EXPLAINING HYPERBOLIC NEURAL NETWORKS VIA GEOMETRY-AWARE RELEVANCE PROPAGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic neural networks introduce geometric operations that require explicit treatment in relevance propagation. Equivalent geometric realizations can produce different feature attributions, even when local relevance is conserved. We study this problem through Geometric Representation Invariance (GRI), a specialization of Implementation Invariance, and zero-curvature consistency, which requires identity relevance propagation when a geometric module approaches the identity. We propose LRP-radial-all for origin-centered radial modules, treating geometric scaling as modulation and assigning relevance entirely to the signal branch. The rule conserves relevance, is invariant to equivalent radial factorizations, and satisfies zero-curvature consistency, yielding GRI for a specified Poincaré–Lorentz logarithmic-map construction. In contrast, a conservative LRP-half baseline can violate both consistency criteria. Experiments on hyperbolic MNIST, sEEG, and CIFAR-10 classifiers assess attribution fidelity, qualitative explanations, and runtime. LRP-radial-all achieves competitive attribution fidelity across datasets with runtime comparable to Gradient$\times$Input and substantially lower than Integrated Gradients. These findings motivate geometry-aware propagation rules that distinguish relevance conservation from consistency across equivalent computations.

1 INTRODUCTION

Explainable artificial intelligence (XAI) seeks to clarify how models arrive at their predictions (Gunning et al., 2019; Samek et al., 2021; Holzinger et al., 2020; Samek et al., 2019). However, most existing XAI methods have been developed primarily for conventional Euclidean neural networks. How these methods should be adapted to models whose computations are governed by non-Euclidean geometry remains much less understood.

This question becomes increasingly important for hyperbolic neural networks, which perform neural network operations in hyperbolic space and have attracted increasing attention (Peng et al., 2021; He et al., 2025). This interest is partly motivated by the mismatch between the exponential growth of tree-like hierarchies and the polynomial volume growth of Euclidean space. Hyperbolic space, by contrast, exhibits exponential volume growth with radius, making it well suited to representing hierarchical structures (Krioukov et al., 2010). Leveraging this geometric advantage, hyperbolic representations have shown promising performance across a range of tasks involving hierarchical structure, including computer vision (Ganea et al., 2018; Khrulkov et al., 2020), natural language processing (Mettes et al., 2024; Tifrea et al., 2019), foundation models (Desai et al., 2023; He et al., 2026), and biosignal analysis (Zhou et al., 2026; Li et al., 2026).

However, hyperbolic geometry poses distinct challenges for feature attribution: geometric maps couple feature coordinates, and equivalent representations can expose different computational factorizations of the same intrinsic function. An explanation method must therefore account for these operations while maintaining consistency across equivalent implementations. Layer-wise relevance propagation (LRP) offers a particularly suitable framework for studying this problem, as it redistributes a prediction backward through individual network modules using explicit local rules (Bach et al., 2015). This modular structure allows us to examine how relevance should propagate through hyperbolic operations and whether local conservation is compatible with Geometric Representation Invariance.

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.

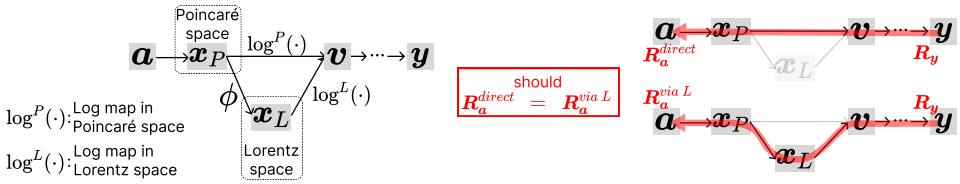


Figure 1: Geometric Representation Invariance (GRI) requires explanations at the same input after equivalent internal coordinate realizations to be the same.

In this work, we propose a relevance propagation framework for origin-centered hyperbolic operations. The origin provides a natural fixed reference point whose tangent space admits a Euclidean representation and is mathematically simple, facilitating efficient computation (Liu et al., 2019). We formulate LRP-radial-all rule for logarithmic maps and exponential maps, treating geometric scaling as modulation and assigning relevance to the signal branch. We study Geometric Representation Invariance (illustrated in Figure 1) as a specialization of Implementation Invariance (Sundararajan et al., 2017), prove consistency for the specified logarithmic-map construction, and show that a conservative LRP-half baseline can violate this property under equivalent radial factorizations. We evaluate the framework through controlled simulations and extensive experiments on various datasets and model architectures, including image and time series. Among the five equivalent hyperbolic models (Cannon et al., 1997), Poincaré and Lorentz are the most popular choices and are covered in this paper (Peng et al., 2021; He et al., 2025).

We make three main contributions.

- We propose *LRP-radial-all* for origin-centered hyperbolic modules, assigning relevance to the signal while treating geometric scaling as modulation.
- We specialize Implementation Invariance to *GRI* and introduce *zero-curvature consistency*. We prove conservation, radial factorization invariance, and zero-curvature consistency for radial-all, including GRI for the specified logarithmic-map test, and show that conservative LRP-half can violate both consistency criteria.
- We evaluate attribution fidelity and runtime on MNIST, sEEG, and CIFAR-10 hyperbolic classifiers, alongside controlled consistency tests.

2 Related Work

Hyperbolic neural networks. Early work on hyperbolic neural networks extended standard neural operations to hyperbolic geometry. Ganea et al. (2018) introduced hyperbolic formulations of multinomial logistic regression and feed-forward layers using the Poincaré ball. Subsequent work broadened the architectural scope: Shimizu et al. (2021) developed hyperbolic fully connected, convolutional, and attention layers, while Bdeir et al. (2024) introduced fully hyperbolic convolutional networks in the Lorentz model for computer vision. Beyond vision, hyperbolic architectures have also been applied to intracranial recordings. For example, Li et al. (2026) incorporated hyperbolic layers into an EEGNet-based model for working-memory load classification from sEEG signals. Together, these developments have made geometry-specific operations such as exponential maps, logarithmic maps, and manifold-valued feature transformations fundamental building blocks of modern hyperbolic architectures.

Explainability for HNN. In contrast to the rapid development of hyperbolic architectures, their explainability has received comparatively limited attention. Most feature-attribution methods were originally formulated for models operating in Euclidean spaces, and therefore do not explicitly account for the geometric operations introduced by HNNs. Recent work has begun to incorporate non-Euclidean geometry into explainability. Manifold Integrated Gradients (MIG) (Zaher et al., 2024) replaces the straight-line path of Integrated Gradients with geodesic paths on a learned Riemannian data manifold, thereby aligning attribution with the intrinsic geometry of the data. Diffeomorphic counterfactual methods use generative models to construct coordinate systems in which gradient-based search produces semantically meaningful changes to the input (Dombrowski et al., 2024). Such approaches demonstrate the relevance of geometry to explainability, but focus primarily on the

geometry of the data manifold, rather than on propagating explanations through the internal geometric operations of a hyperbolic neural network. Our work addresses the complementary problem. We study how relevance should be propagated through the geometric primitives of HNNs themselves, with particular attention to logarithmic and exponential maps and transformations between equivalent hyperbolic representations.

