbolic activations to better capture hierarchical and power-law structure (Gulcehre et al., 2019). HyPE-GT introduces learnable hyperbolic positional encodings into graph Transformers and reports improvements on molecular benchmarks (Bose & Das, 2023). For sequence modeling, HyPE proposes attention biases based on hyperbolic functions for relative positional encoding, with experimental evaluation left for future work (Angelotti, 2023). HoPE uses hyperbolic rotary positional encoding for long-context modeling in large language models (Dai et al., 2025). More recent studies aim to realize a complete Transformer in hyperbolic space. Hypformer instantiates core modules entirely in the Lorentz model and proposes a linear self-attention mechanism that scales to billion-scale graphs and long sequences (Yang et al., 2024). Related lines explore hyperbolic attention for heterogeneous graphs and hyperbolic graph Transformers for recommendation, which further demonstrate the utility of negative curvature in attention-based architectures, although they are tailored to specific data regimes (Park et al., 2024; Yang et al., 2025; Gulcehre et al., 2019). In this paper, we present HBformer, a Homomorphism and Busemann approach to hyperbolic Transformers that operates intrinsically in the Lorentz model.

C Preliminaries

C.1 Operations in the Lorentz model

Setup and notation. Fix a negative curvature $K < 0$. Let $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ denote the Minkowski bilinear form with signature $(-, +, \ldots, +)$ on $\mathbb{R}^{n+1}$, and write $\| \cdot \|_{\mathcal{L}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{L}}}$ for the induced (Riemannian) norm on tangent vectors. The n-dimensional hyperbolic space in the Lorentz (hyperboloid) model is

$$\mathbb{L}_K^n := \Big\{ \mathbf{x} \in \mathbb{R}^{n+1} \ : \ \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = \tfrac{1}{K}, \ \ x_0 > 0 \Big\},$$

and we use $\overline{\mathbf{0}} := \big(\frac{1}{\sqrt{-K}}, \mathbf{0}\big)$ as the pole ("origin"). When no confusion can arise, $\| \cdot \|$ denotes the Euclidean norm in a tangent space.

Distance. For $\mathbf{x}, \mathbf{y} \in \mathbb{L}_K^n$, the geodesic distance inherited from Minkowski space is

$$d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{-K}} \ \operatorname{arcosh}\big(K \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}\big).$$

A useful identity for computations is the "Lorentzian chord" expression (squared distance)

$$D_{\mathcal{L}}^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_{\mathcal{L}}^2 = \frac{2}{K} - 2 \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}.$$

Specializing to the pole $\overline{\mathbf{0}}$,

$$d_{\mathcal{L}}(\mathbf{x}, \overline{\mathbf{0}}) = \big\| \log_{\overline{\mathbf{0}}}^K(\mathbf{x}) \big\|,$$
$$D_{\mathcal{L}}^2(\mathbf{x}, \overline{\mathbf{0}}) = \frac{2}{K} - 2 \, \langle \mathbf{x}, \overline{\mathbf{0}} \rangle_{\mathcal{L}} = \frac{2}{K} + \frac{2 \, x_0}{\sqrt{-K}},$$

and, equivalently,

$$d_{\mathcal{L}}(\mathbf{x}, \overline{\mathbf{0}}) = \frac{1}{\sqrt{-K}} \ \operatorname{arcosh}\big(\sqrt{-K} \, x_0\big).$$

Tangent space. The tangent space at $\mathbf{x} \in \mathbb{L}_K^n$ is the Minkowski-orthogonal complement of $\mathbf{x}$,

$$\mathcal{T}_{\mathbf{x}} \mathbb{L}_K^n = \big\{ \mathbf{v} \in \mathbb{R}^{n+1} \ : \ \langle \mathbf{v}, \mathbf{x} \rangle_{\mathcal{L}} = 0 \big\}.$$

Restricted to $\mathcal{T}_{\mathbf{x}} \mathbb{L}_K^n$, the metric is positive definite and coincides with the Riemannian metric of $\mathbb{L}_K^n$.

Exponential and logarithmic maps. For $\mathbf{z} \in \mathcal{T}_{\mathbf{x}} \mathbb{L}_K^n$,

$$\exp_{\mathbf{x}}^K(\mathbf{z}) = \cosh(\alpha) \, \mathbf{x} \ + \ \sinh(\alpha) \, \frac{\mathbf{z}}{\alpha}, \qquad \alpha = \sqrt{-K} \, \|\mathbf{z}\|_{\mathcal{L}}.$$

The inverse map $\log_{\mathbf{x}}^K : \mathbb{L}_K^n \to \mathcal{T}_{\mathbf{x}} \mathbb{L}_K^n$ sends $\mathbf{y} \in \mathbb{L}_K^n$ to

$$\log_{\mathbf{x}}^K(\mathbf{y}) = \frac{\operatorname{arcosh}(\beta)}{\sqrt{\beta^2 - 1}} \ \big(\mathbf{y} - \beta \, \mathbf{x}\big), \qquad \beta = K \, \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}}.$$

At the pole $\overline{\mathbf{0}}$ these simplify to

$$\exp_{\overline{\mathbf{0}}}^{K}(\mathbf{z}) = \tfrac{1}{\sqrt{-K}}\left(\cosh\!\left(\sqrt{-K}\,\|\mathbf{z}\|\right),\ \sinh\!\left(\sqrt{-K}\,\|\mathbf{z}\|\right)\tfrac{\mathbf{z}}{\|\mathbf{z}\|}\right)$$

$$\log_{\overline{\mathbf{0}}}^{K}(\mathbf{y}) = \begin{cases} \mathbf{0}, & \mathbf{y} = \overline{\mathbf{0}}, \\ \dfrac{\operatorname{arcosh}(\beta)}{\sqrt{\beta^2-1}}\left(\mathbf{y} - \beta\,\overline{\mathbf{0}}\right), & \text{otherwise}, \end{cases} \qquad \beta := K\,\langle \overline{\mathbf{0}}, \mathbf{y}\rangle_{\mathcal{L}}$$

with the convention $\mathbf{z}/\|\mathbf{z}\| = \mathbf{0}$ when $\mathbf{z} = \mathbf{0}$.

Parallel transport. Transporting $\mathbf{v} \in \mathcal{T}_{\mathbf{x}}\mathbb{L}_K^n$ along the geodesic from $\mathbf{x}$ to $\mathbf{y}$ yields

$$\begin{aligned} \mathrm{PT}_{\mathbf{x}\to\mathbf{y}}^{K}(\mathbf{v}) &= \mathbf{v} - \frac{\langle \log_{\mathbf{x}}^{K}(\mathbf{y}),\ \mathbf{v}\rangle_{\mathcal{L}}}{d_{\mathcal{L}}(\mathbf{x},\mathbf{y})^2}\left(\log_{\mathbf{x}}^{K}(\mathbf{y}) + \log_{\mathbf{y}}^{K}(\mathbf{x})\right) \\ &= \mathbf{v} + \frac{\langle \mathbf{y}, \mathbf{v}\rangle_{\mathcal{L}}}{\frac{1}{-K} - \langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}}\,(\mathbf{x} + \mathbf{y}). \end{aligned}$$

Lorentz transformations. A matrix $\mathbf{A} \in \mathbb{R}^{(n+1)\times(n+1)}$ is a Lorentz transformation if it preserves the Minkowski product: $\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y}\rangle_{\mathcal{L}} = \langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}$ for all $\mathbf{x}, \mathbf{y}$. Such matrices form the Lorentz group $\mathbf{O}(1,n)$ (equivalently, $\mathbf{A}^\top \eta\, \mathbf{A} = \eta$ for the Minkowski metric η). Restricting to transformations that map the upper sheet to itself yields the time-orientation–preserving subgroup

$$\mathbf{O}^{+}(1,n) = \left\{\mathbf{A} \in \mathbf{O}(1,n)\ :\ (\mathbf{A}\mathbf{x})_0 > 0 \text{ for all } \mathbf{x} \in \mathbb{L}_K^n\right\},$$

which acts by isometries on $\mathbb{L}_K^n$.