Layer-wise relevance propagation (LRP). Layer-wise relevance propagation (LRP) (Bach et al., 2015) decomposes a model prediction into input relevance scores through architecture-specific backward redistribution rules, typically at a computational cost comparable to a backward computation. Deep Taylor Decomposition provides a theoretical foundation for certain LRP rules through local Taylor expansions of relevance functions (Montavon et al., 2017; Samek et al., 2021). Our work is particularly related to the rules for multiplicative interactions. For gated recurrent networks, Arras et al. (2017) assign all relevance to the source branch and none to the gate, interpreting the latter as a modulation of the signal. Similarly, Ali et al. (2022) treat attention weights as fixed during relevance propagation and route relevance through the value branch, following an LRP-all-style allocation for the attention-value product. Alternatively, LRP-half for LSTM (Arras et al., 2019) propagates half of the relevance to weight and half to signal. AttnLRP (Achtibat et al., 2024) derives a uniform redistribution rule for multiplicative interactions. These approaches motivate different treatments of hyperbolic radial maps.

3 Geometry-aware Criteria

3.1 Hyperbolic Neural Networks

Hyperbolic neural networks (HNNs) generalize neural operations to hyperbolic space (Ganea et al., 2018). We describe a representative architecture in the Poincaré ball $\mathbb{D}_c^d = \{\boldsymbol{x} \in \mathbb{R}^d : c\|\boldsymbol{x}\|_2^2 < 1\}$ with curvature $-c$, where $c > 0$. An input feature vector $\boldsymbol{a} \in \mathbb{R}^d$ is first interpreted as an origin-tangent vector and mapped to the ball: $\boldsymbol{h}^{(0)} = \exp_{\mathbf{0}}^c(\boldsymbol{a})$. Each hidden layer performs a hyperbolic linear transformation, bias addition, and activation:

$$\boldsymbol{m}^{(\ell)} = \exp_{\mathbf{0}}^c(\boldsymbol{W}^{(\ell)} \log_{\mathbf{0}}^c(\boldsymbol{h}^{(\ell-1)})) \oplus_c \exp_{\mathbf{0}}^c(\boldsymbol{b}^{(\ell)}), \tag{1}$$

$$\boldsymbol{h}^{(\ell)} = \exp_{\mathbf{0}}^c(\sigma(\log_{\mathbf{0}}^c(\boldsymbol{m}^{(\ell)}))), \tag{2}$$

where $\boldsymbol{W}^{(\ell)}$ and $\boldsymbol{b}^{(\ell)}$ are trainable parameters, $\oplus_c$ denotes Möbius addition (definition see Eq.65) and σ is an element-wise activation applied in the tangent space. We show the equations of $\exp_{\mathbf{0}}^c(\cdot)$ and $\log_{\mathbf{0}}^c(\cdot)$ in Figure 2. For a tangent-space readout, the final representation is mapped back to Euclidean coordinates and passed to a prediction head g: $\hat{\boldsymbol{y}} = g\big(\log_{\mathbf{0}}^c(\boldsymbol{h}^{(L)})\big)$. Appendix A summarizes the Lorentz model and its correspondence with the Poincaré representation.

3.2 Geometric Representation Invariance

The Poincaré ball and Lorentz hyperboloid are isometric representations of the same hyperbolic geometry. By transforming intermediate representations and the associated operations consistently, a model can be expressed in either space without changing its input-output function. Ideally, explanations in the original input space should also remain unchanged under such a conversion. This requirement is motivated by Implementation Invariance (Sundararajan et al., 2017), which states that functionally equivalent models should yield identical input attributions. We specialize this principle to equivalent hyperbolic representations and refer to the resulting criterion as *Geometric Representation Invariance* (GRI).

Definition 1 (GRI as a restricted Implementation Invariance criterion)**.** *Let $\boldsymbol{a} \in \mathbb{R}^p$ be the original feature vector. Let F_P and F_L implement the same scalar target function on a common input domain, with internal hyperbolic computations related by the specified isometries. For a fixed explanation procedure, GRI requires*

$$\mathcal{E}(F_P, \boldsymbol{a}) = \mathcal{E}(F_L, \boldsymbol{a}), \tag{3}$$

where $\mathcal{E}$ denotes an explanation method.

For gradient-based methods defined solely through evaluations of the input-output function such as Gradient×Input (Ancona et al., 2018) and Integrated Gradients (Sundararajan et al., 2017), GRI

follows directly from functional equivalence. For propagation-based methods such as LRP, however, GRI is non-trivial because relevance redistribution depends on intermediate computations and local propagation rules. Equivalent geometric representations may expose different computational factorizations, so local relevance conservation alone does not guarantee identical input attributions.

A controlled test of GRI. Figure 1 illustrates a controlled test of whether relevance propagation depends on the internal representation of a hyperbolic logarithmic map. Starting from the same original input $\boldsymbol{a}$, a shared encoder produces a Poincaré representation $\boldsymbol{x}_P \in \mathbb{D}_c^d$. We compare two computational paths: the direct path applies the Poincaré logarithmic map, whereas the alternative path first converts $\boldsymbol{x}_P$ to the corresponding Lorentz point $\boldsymbol{x}_L = \phi(\boldsymbol{x}_P)$ and then applies the Lorentz logarithmic map.

After aligning the tangent coordinates, both paths produce the same vector $\boldsymbol{v} \in \mathbb{R}^d$:

$$\boldsymbol{v} = \alpha_P(\boldsymbol{x}_P)\boldsymbol{x}_P = \alpha_L(\boldsymbol{x}_L)\bar{\boldsymbol{x}}_L, \qquad \bar{\boldsymbol{x}}_L = \gamma(\boldsymbol{x}_P)\boldsymbol{x}_P. \tag{4}$$

Here $\boldsymbol{x}_L = (x_{L,0}, \bar{\boldsymbol{x}}_L)$, and α_L includes the factor $1/2$ that converts the ambient Lorentz tangent vector $(0, 2\boldsymbol{v})$ to the common coordinates $\boldsymbol{v}$. The coordinate conversion and aligned logarithmic maps are derived in Appendix D. Both paths then use the same downstream computation $y = G(\boldsymbol{v})$, so they implement the same input-output function.

To test GRI, we initialize relevance at the same scalar output y and propagate it backward along each path, as indicated by the red arrows in Figure 1. The downstream propagation is identical and therefore gives the same relevance $\boldsymbol{R}_{\boldsymbol{v}}$ to both realizations. The direct path propagates this relevance through the Poincaré logarithmic map and the alternative path propagates it through the aligned Lorentz logarithmic map and the coordinate conversion ϕ. Using the same relevance rule through the shared encoder, we compare the resulting explanations at the original input:

$$\boldsymbol{R}_{\boldsymbol{a}}^{\text{direct}} \stackrel{?}{=} \boldsymbol{R}_{\boldsymbol{a}}^{\text{via } L}. \tag{5}$$

This construction isolates the effect of an equivalent geometric realization while keeping the input, prediction, and surrounding computation unchanged.

Appendix E presents a controlled exponential-map test with a shared Lorentz downstream network and an explicitly matched treatment of time-coordinate relevance.

3.3 Zero-curvature consistency

A natural consistency requirement arises when hyperbolic operations approach their Euclidean counterparts as curvature vanishes (Ganea et al., 2018). In particular, the origin-centered Poincaré exponential and logarithmic maps converge to the identity in the coordinate convention used here. Once such a geometric module becomes an identity transformation, it should no longer redistribute relevance among input coordinates. This motivates requiring its backward relevance rule to approach identity propagation as well. We formalize this local requirement as zero-curvature consistency.

Definition 2 (Zero-curvature consistency)**.** *Let f_c be a family of geometric modules expressed in fixed coordinates such that $f_c(\boldsymbol{x}) \to k\boldsymbol{x}$ as $c \to 0$, where $k > 0$ is fixed and independent of $\boldsymbol{x}$. We use identity relevance propagation for the limiting constant scaling, treating k as a fixed modulation factor. Let $\mathcal{R}_c(\boldsymbol{x}, \boldsymbol{R})$ denote the input relevance obtained by propagating a fixed incoming relevance vector $\boldsymbol{R}$ through f_c. The propagation rule is zero-curvature consistent if*

$$\lim_{c \to 0} \mathcal{R}_c(\boldsymbol{x}, \boldsymbol{R}) = \boldsymbol{R} \tag{6}$$

for every input $\boldsymbol{x}$ and incoming relevance $\boldsymbol{R}$.

This definition includes both identity limits and fixed coordinate scalings. The origin Poincaré maps approach the identity, whereas the aligned Lorentz spatial logarithmic and exponential maps approach scaling by $1/2$ and 2, respectively. These constant factors reflect the tangent-coordinate convention and receive no separate relevance.

4 Relevance Propagation Rules

In this section, we define relevance propagation rules for origin-centered hyperbolic modules. We introduce LRP-radial-all and contrast it with a conservative LRP-half baseline, showing that conservation alone does not ensure consistency across equivalent radial factorizations. For tangent-space

linear layers, we use standard Euclidean LRP rules and their propagation formulas and conservation conditions are summarized in Appendix H. For Möbius bias addition, we adopt a separate signal-only convention (see Appendix G).

4.1 Consistent relevance propagation through geometric modulation

Conservation alone does not determine how relevance should be allocated between multiplicative branches. Symmetric splitting can depend on the grouping of multiplication operations, while grouping-invariant symmetric rules require additional assumptions and may have restricted domains (Appendix B). For our geometric modules, we instead distinguish the feature signal from its scalar modulation and adopt an all-to-signal allocation.

Geometry as a modulation branch. For a radial map $\boldsymbol{y} = \alpha(\boldsymbol{x})\boldsymbol{x}$, the two multiplicative branches have different roles: $\boldsymbol{x}$ carries the feature coordinates, while $\alpha(\boldsymbol{x})$ applies a shared geometric scaling. We adopt a signal-based attribution convention that preserves the incoming feature-wise relevance allocation across this modulation. This choice is motivated by two properties. First, splitting the geometric scaling into successive modulation factors should not change the relevance propagated along the designated signal path. Second, when an origin map approaches the identity in the zero-curvature limit, its propagation rule should approach identity propagation as well.

These requirements motivate an all-to-signal allocation, consistent with earlier treatments of source-gate interactions (Arras et al., 2017) and attention-value products (Ali et al., 2022).

Definition 3 (LRP-radial-all)**.** *For a scalar-signal module $\boldsymbol{y} = \alpha(\boldsymbol{x})\boldsymbol{s}$, LRP-radial-all assigns all incoming relevance to the designated signal $\boldsymbol{s}$ and none to the modulation factor:*

$$R_\alpha = 0, \qquad \boldsymbol{R}_{\boldsymbol{s}} = \boldsymbol{R}_{\boldsymbol{y}}. \tag{7}$$

For a radial map of a single vector, $\boldsymbol{s} = \boldsymbol{x}$ and α depends only on its radius.

Proposition 1 (Conservation and signal-path factorization invariance)**.** *Consider functionally equivalent scalar-signal realizations that differ only in the splitting or grouping of scalar modulation factors along a designated signal path. Suppose the signal dimension and coordinate order are preserved, and the incoming relevance is identical. Applying the all-to-signal rule at every multiplication yields $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{y}}$, independently of the number or grouping of modulation factors, and conserves signed total relevance.*

Proof. Each multiplication has backward operator I_d on its designated signal and assigns zero relevance to its modulation branch. A chain of $m \geq 1$ such modules therefore gives $\boldsymbol{R}_{\boldsymbol{x}} = I_d^m \boldsymbol{R}_{\boldsymbol{y}} = \boldsymbol{R}_{\boldsymbol{y}}$, and consequently $\mathbf{1}^\top \boldsymbol{R}_{\boldsymbol{x}} = \mathbf{1}^\top \boldsymbol{R}_{\boldsymbol{y}}$. □

The proposition provides the local consistency used in our controlled GRI test. More generally, necessary conditions for GRI and zero-curvature consistency can be derived (see Appendix C) for the fixed-proportion radial rules

$$R_{x_i} = \eta R_{y_i} + (1-\eta)\frac{x_i^2}{\|\boldsymbol{x}\|_2^2}\sum_j R_{y_j}, \qquad \eta \in [0,1], \tag{8}$$

where the same constant η is used at every module and the modulation relevance is redistributed through the squared norm. This family includes LRP-half at $\eta = 1/2$ and LRP-radial-all at $\eta = 1$. For $d \geq 2$, invariance under equivalent nonzero radial factorizations permits only $\eta = 0$ or $\eta = 1$. Requiring zero-curvature consistency as defined in Definition 2 further selects $\eta = 1$. Thus, within this specified family, radial-all is the unique rule satisfying both criteria.

4.2 LRP-half as an alternative

Motivated by equal-split rules for multiplicative interactions (Arras et al., 2019; Achtibat et al., 2024), we consider an LRP-half baseline that assigns half of the relevance to the signal and half to the radial factor, redistributing the latter through the squared norm (Figures 2 and 7). In the logarithmic-map test of Eq. 4, the direct Poincaré path contains one radial module, whereas the equivalent Lorentz path contains two. For $\boldsymbol{x}_P \neq \mathbf{0}$, their backward allocations are

$$R^P_{x_{P,i}} = \tfrac{1}{2}R_{v_i} + \tfrac{1}{2}\frac{x_{P,i}^2}{\|\boldsymbol{x}_P\|_2^2}\sum_j R_{v_j}, \qquad R^L_{x_{P,i}} = \tfrac{1}{4}R_{v_i} + \tfrac{3}{4}\frac{x_{P,i}^2}{\|\boldsymbol{x}_P\|_2^2}\sum_j R_{v_j}. \tag{9}$$

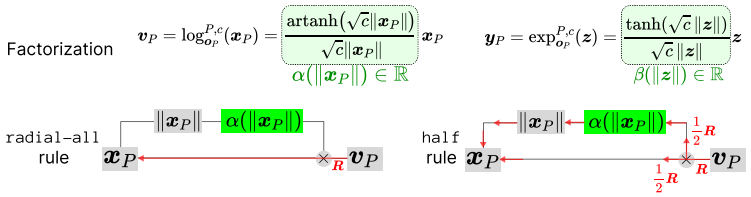


Figure 2: Top: factorization of log and exp map with origin point in Poincaré space. Bottom: our proposed radial-all rule and half rule for LRP applied to such layer.