Further restricting to the proper orthochronous subgroup

$$\mathbf{SO}^{+}(1,n) = \left\{\mathbf{A} \in \mathbf{O}^{+}(1,n) : \det \mathbf{A} = 1\right\},$$

every $\mathbf{A} \in \mathbf{SO}^{+}(1,n)$ admits a unique polar decomposition $\mathbf{A} = \mathbf{R}\mathbf{B}$ into a spatial rotation and a Lorentz boost. The rotation fixes the time axis and rotates spatial coordinates:

$$\mathbf{R} = \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & \tilde{\mathbf{R}} \end{bmatrix}, \qquad \tilde{\mathbf{R}} \in \mathbf{SO}(n).$$

A boost with velocity vector $\mathbf{v} \in \mathbb{R}^n$ ($\|\mathbf{v}\| < 1$) has the block form

$$\mathbf{B} = \begin{bmatrix} \gamma & -\gamma\,\mathbf{v}^\top \\ -\gamma\,\mathbf{v} & \mathbf{I}_n + \dfrac{\gamma^2}{1+\gamma}\,\mathbf{v}\mathbf{v}^\top \end{bmatrix}, \qquad \gamma = \frac{1}{\sqrt{1 - \|\mathbf{v}\|^2}}.$$

(Equivalently, the spatial block can be written $\mathbf{I}_n + \frac{\gamma-1}{\|\mathbf{v}\|^2}\mathbf{v}\mathbf{v}^\top$ when $\mathbf{v} \neq \mathbf{0}$.)

C.2 Hyperbolic gyrovector structures

C.2.1 Gyrogroups

We recall the algebraic notion of a gyrogroup, which extends the concept of a group to settings where associativity is relaxed and corrected by gyrations (Ungar, 2008; 2015).

Definition 1 (Gyrogroup)**.** *Let G be a nonempty set endowed with a binary operation $\oplus\colon G \times G \to G$. The pair $(G, \oplus)$ is a* gyrogroup *if, for all $a, b, c \in G$, the following axioms hold:*

(G1) There exists an element $e \in G$ such that $e \oplus a = a$ (left identity).
(G2) For each $a \in G$ there exists $\ominus a \in G$ such that $\ominus a \oplus a = e$ (left inverse).
(G3) There exists a map $\mathrm{gyr}[a,b]\colon G \to G$ (the gyration generated by a, b) such that

$$a \oplus (b \oplus c) = (a \oplus b) \oplus \mathrm{gyr}[a,b](c) \quad \textit{(left gyroassociative law).}$$

(G4) $\mathrm{gyr}[a,b] = \mathrm{gyr}[a \oplus b, b]$ (left reduction law).

If in addition

$$a \oplus b = \mathrm{gyr}[a, b](b \oplus a), \qquad \forall\, a, b \in G,$$

then $(G, \oplus)$ *is called* gyrocommutative*.*

Gyrogroups generalize groups: when all gyrations are the identity map, $(G, \oplus)$ reduces to a usual group. The lack of strict associativity is compensated by the gyration operators, which encode curvature induced nonlinearity.

C.2.2 From gyrogroups to gyrovector spaces

To model both addition and scalar multiplication in curved geometries, gyrogroups can be enriched to gyrovector spaces (Ungar, 2022; Chen et al., 2025b).

Definition 2 (Gyrovector space)**.** *Let* $(G, \oplus)$ *be a gyrocommutative gyrogroup and let* $\odot\colon \mathbb{R} \times G \to G$ *be a map called scalar multiplication. The triple* $(G, \oplus, \odot)$ *is a* gyrovector space *if, for all* $a, b, c \in G$ *and* $s, t \in \mathbb{R}$*,*

(V1) $1 \odot a = a$*.*
(V2) $(s + t) \odot a = s \odot a \oplus t \odot a$*.*
(V3) $(st) \odot a = s \odot (t \odot a)$*.*
(V4) $\mathrm{gyr}[a, b](t \odot c) = t \odot \mathrm{gyr}[a, b](c)$*.*
(V5) $\mathrm{gyr}[s \odot a, t \odot a]$ *is the identity map on* G*.*

These axioms mirror the familiar properties of vector spaces, with gyrations accounting for the deviation from linearity. In particular, (V2) and (V3) play the role of distributivity and associativity for scalar multiplication, while (V4)–(V5) guarantee a consistent interaction between gyrations and scaling.

C.2.3 Hyperbolic gyrovector spaces

The gyro-structure over the hyperbolic space can be defined by its Riemannian operators (Ganea et al., 2018; Chen et al., 2025b). Given a hyperbolic model $\mathcal{H}^n_K$ along with $x, y, z \in \mathcal{H}^n_K$ and $t \in \mathbb{R}$, the gyroaddition and gyromultiplication are defined as

$$\begin{aligned} x \oplus_{\mathcal{H}} y &= \mathrm{Exp}_x\left(\mathrm{PT}_{e \to x}\left(\mathrm{Log}_e\, y\right)\right), \\ t \odot_{\mathcal{H}} x &= \mathrm{Exp}_e\left(t\, \mathrm{Log}_e\, x\right), \\ \mathrm{gyr}[x, y]z &= \ominus_{\mathcal{H}}\left(x \oplus_{\mathcal{H}} y\right) \oplus_{\mathcal{H}} \left(x \oplus_{\mathcal{H}} \left(y \oplus_{\mathcal{H}} z\right)\right), \end{aligned} \tag{20}$$

where e denotes the origin in $\mathcal{H}^n_K$.