Both conserve the incoming total relevance, but generally disagree coordinate-wise. With an identity encoder, this difference directly violates GRI at the common original input. Appendix F provides the derivation and an explicit counterexample. Appendix I reports a numerical test.

Failure of zero-curvature consistency. Although $\log_{\mathbf{0}}^{P,c}(\boldsymbol{x}_P) \to \boldsymbol{x}_P$ as $c \to 0$, Eq. 9 is independent of curvature for fixed input and incoming relevance, and generally differs from identity propagation. Thus, LRP-half also fails zero-curvature consistency and the same argument also applies to the origin exponential map. LRP-radial-all, by contrast, has identity backward action for every curvature. Conservation alone therefore guarantees neither of the two consistency criteria.

5 Experiments

We compare attribution methods on hyperbolic classifiers for MNIST, sEEG, and CIFAR-10 using sparsity-fidelity metrics and qualitative visualizations. LRP-radial-all achieves competitive attribution fidelity. Runtime comparisons further show that LRP-radial-all is comparable in efficiency to Gradient×Input and substantially faster than Integrated Gradients. We also present a controlled GRI numerical test in Appendix I.

5.1 MNIST

Model and training. We use a Poincaré hyperbolic MLP with dimensions 784-128-10, fixed curvature -1, Möbius biases, and a tangent-space ReLU between the two hyperbolic linear layers. Images have pixel values in $[0, 1]$ and flattened inputs are scaled by 0.1 and mapped to the manifold using the origin exponential map. An origin logarithmic map produces the output logits. The training configuration uses 55000 training and 5000 validation images, Adam with learning rate 10^{-3}, batch size 128 and cross-entropy loss. Training runs for 100 epochs, with checkpoint selection by validation accuracy. The model has 97.44% test accuracy.

Explanation methods. All quantitative comparisons explain the original predicted-class logit of the model. We compare radial-all and the half rule for hyperbolic mapping layers using identical linear LRP-γ rules ($\gamma = 0.25$, $\epsilon = 10^{-9}$) for other layers. Fixed bias branches receive zero relevance. Baselines are Gradient×Input, Integrated Gradients (IG), Patch Occlusion, and random. IG uses a black-image baseline and 512 midpoint integration steps along a straight path in pixel space. Patch Occlusion uses black 4×4 windows with stride 4, assigning the target-logit drop to each pixel in the window.

Sparsity-fidelity protocol. We evaluate randomly sampled 1000 images from test set. In one image, pixels are ranked by their signed attribution scores in descending order. Let S_k contain the top k pixels and let m_{S_k} be their binary pixel mask. For the fixed original predicted class t, we define

$$\text{Sparsity}(k) = 1 - k/K, \qquad s_t(x) = \text{softmax}(f(x))_t, \tag{10}$$

$$\text{Fidelity}^+(k) = s_t(x) - s_t(x \odot (1 - m_{S_k})), \qquad \text{Fidelity}^-(k) = s_t(x) - s_t(x \odot m_{S_k}), \tag{11}$$

where K means the total number of pixels. Higher Fidelity$^+$ indicates greater necessity of selected regions and lower Fidelity$^-$ indicates greater sufficiency. We use 50 regions in sparsity, and compute per-image AUC by integration over actual sparsity values. Random rankings are averaged over 10 repetitions per image. We report mean AUCs and percentile 95% confidence intervals from 1000 image-level bootstrap resamples.

Table 1: Sparsity-fidelity AUC on MNIST, sEEG and CIFAR-10. F^+ and F^- denote Fidelity$^+$ and Fidelity$^-$ AUC. Subscripts report $\pm h$, where $h = \max(\mu - L, U - \mu)$ for the 95% bootstrap interval $[L, U]$. Bold indicates the best mean within each dataset and metric.

	MNIST		sEEG		CIFAR-10	
Method	$F^+ \uparrow$	$F^- \downarrow$	$F^+ \uparrow$	$F^- \downarrow$	$F^+ \uparrow$	$F^- \downarrow$
$\text{LRP}_{\text{radial-all}}$	$\mathbf{0.670}_{\pm 0.005}$	$\mathbf{-0.003}_{\pm 0.005}$	$\mathbf{0.313}_{\pm 0.096}$	$-0.111_{\pm 0.017}$	$0.637_{\pm 0.020}$	$\mathbf{0.277}_{\pm 0.019}$
LRP_{half}	$0.562_{\pm 0.011}$	$0.063_{\pm 0.008}$	$0.310_{\pm 0.095}$	$-0.106_{\pm 0.017}$	$0.588_{\pm 0.022}$	$0.505_{\pm 0.022}$
Gradient×Input	$0.410_{\pm 0.017}$	$0.202_{\pm 0.014}$	$0.303_{\pm 0.096}$	$\mathbf{-0.113}_{\pm 0.017}$	$\mathbf{0.703}_{\pm 0.020}$	$0.547_{\pm 0.023}$
Integrated Grad.	$0.664_{\pm 0.005}$	$0.000_{\pm 0.005}$	$0.309_{\pm 0.096}$	$\mathbf{-0.113}_{\pm 0.017}$	$0.683_{\pm 0.020}$	$0.495_{\pm 0.022}$
Patch Occlusion	$0.504_{\pm 0.013}$	$0.067_{\pm 0.007}$	-	-	-	-
Random	$0.193_{\pm 0.005}$	$0.193_{\pm 0.005}$	$0.057_{\pm 0.023}$	$0.058_{\pm 0.022}$	$0.664_{\pm 0.022}$	$0.663_{\pm 0.022}$

Table 2: Runtime comparison of explanation methods. We average the runtime on 3 runs.

Dataset	LRP-radial-all	Gradient×Input	Occlusion	Integrated Gradients
MNIST (ms/image)	0.505	0.960	12.676	41.511
sEEG (ms/sample)	5.236	3.813	-	94.431
CIFAR-10 (ms/image)	15.688	12.730	-	66.375

Results. As shown in Table 1 and Figure 9, LRP-radial-all achieves the highest Fidelity$^+$ AUC and the lowest Fidelity$^-$ AUC among the evaluated methods. Although the improvement from IG is marginal, LRP-radial-all is substantially faster than IG as in Table 2. Figure 3 shows two example heatmaps by our method and baselines. LRP-radial-all highlights the important curves and intersections in the number pictures. Integrated Gradients can capture the important parts as well, but the heatmaps appear noisier.

5.2 sEEG

sEEG captures interactions across electrodes and temporal scales, giving rise to hierarchical structure that can be naturally represented in hyperbolic space. Recent hyperbolic approaches to sEEG modeling have achieved state-of-the-art performance and substantially outperformed their Euclidean counterparts (Guillemaud et al., 2025; Li et al., 2026). Following Li et al. (2026), we evaluate on the working-memory sEEG dataset introduced by Boran et al. (2020). The dataset contains 9 subjects and many trials for each subject. Each trial is represented as a multichannel time series $\mathbf{X} \in \mathbb{R}^{C \times T}$, where C denotes the number of electrodes and T the number of temporal samples, and is associated with a binary label indicating whether the working-memory load is high or low. Following Li et al. (2026), we trained a variant of EEGNet in which the multi-layer perceptron is replaced with its hyperbolic counterpart. On each subject we trained the model using 5-fold cross-validation.