On the Poincaré ball $\mathbb{P}^n_K$, such gyro-structure is known as the Möbius gyrovector space (Ungar, 2022, Ch. 6.14):

$$\begin{aligned} x \oplus_{\mathrm{M}} y &= \frac{\left(1 - 2K\langle x, y\rangle - K\|y\|^2\right) x + \left(1 + K\|x\|^2\right) y}{1 - 2K\langle x, y\rangle + K^2\|x\|^2\|y\|^2}, \\ t \odot_{\mathrm{M}} x &= \frac{\tanh\left(t \tanh^{-1}(\sqrt{|K|}\|x\|)\right)}{\sqrt{|K|}} \frac{x}{\|x\|} \end{aligned}$$

where $\ominus_{\mathrm{M}} x = -1 \odot_{\mathrm{M}} x = -x$ is the gyroinverse and $\mathbf{0}$ is the gyro identity: $\mathbf{0} \oplus_{\mathrm{M}} x = x, \forall x \in \mathbb{P}^n_K$. As shown by Chen et al. (2025b, Props. 24-25), the Lorentz gyroaddition and gyromultiplication also admit closed-form expressions:

$$x \oplus_{\mathbb{L}} y = \begin{cases} x, & y = \overline{\mathbf{0}}, \\ y, & x = \overline{\mathbf{0}}, \\ \begin{bmatrix} \frac{1}{\sqrt{|K|}} \frac{D - KN}{D + KN} \\ \frac{2(A_s x_s + A_y y_s)}{D + KN} \end{bmatrix}, & \text{Otherwise.} \end{cases} \tag{21}$$

$$t \odot_{\mathbb{L}} x = \begin{cases} \overline{\mathbf{0}}, & t = 0 \vee x = \overline{\mathbf{0}} \\ \frac{1}{\sqrt{|K|}} \begin{bmatrix} \cosh(t\theta) \\ \frac{\sinh(t\theta)}{\|x_s\|} x_s \end{bmatrix}, & \text{Otherwise,} \end{cases} \tag{22}$$

Here, $\theta = \cosh^{-1}(\sqrt{|K|}x_t)$, $A_s = ab^2 - 2Kbs_{xy} - Kan_y$ and $A_y = b(a^2 + Kn_x)$ with the following notation:

$$\begin{aligned} &a = 1 + \sqrt{|K|}x_t, b = 1 + \sqrt{|K|}y_t, \\ &n_x = \|x_s\|^2, n_y = \|y_s\|^2, s_{xy} = \langle x_s, y_s \rangle, \\ &D = a^2b^2 - 2Kabs_{xy} + K^2 n_x n_y, \\ &N = a^2 n_y + 2abs_{xy} + b^2 n_x. \end{aligned} \tag{23}$$

In particular, the gyro identity is $\overline{\mathbf{0}}$ and the gyroinverse is $\ominus_{\mathbb{L}} x = -1 \odot_{\mathbb{L}} x = [x_t, -x_s^\top]^\top$.

C.3 Metric geometry

Geodesic, geodesic ray, and asymptotic rays. Let (X, d) be a geodesic metric space. Geodesics are length-minimizing curves, and we consider unit-speed geodesics. A geodesic ray is a geodesic $\gamma : [0, \infty) \to X$ such that

$$d(\gamma(s), \gamma(t)) = |t - s|, \qquad s, t \geq 0. \tag{24}$$

Two geodesic rays $\gamma_1, \gamma_2 : [0, \infty) \to X$ are *asymptotic* if

$$\sup_{t \geq 0} d(\gamma_1(t), \gamma_2(t)) < \infty. \tag{25}$$

This extends the notion of Euclidean parallel lines in two ways (Ratcliffe, 2006): Euclidean parallels have constant separation, while asymptotic rays only require bounded separation. Euclidean parallels are disjoint, whereas asymptotic rays in hyperbolic space need not be disjoint.

Busemann function. Let γ be a geodesic ray and let $d(\cdot, \cdot)$ denote the geodesic distance. The Busemann function associated with γ is defined by

$$\forall x \in X, \qquad B^\gamma(x) = \lim_{t \to \infty} \left(d(x, \gamma(t)) - t \right). \tag{26}$$

In hyperbolic spaces, this limit exists (Ratcliffe, 2006; Atigh et al., 2021). In Euclidean space, for $\gamma(t) = tv$ with $v \in \mathbb{S}^{n-1}$ and $x \in \mathbb{R}^n$,

$$B^v(x) = -\langle x, v \rangle. \tag{27}$$

Hence, up to sign, the Busemann function recovers the inner product and provides an intrinsic generalization of directional similarity to curved spaces. For asymptotic geodesic rays, Busemann functions differ only by an additive constant (Ratcliffe, 2006; Atigh et al., 2021).

Origin-based representative and closed forms. Therefore, up to an additive constant, we may choose rays that emanate from the canonical origin $\overline{\mathbf{0}}$. For $v \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^n$ with $\|v\| = 1$, define

$$\gamma_v(t) = \exp_{\overline{\mathbf{0}}}^K(tv), \qquad t \geq 0, \qquad B^v(x) := B^{\gamma_v}(x). \tag{28}$$

For $v \in \mathbb{S}^{n-1}$ and $x \in \mathbb{P}_K^n$, the Poincaré closed form is (Bonet et al., 2025)

$$B^v(x) = \frac{1}{\sqrt{-K}} \log\left(\frac{\|v - \sqrt{-K}x\|^2}{1 + K\|x\|^2} \right). \tag{29}$$

For $x = (x_t, x_s) \in \mathbb{L}_K^n$, the Lorentz closed form is

$$B^v(x) = \frac{1}{\sqrt{-K}} \log\left(-\sqrt{-K} \left\langle x, \sqrt{-K}\overline{\mathbf{0}} + v \right\rangle_{\mathcal{L}} \right) = \frac{1}{\sqrt{-K}} \log\left(x_t - \sqrt{-K} \langle x_s, v \rangle \right). \tag{30}$$

Its level sets

$$H_\tau^v = \{ x \in \mathbb{L}_K^n \mid B^v(x) = \tau \} \tag{31}$$

are horospheres (Fan et al., 2023).

D LORENTZ MLR'S OUTPUT LOGIT $v_{z_k,a_k}(x)$

For completeness, we recall the definition of the Lorentz multinomial logistic regression (Lorentz MLR) logit $v_{z_k,a_k}(x)$ that is used in Sec. 3.3. We follow the construction of Bdeir et al. (2024) and adapt their notation to our setting.

Let $K < 0$ and let $x = (x_t, x_s) \in \mathbb{L}_K^n$ with time coordinate $x_t \in \mathbb{R}$ and spatial coordinates $x_s \in \mathbb{R}^n$. For learnable parameters $z \in \mathbb{R}^n$ and $a \in \mathbb{R}$, the Lorentz hyperplane $\tilde{H}_{z,a}$ is

$$\tilde{H}_{z,a} = \Big\{ x \in \mathbb{L}_K^n \;\Big|\; \cosh\Big(\sqrt{-K}\, a\Big)\, \langle z, x_s\rangle - \sinh\Big(\sqrt{-K}\, a\Big)\, \|z\|_2\, x_t = 0 \Big\} \tag{32}$$

which is the image of a tangent vector at the origin transported to a point at signed distance a along the geodesic in direction z.

For $x \in \mathbb{L}_K^n$ the Lorentzian distance from x to $\tilde{H}_{z,a}$ admits the closed form

$$d_{\mathcal{L}}\big(x, \tilde{H}_{z,a}\big) = \frac{1}{\sqrt{-K}} \Big|\sinh^{-1}\Big(\sqrt{-K}\, \frac{\alpha(z,a;x)}{\beta(z,a)}\Big)\Big|, \tag{33}$$

where

$$\alpha(z,a;x) = \cosh\Big(\sqrt{-K}\, a\Big)\, \langle z, x_s\rangle - \sinh\Big(\sqrt{-K}\, a\Big)\, \|z\|_2\, x_t, \tag{34}$$

$$\beta(z,a) = \sqrt{\Big\|\cosh\Big(\sqrt{-K}\, a\Big)\, z\Big\|_2^2 - \Big(\sinh\Big(\sqrt{-K}\, a\Big)\, \|z\|_2\Big)^2}. \tag{35}$$

Here $\alpha(z,a;x) = \langle w_{z,a}, x\rangle_{\mathcal{L}}$ and $\beta(z,a) = \|w_{z,a}\|_{\mathcal{L}}$ are, respectively, the Lorentz inner product and Lorentz norm of a suitably parameterized normal vector $w_{z,a}$ to the hyperplane.