Qualitative results. We take subject 2 as an example and plot the signals and LRP-radial-all heatmaps averaged on all samples of it by the label in Figure 4. In the low-load condition, a prominent positive relevance region is concentrated within a relatively limited subset of channels, indicating a localized contribution to the model prediction. In contrast, the high-load condition exhibits more distributed relevance across multiple channels during the middle portion of the time window. Stronger relevance concentrations also emerge toward the later portion of the high-load window, indicating a greater contribution from these temporal segments. We validate the importance of the relevant time windows to the prediction in Appendix J. In addition, matched upper-channel regions show opposite relevance polarities between the two conditions, indicating a class-dependent difference in how the model uses the same spatiotemporal region. The annotated boxes highlight these representative attribution patterns. The more distributed relevance observed under the high-load condition is consistent with recent intracranial evidence (Yang et al., 2025) showing increased inter-regional information sharing as working-memory demands increase.

Applying DFT-LRP (Vielhaben et al., 2024), we propagate relevance to the frequency space and visualize the relevance in Figure 5. The results show that frequency-domain relevance is mainly concentrated in the lower-frequency range, particularly the delta band (0.5 - 4 Hz). Besides, the

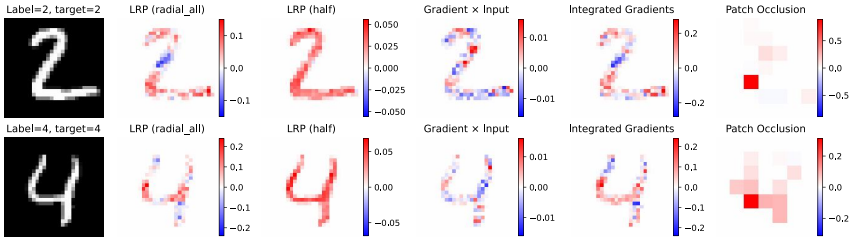


Figure 3: Heatmaps on images of MNIST.

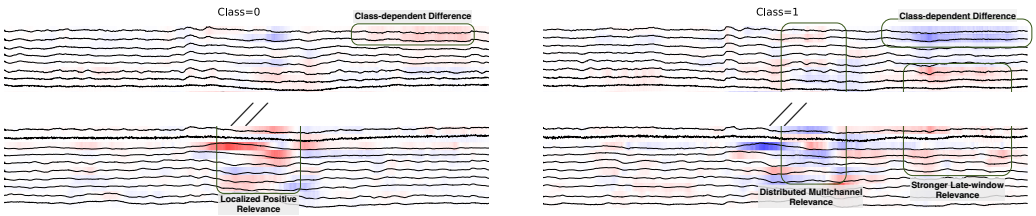


Figure 4: Averaged signals and relevance heatmaps for class 0 (low working-memory load) and class 1 (high working-memory load). Boxes highlight localized relevance, distributed multichannel relevance, stronger late-window relevance, and class-dependent attribution differences. Each time series stands for a sensor. Time range is 3-second with 3000 points. Red and blue indicate positive and negative relevance for the explained class score, respectively. The full figure is in Figure 11.

magnitude of delta-band relevance differs between the two working-memory load classes. These observations suggest the classifier mainly exploits low-frequency activity, with such activity contributing differently to the decision of the model under two working-memory load conditions.

Quantitative evaluation. We evaluate sensor-level attribution fidelity using all trials from all subjects in the dataset. We follow the setting in (Li et al., 2026) and use subject-specific HEEGNet0 checkpoints from stratified five-fold within-subject cross-validation. For an input $x \in \mathbb{R}^{C\times T}$, we fix the target class to the model's prediction and compute attributions for the target logit. The relevance of sensor c is $R_c = \sum_{t=1}^{T} R_{c,t}$. Sensors are ranked in descending order of R_c. We compare LRP-radial-all, LRP-half, Gradient×Input, Integrated Gradients (IG), and random rankings. Both LRP variants use the ϵ-rule with $\epsilon = 10^{-9}$ for layers apart from hyperbolic layers. IG uses a zero baseline and integration over 128 steps. The random baseline averages five random rankings.

We replace perturbed sensors with the timepoint-wise mean across all sensors in the original trial. We perturb the input by sparsity from 0% to 100% in 5% increments, and calculate the fidelity as defined in Eq.10,11, as well as the area under the sparsity-fidelity curves. We average trial-level AUCs across the held-out folds within each subject and then average the nine subject means, and the results are summarized in Table 1. LRP-radial-all achieves the best average AUC of Fidelity$^+$ and comparable AUC of Fidelity$^-$ to the best baseline.

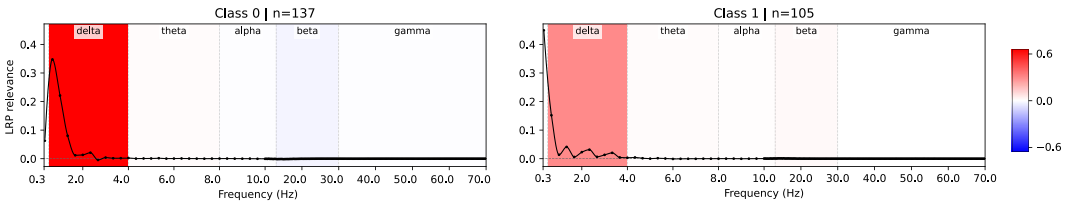


Figure 5: Relevance of sEEG prediction by class in frequency space. The relevance score is averaged on all trials of subject 2 in the dataset. We grouped the frequency-domain relevance into conventional electrophysiological frequency bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–70 Hz).

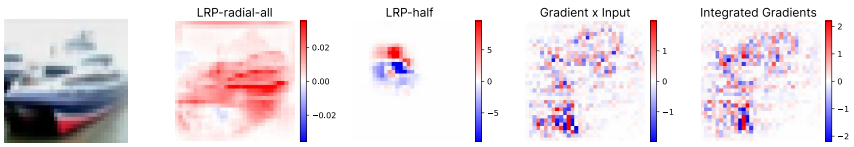


Figure 6: Qualitative attribution comparison on a correctly classified CIFAR-10 image of ship. Columns show the input, heatmaps for LRP-radial-all, LRP-half, Gradient×Input, and Integrated Gradients. RGB-channel attributions are summed, with red and blue indicating positive and negative values, respectively. More examples in Figure 10.

5.3 CIFAR-10

We evaluate our relevance propagation method on CIFAR-10 using a lightweight classifier with six Lorentz convolutional layers inspired by Bdeir et al. (2024). The model architecture and training details are in Appendix K. We assess the resulting explanations through qualitative heatmaps and quantitative fidelity and sparsity metrics.

Qualitative results. Figure 6 compares explanations for a CIFAR-10 image of ship. With the same $z^{\mathcal{B}}$-γ-ϵ configuration (Montavon et al., 2017; Samek et al., 2019) for layers apart from hyperbolic layers, LRP-radial-all produces broader, spatially coherent attribution patterns that overlap visible object regions, whereas LRP-half concentrates attribution into smaller regions with positive and negative contributions. Gradient×Input and Integrated Gradients exhibit much noisier patterns. These examples illustrate how the radial propagation rule affects attribution structure.

Quantitative evaluation. We evaluate attribution fidelity on 512 randomly sampled CIFAR-10 test images using pixel-wise mean-color replacement and predicted-class probabilities. As shown in Table 1, LRP-radial-all achieves the lowest Fidelity$^-$ AUC, outperforming other methods, indicating stronger prediction preservation when retaining highly ranked pixels. However, its Fidelity$^+$ AUC falls even below Random. These results indicate stronger retention sufficiency but weaker deletion sensitivity under this perturbation protocol.

5.4 Runtime evaluation

We compare attribution runtimes for MNIST on an AMD EPYC 7453, sEEG and CIFAR-10 on a NVIDIA A100. As in Table 2, LRP-radial-all incurs substantially lower computational cost than Integrated Gradients, and comparable with Gradient×Input that only includes one forward and backward computation of the model.

6 Conclusion

We developed a relevance propagation framework for origin-centered hyperbolic operations, guided by Geometric Representation Invariance and zero-curvature consistency. Our analysis shows that relevance conservation alone does not ensure either property, as the specified LRP-half baseline dependent on equivalent radial factorizations fails to approach identity propagation as curvature vanishes. By treating geometric scaling as modulation, LRP-radial-all conserves relevance locally, preserves the incoming allocation across radial factorizations, and satisfies zero-curvature consistency. These properties guarantee GRI for the specified Poincaré–Lorentz logarithmic-map construction. Experiments on MNIST, sEEG, and CIFAR-10 demonstrate competitive attribution fidelity with runtime comparable to Gradient×Input and substantially lower than Integrated Gradients, while revealing a trade-off between retention and deletion performance on CIFAR-10.

Limitations and future work. Our theoretical guarantees apply to the specified module realizations and propagation conventions. The controlled exponential map test additionally requires a treatment of time-coordinate relevance, and invariance under alternative treatments remains an open question. Future work includes extending the framework to non-origin base points and structured explanations of message-passing paths and higher-order interactions as in Xiong et al. (2026).

AI USE STATEMENT

We used generative AI tools to assist with language polishing, identifying relevant literature, providing feedback on research methodology and experimental design, and implementing methods. We also used these tools as discussion and editing aids for refining theoretical arguments and proofs. We did not use generative AI to interpret experimental results, generate synthetic datasets or to clean or reformat datasets. The authors reviewed all AI-assisted content and verified cited references against their original sources. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 5 describes the evaluation protocols, attribution baselines, and experimental settings. The appendices provide mathematical derivations and proofs, relevance propagation rules, and additional experimental details, including the CIFAR-10 architecture and training configuration in Appendix K. The complete codebase will be released upon acceptance.

ACKNOWLEDGMENTS

REFERENCES

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers. In *ICML*, pp. 135–168. PMLR / OpenReview.net, 2024.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162, pp. 435–451. PMLR, 2022.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *ICLR (Poster)*. OpenReview.net, 2018.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *WASSA@EMNLP*, pp. 159–168. Association for Computational Linguistics, 2017.

Leila Arras, Jose A. Arjona-Medina, Michael Widrich, Grégoire Montavon, Michael Gillhofer, Klaus-Robert Müller, Sepp Hochreiter, and Wojciech Samek. Explaining and interpreting LSTMs. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700, pp. 211–238. Springer, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

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

Ece Boran, Tommaso Fedele, Adrian Steiner, Peter Hilfiker, Lennart Stieglitz, Thomas Grunwald, and Johannes Sarnthein. Dataset of human medial temporal lobe neurons, scalp and intracranial eeg during a verbal working memory task. *Scientific data*, 7(1):30, 2020. doi: 10.1038/s41597-020-0364-3.

James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Ann-Kathrin Dombrowski, Jan E. Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic counterfactuals with generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5):3257–3274, 2024.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *NeurIPS*, pp. 5350–5360, 2018.

Martin Guillemaud, Louis Cousyn, Vincent Navarro, and Mario Chavez. Hyperbolic embedding of brain networks as a tool for epileptic seizures forecasting. *Physical Review Research*, 7(2): 023182, 2025.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. Xai—explainable artificial intelligence. *Science Robotics*, 4(37), 2019.

Neil He, Hiren Madhu, Ngoc Bui, Menglin Yang, and Rex Ying. Hyperbolic deep learning for foundation models: A survey. In *KDD (2)*, pp. 6021–6031. ACM, 2025.

Neil He, Rishabh Anand, Hiren Madhu, Ali Maatouk, Smita Krishnaswamy, Leandros Tassiulas, Menglin Yang, and Rex Ying. Helm: Hyperbolic large language models via mixture-of-curvature experts. *Advances in Neural Information Processing Systems*, 38:142604–142635, 2026.

Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable AI methods - A brief overview. In *xxAI@ICML*, pp. 13–38. Springer, 2020.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6417–6427. IEEE, 2020.

Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(3):036106, 2010. doi: 10.1103/PhysRevE.82.036106.

Shanglin Li, Chu Shiwen, Okan Koç, Yi Ding, Qibin Zhao, Motoaki Kawanabe, and Ziheng Chen. HEEGNet: Hyperbolic embeddings for EEG. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CNDNRjpVIL.

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

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132 (9):3484–3508, 2024. doi: 10.1007/s11263-024-02043-5.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.*, 65:211–222, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3776–3785. PMLR, 2018.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(12):10023–10044, 2021. doi: 10.1109/TPAMI.2021.3136921.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller (eds.). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*. Springer, 2019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE*, 109(3):247–278, 2021.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *ICLR (Poster)*. OpenReview.net, 2019.

Johanna Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek. Explainable AI for time series via virtual inspection layers. *Pattern Recognit.*, 150:110309, 2024.

Ping Xiong, Thomas Schnake, Grégoire Montavon, Klaus-Robert Müller, and Shinichi Nakajima. Normalized relevance measure as a unifying framework to explain neural network latent structures. *CoRR*, abs/2606.00557, 2026.

Jiayi Yang, Dan Cao, Chunyan Guo, Lennart Stieglitz, Debora Ledergerber, Johannes Sarnthein, and Jin Li. Enhanced role of the entorhinal cortex in adapting to increased working memory load. *Nature Communications*, 16(1):5798, Jul 2025. ISSN 2041-1723.

Eslam Zaher, Maciej Trzaskowski, Quan Nguyen, and Fred Roosta. Manifold integrated gradients: Riemannian geometry for feature attribution. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58090–58104. PMLR / OpenReview.net, 2024.

Runhe Zhou, Shanglin Li, Guanxiang Huang, Xinliang Zhou, Qibin Zhao, Motoaki Kawanabe, Yi Ding, and Cuntai Guan. Eeg-based multimodal learning via hyperbolic mixture-of-curvature experts. *International Conference on Machine Learning*, 2026.