In Euclidean multinomial logistic regression, the logit of class k can be written as a signed distance to an affine hyperplane multiplied by the norm of its normal vector as follows:

$$\begin{aligned} p(y = k \mid \boldsymbol{x}) &\propto \exp(v_k(\boldsymbol{x})), \\ v_k(\boldsymbol{x}) &= \operatorname{sign}(\langle \boldsymbol{a}_k, \boldsymbol{x} - \boldsymbol{p}_k\rangle) \|\boldsymbol{a}_k\| d(\boldsymbol{x}, H_{\boldsymbol{a}_k,\boldsymbol{p}_k}), \;\; \boldsymbol{a}_k \in \mathbb{R}^n, \end{aligned} \tag{36}$$

Replacing the Euclidean distance by the Lorentzian distance Eq. (33) yields the Lorentz MLR logit (Bdeir et al., 2024)

$$v_{z,a}(x) = \tfrac{1}{\sqrt{-K}}\, \operatorname{sign}\big(\alpha(z,a;x)\big)\, \beta(z,a) \Big|\sinh^{-1}\Big(\sqrt{-K}\, \tfrac{\alpha(z,a;x)}{\beta(z,a)}\Big)\Big|. \tag{37}$$

In our notation, the output logit associated with the k-th class or output channel is obtained by taking $(z,a) = (z_k, a_k)$ and writing $v_{z_k,a_k}(x) = v_{z,a}(x)\big|_{z=z_k,\; a=a_k}$, which is exactly the quantity used in Theorem 2 and in the point-to-hyperplane transformation of Sec. 3.3.

E PROOFS

E.1 PROOF OF PROPOSITION 1

We recall that the Lorentz group is defined independently of the curvature parameter. Let

$$J = \operatorname{diag}(-1, 1, \ldots, 1) \in \mathbb{R}^{(n+1)\times(n+1)},$$

and set

$$SO^+(1,n) = \big\{ A \in \mathbb{R}^{(n+1)\times(n+1)} : A^\top J A = J,\; \det(A) = 1,\; A_{00} > 0 \big\}.$$

For $K = -1$, the hyperboloid model is

$$\mathbb{L}_{-1}^n = \big\{ x \in \mathbb{R}^{n+1} : \langle x, x\rangle_{\mathcal{L}} = -x_0^2 + \|x_s\|_2^2 = -1,\; x_0 > 0 \big\},$$

and it is a standard fact that the positive special Lorentz group acts transitively on $\mathbb{L}_{-1}^n$ by matrix multiplication

$$A \cdot x = Ax, \qquad A \in SO^+(1,n),\; x \in \mathbb{L}_{-1}^n,$$

that is, for any $u, v \in \mathbb{L}_{-1}^n$ there exists $A \in SO^+(1,n)$ such that $Au = v$ as shown in Fact 1 of Fan et al. (2022).

We now extend this transitivity to the curvature-K Lorentz model

$$\mathbb{L}_K^n = \big\{ x \in \mathbb{R}^{n+1} : \langle x, x\rangle_{\mathcal{L}} = \tfrac{1}{K},\; x_0 > 0 \big\}, \qquad K < 0.$$

Lemma 1. *For every $K < 0$ the linear map*

$$S_K : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}, \qquad S_K(x) = \sqrt{-K}\, x,$$

induces a bijection between $\mathbb{L}_K^n$ and $\mathbb{L}_{-1}^n$.

Proof. Let $x \in \mathbb{L}_K^n$. Then

$$\big\langle S_K(x), S_K(x) \big\rangle_{\mathcal{L}} = (-K)\, \langle x, x \rangle_{\mathcal{L}} = (-K) \cdot \tfrac{1}{K} = -1,$$

and the time component is $(S_K(x))_0 = \sqrt{-K}\, x_0 > 0$. Thus $S_K(x) \in \mathbb{L}_{-1}^n$, so $S_K(\mathbb{L}_K^n) \subset \mathbb{L}_{-1}^n$. The inverse map is $S_K^{-1}(y) = \frac{1}{\sqrt{-K}}\, y$, and the same calculation shows that $S_K^{-1}(\mathbb{L}_{-1}^n) \subset \mathbb{L}_K^n$. Therefore S_K restricts to a bijection $S_K : \mathbb{L}_K^n \to \mathbb{L}_{-1}^n$ with inverse $S_K^{-1} : \mathbb{L}_{-1}^n \to \mathbb{L}_K^n$. □

Proposition 1. Let $K < 0$ and fix arbitrary $x, y \in \mathbb{L}_K^n$. Define

$$u = S_K(x), \qquad v = S_K(y),$$

which belong to $\mathbb{L}_{-1}^n$ by Lemma 1. By transitivity of $SO^+(1,n)$ on $\mathbb{L}_{-1}^n$, there exists $A \in SO^+(1,n)$ such that $Au = v$.

Since S_K is a scalar multiple of the identity matrix, it commutes with A. Therefore

$$S_K(Ax) = AS_K(x) = Au = v = S_K(y).$$

Applying S_K^{-1} to both sides yields $Ax = y$. Finally, A preserves the Lorentz product, hence $\langle Ax, Ax \rangle_{\mathcal{L}} = \langle x, x \rangle_{\mathcal{L}} = \frac{1}{K}$, so A indeed maps $\mathbb{L}_K^n$ to itself.

Thus for any $x, y \in \mathbb{L}_K^n$ there exists $A \in SO^+(1,n)$ with $Ax = y$, which proves the transitivity of the $SO^+(1,n)$ action on $\mathbb{L}_K^n$ for every $K < 0$. □

E.2 Proof of Theorem 1

We identify $\mathbb{L}_K^n$ with the radius model $\mathcal{M}_K^n$ and write $\oplus_{\mathcal{L}}$ and $\otimes_{\mathcal{L}}$ for its gyroaddition and scalar multiplication, respectively. The stereographic model st_K^n carries gyrooperations $\oplus_K$ and $\otimes_K$. By Chen et al. (2025a), the stereographic projection $\pi_{\mathcal{M}\to st}\colon \mathcal{M}_K^n \to st_K^n$ and its inverse $\pi_{st\to\mathcal{M}}$ are gyro isomorphisms and isometries:

$$\begin{aligned} \pi_{\mathcal{M}\to st}(x \oplus_{\mathcal{L}} y) &= \pi_{\mathcal{M}\to st}(x) \oplus_K \pi_{\mathcal{M}\to st}(y), \\ \pi_{st\to\mathcal{M}}(u \oplus_K v) &= \pi_{st\to\mathcal{M}}(u) \oplus_{\mathcal{L}} \pi_{st\to\mathcal{M}}(v), \end{aligned} \tag{38}$$

and analogously for scalar multiplication.

For $R \in SO(n)$ we set $Q = \mathrm{diag}(1, R)$ and $\Phi_R(x) = Qx$. We prove the three claims in Theorem 1.