A Poincaré and Lorentz Representations

For curvature $-c$ with $c > 0$, the Poincaré ball and Lorentz hyperboloid are

$$\mathbb{D}_c^d = \{\boldsymbol{x}_P \in \mathbb{R}^d : c\|\boldsymbol{x}_P\|_2^2 < 1\}, \tag{12}$$

$$\mathbb{L}_c^d = \{(x_{L,0}, \bar{\boldsymbol{x}}_L) \in \mathbb{R}^{d+1} : -x_{L,0}^2 + \|\bar{\boldsymbol{x}}_L\|_2^2 = -1/c,\ x_{L,0} > 0\}. \tag{13}$$

Their origins are $\boldsymbol{o}_P = \mathbf{0}$ and $\boldsymbol{o}_L = (c^{-1/2}, \mathbf{0})$. The standard Poincaré-Lorentz isometry and its inverse (Nickel & Kiela, 2018), rescaled here to curvature $-c$, are

$$\phi(\boldsymbol{x}_P) = \left(\frac{1 + c\|\boldsymbol{x}_P\|_2^2}{\sqrt{c}(1 - c\|\boldsymbol{x}_P\|_2^2)}, \frac{2\boldsymbol{x}_P}{1 - c\|\boldsymbol{x}_P\|_2^2}\right), \tag{14}$$

$$\phi^{-1}(\boldsymbol{x}_L) = \frac{\bar{\boldsymbol{x}}_L}{\sqrt{c}\, x_{L,0} + 1}. \tag{15}$$

Since $x_{L,0} = \sqrt{c^{-1} + \|\bar{\boldsymbol{x}}_L\|_2^2}$, both spatial conversions are radial:

$$\bar{\boldsymbol{x}}_L = \gamma(\boldsymbol{x}_P)\boldsymbol{x}_P, \qquad \gamma(\boldsymbol{x}_P) = \frac{2}{1 - c\|\boldsymbol{x}_P\|_2^2}, \qquad \boldsymbol{x}_P = \frac{\bar{\boldsymbol{x}}_L}{\sqrt{1 + c\|\bar{\boldsymbol{x}}_L\|_2^2} + 1}. \tag{16}$$

B Multiplicative Attribution and Grouping Invariance

Origin-centered hyperbolic maps and spatial coordinate conversions admit a scalar–signal structure, as in Eq. 4. This raises an attribution question: should relevance be shared between the geometric factor and the signal, or assigned entirely to the signal? Conservation alone does not determine this choice. To clarify the role of multiplicative structure, we first study attribution to scalar factors with equal explanatory status. This analysis complements the signal-only formulation used by LRP-radial-all.

B.1 Local rules and grouping invariance

For a scalar product $y = ab$ with $a, b > 1$, consider a local rule that is linear in incoming relevance and whose coefficients depend only on the current factor values. Every conservative rule in this class has the form

$$R_a = w(a, b)R_y, \qquad R_b = [1 - w(a, b)]R_y. \tag{17}$$

The same coefficient function w is used at every multiplication. Symmetric treatment of the factors additionally requires $w(a, b) = 1 - w(b, a)$.

Definition 4 (Multiplicative grouping invariance)**.** *A local relevance rule is multiplicatively grouping-invariant if, for every fixed ordered list of factors and incoming relevance, all binary parenthesizations of their product yield identical relevance at each original factor.*

This criterion depends on associative regrouping with the original explanatory factors held fixed. Equal splitting is symmetric and conservative but fails this criterion: for $y = abc$, the parenthesizations $(ab)c$ and $a(bc)$ assign relevance proportions $(1/4, 1/4, 1/2)$ and $(1/2, 1/4, 1/4)$, respectively.

More generally, propagation through $(ab)c$ assigns the coefficients

$$\big(w(ab, c)w(a, b),\ w(ab, c)[1 - w(a, b)],\ 1 - w(ab, c)\big) \tag{18}$$

to (R_a, R_b, R_c) relative to R_y, whereas propagation through $a(bc)$ assigns

$$\big(w(a, bc),\ [1 - w(a, bc)]w(b, c),\ [1 - w(a, bc)][1 - w(b, c)]\big). \tag{19}$$

Grouping invariance therefore requires, for every $a, b, c > 1$,

$$w(ab, c)w(a, b) = w(a, bc), \tag{20}$$

$$w(ab, c)[1 - w(a, b)] = [1 - w(a, bc)]w(b, c), \tag{21}$$

$$1 - w(ab, c) = [1 - w(a, bc)][1 - w(b, c)]. \tag{22}$$

These identities are also sufficient: any two binary parenthesizations are connected by elementary reassociations, each of which preserves the relevance entering its three sub-expressions.

B.2 Characterization of symmetric factor attribution

Theorem 1 (Continuous symmetric grouping-invariant attribution)**.** *Let $w : (1,\infty)^2 \to \mathbb{R}$ be continuous. The conservative local rule in Eq. 17 is symmetric and multiplicatively grouping-invariant if and only if*

$$R_a = \frac{\log a}{\log a + \log b} R_y, \qquad R_b = \frac{\log b}{\log a + \log b} R_y. \tag{23}$$

Proof. Take incoming relevance $R_y = 1$ and consider a product of N identical factors $t > 1$. Any adjacent pair can be made siblings by changing the parenthesization. Since symmetry gives $w(t,t) = 1/2$, the two factors receive equal relevance in that parenthesization. Grouping invariance transfers this equality to every parenthesization. All adjacent factors therefore receive equal relevance, and conservation implies an allocation of $1/N$ to each factor.

Now consider $m + n$ identical factors, where m, n are positive integers, and choose a tree whose root separates the first m factors from the remaining n. Because the rule depends only on the two current factor values, the root assigns relevance $w(t^m, t^n)$ to the first subtree. Conservation within this subtree gives

$$w(t^m, t^n) = \frac{m}{m+n}. \tag{24}$$

Consequently, whenever $\log a / \log b = m/n$, choosing $t = a^{1/m} = b^{1/n}$ yields

$$w(a,b) = \frac{\log a}{\log a + \log b}. \tag{25}$$

This argument uses grouping invariance on the fixed list of identical factors and locality at the root and does not require a separate assumption of invariance under factor expansion.

For arbitrary $a, b > 1$, choose positive rational numbers $q_r \to \log a / \log b$. Since $b^{q_r} \to a$, continuity extends the identity $w(b^{q_r}, b) = q_r/(1+q_r)$ to $w(a,b) = \log a/(\log a + \log b)$.

Conversely, the logarithmic rule is continuous, symmetric, and conservative on the stated domain. For any product $y = \prod_{i=1}^N a_i$, propagation through an arbitrary binary tree yields

$$R_{a_i} = \frac{\log a_i}{\sum_{j=1}^N \log a_j} R_y, \tag{26}$$

because intermediate logarithmic sums cancel along each root-to-leaf path. The allocation is therefore independent of parenthesization. □

B.3 Connection to radial attribution

This characterization applies to continuous, symmetric, value-based rules on $(1,\infty)^2$. The logarithmic formula does not directly cover general neural activations, since it is undefined for nonpositive factors and singular when positive factors have product one. Our geometric modules instead distinguish a vector signal from its scalar modulation, motivating the asymmetric LRP-radial-all convention. Its invariance under splitting or merging radial modules differs from regrouping a fixed list of explanatory factors. Appendix C analyzes this setting within the fixed-proportion radial family.