(1) Isometry and preservation of the hyperboloid. For $x = (x_0, x_s)$ and $y = (y_0, y_s)$ in $\mathbb{R}^{n+1}$,

$$\begin{aligned} \langle Qx, Qy \rangle_{\mathcal{L}} &= -x_0 y_0 + \langle Rx_s, Ry_s \rangle \\ &= -x_0 y_0 + \langle x_s, y_s \rangle = \langle x, y \rangle_{\mathcal{L}}, \end{aligned}$$

since R is orthogonal. Thus $Q^\top JQ = J$, $\det(Q) = \det(R) = 1$, and $Q_{00} = 1 > 0$, so $Q \in SO^+(1,n)$. Since $Q^\top JQ = J$, Q preserves the ambient Lorentz product and hence the induced metric g_K. Moreover, $Qx = (x_0, Rx_s)$ preserves both the level set $\{x \mid \langle x, x \rangle_{\mathcal{L}} = 1/K\}$ and the forward sheet $x_0 > 0$. Hence Φ_R is an isometry and maps $\mathbb{L}_K^n$ to itself.

(2) Gyrogroup homomorphism. On st_K^n we consider the map $\psi_R\colon st_K^n \to st_K^n$, $\psi_R(u) = Ru$. The explicit formulas of $\oplus_K$ and $\otimes_K$ in Chen et al. (2025a) depend only on Euclidean inner products and norms, which are invariant under $R \in SO(n)$. Consequently R is a gyrovector homomorphism on st_K^n,

$$\begin{aligned} \psi_R(u \oplus_K v) &= \psi_R(u) \oplus_K \psi_R(v), \\ \psi_R(t \otimes_K u) &= t \otimes_K \psi_R(u). \end{aligned} \tag{39}$$

Using the closed-form stereographic maps of Chen et al. (2025a) and $\|Rz\| = \|z\|$, the inverse projection satisfies $\pi_{st\to\mathcal{M}}(Rz) = Q\pi_{st\to\mathcal{M}}(z)$. Taking $z = \pi_{\mathcal{M}\to st}(x)$ gives

$$\pi_{st\to\mathcal{M}}\big(R\,\pi_{\mathcal{M}\to st}(x)\big) = Qx = \Phi_R(x), \qquad x \in \mathcal{M}_K^n, \tag{40}$$

so Φ_R is the conjugate of ψ_R by $\pi_{\mathcal{M}\to st}$. For $x, y \in \mathbb{L}_K^n$,

$$\begin{aligned}
\Phi_R(x \oplus_{\mathcal{L}} y) &= \pi_{st\to\mathcal{M}}\Big(R\,\pi_{\mathcal{M}\to st}(x \oplus_{\mathcal{L}} y)\Big) && \text{by Eq. (40)}\\
&= \pi_{st\to\mathcal{M}}\Big(R\big(\pi_{\mathcal{M}\to st}(x) \oplus_K \pi_{\mathcal{M}\to st}(y)\big)\Big) && \text{by Eq. (38)}\\
&= \pi_{st\to\mathcal{M}}\Big(R\pi_{\mathcal{M}\to st}(x) \oplus_K R\pi_{\mathcal{M}\to st}(y)\Big) && \text{by Eq. (39)}\\
&= \pi_{st\to\mathcal{M}}\big(R\pi_{\mathcal{M}\to st}(x)\big) \oplus_{\mathcal{L}} \pi_{st\to\mathcal{M}}\big(R\pi_{\mathcal{M}\to st}(y)\big) && \text{by Eq. (38)}\\
&= \Phi_R(x) \oplus_{\mathcal{L}} \Phi_R(y),
\end{aligned}$$

which is exactly Eq. (10) (recalling that $\oplus_K$ in Theorem 1 coincides with $\oplus_{\mathcal{L}}$ on $\mathbb{L}_K^n$).

(3) Compatibility with scalar multiplication. The argument for scalar multiplication is analogous. For $t \in \mathbb{R}$ and $x \in \mathbb{L}_K^n$,

$$\begin{aligned}
\Phi_R(t \otimes_{\mathcal{L}} x) &= \pi_{st\to\mathcal{M}}\Big(R\,\pi_{\mathcal{M}\to st}(t \otimes_{\mathcal{L}} x)\Big)\\
&= \pi_{st\to\mathcal{M}}\Big(R\,(t \otimes_K \pi_{\mathcal{M}\to st}(x))\Big) && \text{by Eq. (38)}\\
&= \pi_{st\to\mathcal{M}}\Big(t \otimes_K R\,\pi_{\mathcal{M}\to st}(x)\Big) && \text{by Eq. (39)}\\
&= t \otimes_{\mathcal{L}} \pi_{st\to\mathcal{M}}\big(R\,\pi_{\mathcal{M}\to st}(x)\big) && \text{by Eq. (38)}\\
&= t \otimes_{\mathcal{L}} \Phi_R(x),
\end{aligned}$$

which is Eq. (11).

Combining the three parts completes the proof of Theorem 1.

E.3 Proof of Theorem 2

Proof. **Identify the output space with Euclidean space.** Method in Sec. 3.3 starts from the Euclidean signed distance characterization in Eq. (17) and keeps its left-hand side unchanged. For the Lorentz-to-Euclidean layer, the output lies in $T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$, where $\overline{\mathbf{0}} = [(-K)^{-1/2}, 0]$. This tangent space admits the canonical identification

$$T_{\overline{\mathbf{0}}}\mathbb{L}_K^m = \big\{[0, u] \in \mathbb{R}^{1+m} \mid u \in \mathbb{R}^m\big\} \cong \mathbb{R}^m. \tag{41}$$

Under this identification, every output can be written as $y = [0, y_s]$ with $y_s \in \mathbb{R}^m$, and the induced metric on $T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$ is exactly the Euclidean metric on y_s.

Let $e^{(k)} \in \mathbb{R}^m$ be the k-th standard basis vector and define $\tilde{e}^{(k)} = [0, e^{(k)}] \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$. The corresponding coordinate hyperplane in the tangent space is

$$\bar{H}^K_{\tilde{e}^{(k)},0} = \Big\{[0, u] \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^m \mid \langle e^{(k)}, u\rangle = u_k = 0\Big\}. \tag{42}$$

Via Eq. (41), this is exactly the Euclidean hyperplane

$$H_{e^{(k)},\mathbf{0}} = \Big\{u \in \mathbb{R}^m \mid \langle e^{(k)}, u - \mathbf{0}\rangle = 0\Big\}. \tag{43}$$

Keep the Euclidean signed distance and replace only the score. Method Sec. 3.3 explicitly keeps the left hand side of Eq. (17) and replaces the Euclidean fully connected score by its Lorentz counterpart. Therefore, for an input $x \in \mathbb{L}_K^n$ and output $y = [0, y_s] \in T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$, we require

$$\operatorname{sign}\Big(\langle e^{(k)}, y_s - \mathbf{0}\rangle\Big)\, d\big(y_s, H_{e^{(k)},\mathbf{0}}\big) = v_{z_k,a_k}(x), \qquad k = 1, \dots, m, \tag{44}$$

where $v_{z_k,a_k}(x)$ is the Lorentz MLR logit recalled in Appendix D.

Solve the system coordinatewise. For the coordinate hyperplane $H_{e^{(k)},\mathbf{0}}$, the signed Euclidean distance of y_s to this hyperplane is exactly its k-th coordinate:

$$\operatorname{sign}\Big(\langle e^{(k)}, y_s - \mathbf{0}\rangle\Big)\, d\big(y_s, H_{e^{(k)},\mathbf{0}}\big) = \langle e^{(k)}, y_s\rangle = y_{s,k}. \tag{45}$$

Substituting Eq. (45) into Eq. (44) yields

$$y_{s,k} = v_{z_k,a_k}(x), \qquad k = 1, \dots, m. \tag{46}$$

Hence

$$y_s = (v_{z_1,a_1}(x), \dots, v_{z_m,a_m}(x))^\top. \tag{47}$$

Lifting y_s back to the tangent space gives

$$y = \left[0, (v_{z_k,a_k}(x))_{k=1}^m\right], \tag{48}$$

which is exactly Eq. (18). Therefore, the point-to-hyperplane transformation in Sec. 3.3 defines a unique element of $T_{\overline{\mathbf{0}}}\mathbb{L}_K^m$ whose coordinates are the Lorentz logits, which completes the proof. □

E.4 Euclidean inner product from the Busemann definition

Proposition 2 (Euclidean Busemann identity)**.** *Let $u \in \mathbb{R}^m$ satisfy $\|u\|_2 = 1$, and define the unit speed ray*

$$\delta_u(t) = tu, \qquad t \geq 0. \tag{49}$$

Define the Euclidean Busemann function by

$$B_u(x) := \lim_{t\to\infty} \left(\|x - tu\|_2 - t\right), \qquad x \in \mathbb{R}^m. \tag{50}$$

Then

$$B_u(x) = -\langle x, u\rangle. \tag{51}$$

As a consequence, for any $y \in \mathbb{R}^m$ with $y \neq 0$,

$$\langle x, y\rangle = -\|y\|_2 B_{y/\|y\|_2}(x), \tag{52}$$

which is the identity used in Eq. (12)*.*

Proof. Fix $x \in \mathbb{R}^m$ and set

$$f(t) := \|x - tu\|_2 - t, \qquad t > 0. \tag{53}$$

Using the difference of squares,

$$f(t) = \frac{\|x - tu\|_2^2 - t^2}{\|x - tu\|_2 + t}. \tag{54}$$

Since $\|u\|_2 = 1$,

$$\|x - tu\|_2^2 = \|x\|_2^2 - 2t\langle x, u\rangle + t^2, \tag{55}$$

hence

$$f(t) = \frac{\|x\|_2^2 - 2t\langle x, u\rangle}{\|x - tu\|_2 + t}. \tag{56}$$

From the triangle inequality,

$$t - \|x\|_2 \leq \|x - tu\|_2 \leq t + \|x\|_2. \tag{57}$$

Therefore

$$1 - \frac{\|x\|_2}{t} \leq \frac{\|x - tu\|_2}{t} \leq 1 + \frac{\|x\|_2}{t}, \tag{58}$$

and by squeeze,

$$\lim_{t\to\infty} \frac{\|x - tu\|_2}{t} = 1, \qquad \lim_{t\to\infty} \frac{\|x - tu\|_2 + t}{t} = 2. \tag{59}$$

Also,

$$\lim_{t\to\infty} \frac{\|x\|_2^2 - 2t\langle x, u\rangle}{t} = -2\langle x, u\rangle. \tag{60}$$

Divide numerator and denominator in Eq. (56) by t, then use Eqs. (59) and (60) to obtain

$$\lim_{t\to\infty} f(t) = -\langle x, u\rangle. \tag{61}$$

By Eqs. (50) and (53), this proves Eq. (51).

Now let $y \neq 0$ and set $u = y/\|y\|_2$. Then

$$B_{y/\|y\|_2}(x) = -\left\langle x, \frac{y}{\|y\|_2}\right\rangle = -\frac{\langle x, y\rangle}{\|y\|_2}, \tag{62}$$

which is equivalent to Eq. (52). □

Equivalent definition. In Euclidean space, one can also write

$$B_u(x) = \lim_{t\to\infty} \left(d\left(x, tu\right) - d\left(0, tu\right)\right). \tag{63}$$

Since $d\left(0, tu\right) = \left\|tu\right\|_2 = t$, this is equivalent to Eq. (50).

E.5 Flat limit of Busemann Function

Proposition 3 (Flat limit of normalized Busemann potential)**.** *Let $K < 0$ and set $r = \sqrt{-K}$. Fix $x_s \in \mathbb{R}^d$ and a unit direction $v \in \mathbb{S}^{d-1}$. Define*

$$o_K = \left(\frac{1}{r}, 0\right), \qquad \xi_v = (1, v), \qquad x_K = \left(\sqrt{\frac{1}{r^2} + \left\|x_s\right\|_2^2}, x_s\right) \in \mathbb{L}_K^d. \tag{64}$$

Consider the normalized Busemann potential

$$B_v^{(K)}(x) = \frac{1}{r} \log\left(\frac{-\left\langle x, \xi_v\right\rangle_{\mathcal{L}}}{-\left\langle o_K, \xi_v\right\rangle_{\mathcal{L}}}\right). \tag{65}$$

Then

$$\lim_{K\to 0^-} B_v^{(K)}(x_K) = -\left\langle v, x_s\right\rangle. \tag{66}$$

Proof. Since $\left\langle o_K, \xi_v\right\rangle_{\mathcal{L}} = -1/r$, Eq. (65) gives

$$B_v^{(K)}(x_K) = \frac{1}{r}\log(r\left(-\left\langle x_K, \xi_v\right\rangle_{\mathcal{L}}\right)) = \frac{1}{r}\log(r\left(x_{K,0} - \left\langle v, x_s\right\rangle\right)). \tag{67}$$

Using Eq. (64),

$$B_v^{(K)}(x_K) = \frac{1}{r}\log\left(\sqrt{1 + r^2\left\|x_s\right\|_2^2} - r\left\langle v, x_s\right\rangle\right). \tag{68}$$

Let

$$a = \left\|x_s\right\|_2^2, \qquad b = \left\langle v, x_s\right\rangle, \qquad h(r) = \log\left(\sqrt{1 + r^2 a} - rb\right). \tag{69}$$

Then $h(0) = 0$ and Eq. (68) becomes $B_v^{(K)}(x_K) = h(r)/r$. Moreover,

$$h'(r) = \frac{\frac{ra}{\sqrt{1+r^2a}} - b}{\sqrt{1 + r^2 a} - rb}, \qquad h'(0) = -b. \tag{70}$$

Hence

$$\lim_{K\to 0^-} B_v^{(K)}(x_K) = \lim_{r\to 0^+} \frac{h(r)}{r} = h'(0) = -b = -\left\langle v, x_s\right\rangle, \tag{71}$$

which proves Eq. (66). □

E.6 Flat limit of Lorentzian midpoint

Proposition 4 (Flat limit in space-like coordinates)**.** *Fix a query index i. Let $\alpha_{ij} \geq 0$ and $\sum_{j=1}^N \alpha_{ij} = 1$. Let $v_j \in \mathbb{R}^d$ be fixed Euclidean vectors, and define their curvature-dependent lifts*

$$V_j(K) = (t_j(K), v_j) \in \mathbb{L}_K^d, \qquad t_j(K) = \sqrt{\frac{1}{|K|} + \left\|v_j\right\|_2^2}. \tag{72}$$

Define the Lorentzian midpoint output by

$$Y_i(K) = \frac{\sum_{j=1}^N \alpha_{ij} V_j(K)}{\sqrt{|K|\left|\left\langle \sum_{j=1}^N \alpha_{ij} V_j(K), \sum_{j=1}^N \alpha_{ij} V_j(K)\right\rangle_{\mathcal{L}}\right|}}. \tag{73}$$

Then the space-like component satisfies

$$\lim_{K\to 0^-} \left(Y_i(K)\right)_s = \sum_{j=1}^N \alpha_{ij} v_j. \tag{74}$$

Proof. Set

$$m_i := \sum_{j=1}^{N} \alpha_{ij} v_j, \qquad B_i(K) := \sum_{j=1}^{N} \alpha_{ij} \sqrt{1 + |K| \, \|v_j\|_2^2}. \tag{75}$$

From $\sum_{j=1}^{N} \alpha_{ij} = 1$, we have $B_i(K) \to 1$ as $K \to 0^-$. Let

$$S_i(K) := \sum_{j=1}^{N} \alpha_{ij} V_j(K). \tag{76}$$

Using Eq. (72),

$$S_i(K) = \left(\frac{B_i(K)}{\sqrt{|K|}}, m_i \right). \tag{77}$$

Hence

$$\langle S_i(K), S_i(K) \rangle_{\mathcal{L}} = -\frac{B_i(K)^2}{|K|} + \|m_i\|_2^2 . \tag{78}$$

For K sufficiently close to 0^-, the right-hand side in Eq. (78) is negative, so

$$|K| \, |\langle S_i(K), S_i(K) \rangle_{\mathcal{L}}| = B_i(K)^2 - |K| \, \|m_i\|_2^2 . \tag{79}$$

Substituting Eq. (77) and Eq. (79) into Eq. (73) yields

$$\left(Y_i(K)\right)_s = \frac{m_i}{\sqrt{B_i(K)^2 - |K| \, \|m_i\|_2^2}}. \tag{80}$$

Since $B_i(K) \to 1$ and $|K| \, \|m_i\|_2^2 \to 0$, the denominator in Eq. (80) converges to 1. Therefore

$$\lim_{K \to 0^-} \left(Y_i(K)\right)_s = m_i = \sum_{j=1}^{N} \alpha_{ij} v_j, \tag{81}$$

which proves Eq. (74). □

F Discussion of Our Homomorphic Transformations

F.1 Current maximality and the choice of $SO(n)$.

In Euclidean space, linear homomorphisms admit the general and tractable form $x \mapsto Ax$. In a gyrovector space, however, a comparable general characterization of gyro-homomorphisms is not currently available. We therefore identify Lorentz gyro-homomorphisms through the sufficient condition of being an origin-preserving isometry. In Lorentz space, the maximal family satisfying this condition can be shown to be the orthogonal group $\{x \mapsto Rx \mid R \in O(n)\}$. Although [(Helfrich et al., 2018), Thm. 3.1] provides a scaled Cayley parameterization of $O(n)$, it introduces an additional diagonal sign matrix D, with the number ρ of its negative entries treated as a hyperparameter. For simpler optimization, we use $SO(n)$.

F.2 Expressiveness, dimension-changing cases, and future work.

Our sufficient-condition construction identifies only a tractable subclass of Lorentz gyro-homomorphisms and may therefore sacrifice expressiveness relative to the full, as-yet-uncharacterized class. Moreover, a dimension-reducing map with $d_k < d$ is not covered by our current homomorphic construction, since the present work considers only equal-dimensional transformations. However, characterizing broader Lorentz gyro-homomorphisms and developing dimension-changing homomorphic maps are two important directions for future work.

G Additional Ablation Studies

G.1 Benefits of Native Homomorphic Mappings and Dot-Product Attention

The Euclidean-limit property is a key geometric motivation of our design. As $K \to 0^-$, the Busemann similarity reduces to an inner-product form, the homomorphism-constrained query and value transformations reduce to Euclidean orthogonal linear maps, and Lorentz midpoint aggregation becomes Euclidean weighted averaging. These properties connect our hyperbolic attention to its Euclidean counterpart, with proofs for the similarity and aggregation given in Sections E.5 and E.6.

To examine performance as curvature approaches zero, we compare HBformer with Hypformer on 20News-Groups at five Lorentz parameter values $k_{\text{out}} \in \{0.5, 1, 2, 4, 20\}$, corresponding to $K = -1/k_{\text{out}}$. As shown in Table 10, HBformer achieves higher mean accuracy at every tested curvature, with gains of 0.41, 0.84, 0.63, 0.99, and 1.98 percentage points, respectively. The largest gain occurs at $K = -0.05$, the setting closest to zero in this sweep, where HBformer reaches 69.32% compared with Hypformer's 67.34%. Overall, HBformer's improvement over Hypformer tends to increase as curvature approaches zero, the Euclidean limit.

Table 10: Accuracy (%) on 20News-Groups at different curvatures $K = -1/k_{\text{out}}$. Subscripts report standard deviations, and gain denotes the difference in mean accuracy in percentage points. Bold marks the higher mean accuracy in each row.

k_{out}	Curvature K	Hypformer	HBformer	Gain
0.5	-2	$68.35_{\pm 0.34}$	$\mathbf{68.76}_{\pm 0.55}$	$+0.41$
1	-1	$67.98_{\pm 0.54}$	$\mathbf{68.82}_{\pm 0.51}$	$+0.84$
2	-0.5	$68.21_{\pm 0.44}$	$\mathbf{68.84}_{\pm 0.78}$	$+0.63$
4	-0.25	$68.10_{\pm 0.96}$	$\mathbf{69.09}_{\pm 0.55}$	$+0.99$
20	-0.05	$67.34_{\pm 1.05}$	$\mathbf{69.32}_{\pm 0.54}$	$+1.98$

G.2 Ablation on L2EFC

We isolate the contribution of L2EFC by replacing it with the logarithmic map while retaining the homomorphism-constrained transformations and Busemann attention. An additional FC layer is applied before the logarithmic map to match the number of parameters. As shown in Table 11, L2EFC achieves higher mean accuracy on all four datasets, with gains of 0.26, 0.16, 0.40, and 0.02 percentage points on Cora, PubMed, Mini-ImageNet, and 20News-Groups, respectively. These results support the effectiveness of the point-to-hyperplane key transformation, although the gain on 20News-Groups is marginal.

Table 11: Ablation on L2EFC. Both variants use homomorphism-constrained transformations and Busemann attention. Accuracy (%) is reported with standard deviations as subscripts. Mini-ImageNet and 20News-Groups use $k = 20$.

Key transformation	Cora	PubMed	Mini-ImageNet	20News-Groups
FC + logmap	$85.74_{\pm 0.25}$	$83.14_{\pm 0.26}$	$87.71_{\pm 0.30}$	$69.30_{\pm 1.03}$
L2EFC	$\mathbf{86.00}_{\pm 0.30}$	$\mathbf{83.30}_{\pm 0.23}$	$\mathbf{88.11}_{\pm 0.14}$	$\mathbf{69.32}_{\pm 0.54}$

H More Implementation Details

H.1 Dataset Details

Graph datasets. We evaluate on seven graph benchmarks, including AIRPORT (Chami et al., 2019), CITESEER, CORA (Sen et al., 2008), PUBMED (Namata et al., 2012), ogbn-proteins, Amazon2M, and ogbn-arxiv (Hu et al., 2020). The citation networks CORA, CITESEER, and PUBMED are standard single-label node classification benchmarks with bag-of-words node features. We follow the

conventional public splits used in prior graph learning work. The AIRPORT graph models air transportation connectivity, where node features and labels characterize airport properties and traffic categories. For AIRPORT, we follow the split protocol used in prior hyperbolic graph learning studies (Chami et al., 2019). Each configuration is run with five random seeds, and we report mean and standard deviation. For large-scale evaluation, we use ogbn-proteins and ogbn-arxiv with the official OGB splits (Hu et al., 2020), together with Amazon2M. On ogbn-proteins, node features are obtained by aggregating edge-level attributes, and we report ROC-AUC following the official protocol. Amazon2M and ogbn-arxiv are evaluated by node classification accuracy.

Text and vision datasets. We evaluate semi-supervised classification on Mini-ImageNet and 20News-Groups. For Mini-ImageNet, we follow the standard protocol with 30 classes and 600 images per class. We use the same feature extractor as NodeFormer (Wu et al., 2022), namely a four-layer convolutional network followed by a fully connected layer, which produces 128-dimensional features. For 20News-Groups, we use fixed-dimensional bag-of-words-style features following prior graph Transformer studies (Wu et al., 2023; Yang et al., 2024). For both datasets, we use a fixed split ratio of 50% training, 25% validation, and 25% test. Because these datasets do not provide native graph structures, we construct a k-nearest-neighbor graph in feature space for message passing following Yang et al. (2024). We evaluate $k \in \{5, 10, 15, 20\}$ and symmetrize the resulting adjacency. We also report the graph-free variant where the backbone operates directly on the feature set with global attention as in (Wu et al., 2023).

H.2 ADDITIONAL VISION CLASSIFICATION EXPERIMENTS

Setup. We further evaluate vision classification via a drop-in replacement in LViT, following He et al. (2025). Concretely, we replace the self-attention module in each LViT block with our attention, while keeping the remaining architecture unchanged, including the patch embedding, MLP blocks, depth, hidden width, and number of heads. We also keep the same training pipeline as the LViT baseline, including data preprocessing and optimization settings, and only change the attention computation. We report results on CIFAR-10 and CIFAR-100.

Results. Table 4 shows that replacing the original self-attention in LViT with our hyperbolic attention yields consistent gains on both benchmarks. Specifically, the test accuracy improves from 85.02 to 86.18 on CIFAR-10 and from 69.11 to 70.36 on CIFAR-100. The improvement is observed under the same backbone, training schedule, and optimization pipeline, which suggests that the gain is attributable to the proposed attention mechanism rather than to auxiliary architectural changes. These results further indicate that our method can be integrated into a standard vision transformer backbone and remain effective beyond graph and text settings.

H.3 GRAPH EXPERIMENTS

Baselines. On all graph datasets, we compare HBformer with a broad spectrum of baselines that include classical Euclidean graph neural networks, recent graph Transformers, and hyperbolic message passing models. Specifically, we include multilayer perceptrons, GCN (Kipf & Welling, 2017), GAT (Velickovic et al., 2018), and SGC (Wu et al., 2019) as representative Euclidean GNNs, as well as sampling-based variants such as GCN-NSampler and GAT-NSampler, which are GCN (Kipf & Welling, 2017) and GAT (Velickovic et al., 2018) trained with neighbor sampling, and the SIGN (Frasca et al., 2020) model. For graph Transformers, we consider Graphormer (Ying et al., 2021), GraphTrans (Wu et al., 2021), GraphGPS (Rampášek et al., 2022), NodeFormer (Wu et al., 2022), SGFormer (Wu et al., 2023), and Hypformer (Yang et al., 2024). For hyperbolic architectures, we compare against HGCN (Chami et al., 2019), HGNN (Liu et al., 2019), HAN (Gulcehre et al., 2019), HNN++ (Shimizu et al., 2021), F-HNN (Chen et al., 2022), and FPS-T (Cho et al., 2023). All baselines are trained under the same data splits and evaluation metrics as HBformer. For methods with publicly available implementations, we adopt the authors' recommended hyperparameters or tune within the same ranges as for our model.

Large-scale implementation and additional analysis. Across all seven benchmarks in Table 2, HBformer achieves consistent improvements over both Euclidean and hyperbolic baselines, including Hypformer. On the AIRPORT dataset, where several strong methods already reach accuracies above ninety-five percent, HBformer improves the best previous result from 96.0 ± 0.6 to 96.72 ± 0.3 and outperforms Hypformer by more than one percentage point. On citation networks, HBformer

improves over Hypformer on all three datasets, with a one-point gain on CORA and a two-point gain on PUBMED, while maintaining a small but consistent advantage on CITESEER. On large-scale benchmarks, HBformer attains the best ROC-AUC on ogbn-proteins and the best accuracy on Amazon2M and ogbn-arxiv, improving Hypformer from 89.4 ± 0.3 to 91.55 ± 0.10 on Amazon2M while retaining low variance. For all three large-scale benchmarks, ogbn-proteins, Amazon2M, and ogbn-arxiv, we use HBformer with Busemann Linear Attention (BLA), as defined in Section 3.4. BLA avoids constructing an $N \times N$ attention matrix and requires $O(Nd^2)$ computation, which is linear in the number of nodes N for fixed feature dimension d. The uniformly low variance across datasets suggests that our design improves both average performance and training stability.

H.4 TEXT AND VISION EXPERIMENTS

Baselines and protocol. We compare HBformer against GCN (Kipf & Welling, 2017), GAT (Velickovic et al., 2018), DropEdge (Rong et al., 2020), IDGL (Chen et al., 2020), LDS (Franceschi et al., 2019), NodeFormer (Wu et al., 2022), SGFormer (Wu et al., 2023), Hypformer (Yang et al., 2024), and, where applicable, their variants without an explicit graph. Following Yang et al. (2024), we construct a k-nearest-neighbor graph on top of the input features and additionally evaluate a graph-free variant in which the backbone operates directly on the feature set with global attention as in (Wu et al., 2023). All models share the same feature representations and the same training, validation, and test splits. We report test accuracy for each choice of k, together with the mean and standard deviation over multiple random initializations, as summarized in Table 3.

Additional analysis. HBformer consistently outperforms all baselines on both Mini-ImageNet and 20News-Groups across all values of k. On Mini-ImageNet, HBformer improves over Hypformer by approximately 0.6 to 0.8 percentage points for all graph-based settings and also achieves the best performance without an explicit graph, while maintaining small standard deviations. The gains on 20News-Groups are more pronounced, where HBformer exceeds the best baseline by approximately 1.5 to 2.7 percentage points depending on k and achieves a clear improvement in the graph-free setting. This larger margin is consistent with the stronger latent hierarchy and higher sparsity of textual data, for which the Busemann similarity and intrinsic Lorentz aggregation provide a better inductive bias. Moreover, HBformer provides an intrinsic lift of scaled dot-product attention to the Lorentz model and recovers the classical Euclidean formulation in the flat limit $K \to 0^-$. Competing graph Transformers and hyperbolic models exhibit more noticeable performance fluctuations as k varies, whereas HBformer remains uniformly strong, which suggests that Busemann similarity together with the point-to-hyperplane key transformation yields representations that are less sensitive to the particular graph construction.