C Characterization of Fixed-Proportion Radial Rules

C.1 Rule family and repeated propagation

Consider nonzero radial modules $\boldsymbol{y} = \alpha(\|\boldsymbol{x}\|_2)\boldsymbol{x}$ in dimension $d \geq 2$, with nonzero intermediate signals. We restrict attention to the rule family in Eq. 8, where $\eta \in [0,1]$ is a fixed constant shared across modules. The signal branch receives proportion η of each output relevance, while the scalar branch receives the remaining total relevance and redistributes it according to squared input coordinates.

For compactness, define the matrix

$$P_{\boldsymbol{x}} = \frac{\boldsymbol{x} \odot \boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} \mathbf{1}^\top, \qquad T_{\eta,\boldsymbol{x}} = \eta I_d + (1-\eta) P_{\boldsymbol{x}}, \tag{27}$$

where $\odot$ denotes element-wise multiplication. The rule is $\boldsymbol{R}_{\boldsymbol{x}} = T_{\eta,\boldsymbol{x}}\boldsymbol{R}_{\boldsymbol{y}}$. Since the normalized squared coordinates sum to one,

$$P_{\boldsymbol{x}}^2 = P_{\boldsymbol{x}}, \qquad \mathbf{1}^\top P_{\boldsymbol{x}} = \mathbf{1}^\top. \tag{28}$$

Furthermore, $P_{a\boldsymbol{x}} = P_{\boldsymbol{x}}$ for every nonzero scalar a. Thus every module along a nonzero radial chain has the same squared-coordinate redistribution matrix.

Proposition 2 (Factorization invariance within the fixed-proportion family)**.** *For $d \geq 2$, a rule in Eq. 8 is invariant, for every nonzero input and incoming relevance, under replacement of one radial module by an equivalent chain of nonzero radial modules if and only if $\eta \in \{0, 1\}$.*

Proof. Since $P_{\boldsymbol{x}}$ is idempotent,

$$T_{\eta,\boldsymbol{x}}^m = \eta^m I_d + (1 - \eta^m)P_{\boldsymbol{x}} \tag{29}$$

for every $m \geq 1$. In particular, invariance between one and two modules requires

$$T_{\eta,\boldsymbol{x}}^2 - T_{\eta,\boldsymbol{x}} = (\eta^2 - \eta)(I_d - P_{\boldsymbol{x}}) = 0. \tag{30}$$

For $d \geq 2$, $P_{\boldsymbol{x}}$ has rank one and cannot equal I_d. Therefore $\eta^2 = \eta$, giving $\eta = 0$ or $\eta = 1$. Conversely, both choices yield idempotent backward operators, so repeated application does not change the propagated relevance. □

The case $\eta = 1$ is radial-all. The case $\eta = 0$ assigns all relevance to the scalar branch and then redistributes it according to squared input coordinates. Both rules conserve relevance and satisfy the stated radial factorization invariance, demonstrating that these two properties alone do not uniquely identify radial-all.

C.2 Adding zero-curvature consistency

Corollary 1 (Unique joint consistency within the fixed-proportion family)**.** *For $d \geq 2$, LRP-radial-all is the unique member of the fixed-proportion family in Eq. 8 satisfying both radial factorization invariance and zero-curvature consistency for the origin Poincaré maps in the fixed coordinates of Definition 2.*

Proof. Proposition 2 restricts the candidates to $\eta = 0$ and $\eta = 1$. For fixed nonzero input and incoming relevance, $T_{\eta,\boldsymbol{x}}$ is independent of curvature. Zero-curvature (as in Definition 2) consistency therefore requires $T_{\eta,\boldsymbol{x}}\boldsymbol{R} = \boldsymbol{R}$ for every $\boldsymbol{R}$. The choice $\eta = 1$ satisfies this requirement. The choice $\eta = 0$ does not, since $P_{\boldsymbol{x}} \neq I_d$ for $d \geq 2$. Hence only $\eta = 1$ satisfies both criteria. □

D Origin Logarithmic Maps and the GRI Test

D.1 Poincaré branch

Let $r_P = \|\boldsymbol{x}_P\|_2$. The origin logarithmic map has the radial form

$$\boldsymbol{v} = \log_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{x}_P) = \alpha_P(\boldsymbol{x}_P)\boldsymbol{x}_P, \qquad \alpha_P(\boldsymbol{x}_P) = \frac{\operatorname{artanh}(\sqrt{c}\, r_P)}{\sqrt{c}\, r_P}. \tag{31}$$

The continuous value at the origin is $\alpha_P(\mathbf{0}) = 1$.

D.2 Lorentz branch and tangent alignment

Let $\theta_L = \operatorname{arcosh}(\sqrt{c}\, x_{L,0})$. The Lorentz logarithmic map is

$$\log_{\boldsymbol{o}_L}^{L,c}(\boldsymbol{x}_L) = \frac{\theta_L}{\sinh\theta_L}\left(\boldsymbol{x}_L - \sqrt{c}\, x_{L,0}\boldsymbol{o}_L\right) = \left(0, \frac{\theta_L}{\sinh\theta_L}\bar{\boldsymbol{x}}_L\right). \tag{32}$$

To compare this output with the Poincaré branch, we express both in common tangent coordinates. The Jacobian of the isometry in Eq. 14 satisfies

$$\left.\frac{\partial\phi(\boldsymbol{x}_P)}{\partial\boldsymbol{x}_P}\right|_{\boldsymbol{x}_P=\mathbf{0}} \boldsymbol{v} = \begin{pmatrix}\mathbf{0}^\top \\ 2I_d\end{pmatrix}\boldsymbol{v} = (0, 2\boldsymbol{v}). \tag{33}$$

Thus, the Lorentz spatial tangent component must be divided by two to recover the common coordinates:

$$\boldsymbol{v} = \frac{1}{2}\left[\log_{\boldsymbol{o}_L}^{L,c}(\boldsymbol{x}_L)\right]_{\mathrm{sp}} = \alpha_L(\boldsymbol{x}_L)\bar{\boldsymbol{x}}_L, \qquad \alpha_L(\boldsymbol{x}_L) = \frac{\theta_L}{2\sinh\theta_L}. \tag{34}$$

Here $[\cdot]_{\mathrm{sp}}$ extracts the spatial components. Let $r_L = \|\bar{\boldsymbol{x}}_L\|_2$, the hyperboloid constraint gives $\sqrt{c}\,x_{L,0} = \sqrt{1 + c r_L^2}$ and $\sinh\theta_L = \sqrt{c}\,r_L$. Consequently,

$$\alpha_L(\boldsymbol{x}_L) = \frac{\operatorname{arsinh}(\sqrt{c}\,r_L)}{2\sqrt{c}\,r_L}, \tag{35}$$

so the aligned Lorentz logarithmic map is radial in the spatial coordinates, with $\alpha_L(\boldsymbol{o}_L) = 1/2$.

D.3 Forward equivalence

For $\boldsymbol{x}_L = \phi(\boldsymbol{x}_P)$ and $t = \sqrt{c}\,r_P \in [0, 1)$, the conversion formulas give

$$\sqrt{c}\,x_{L,0} = \frac{1 + t^2}{1 - t^2}, \qquad \theta_L = 2\operatorname{artanh}(t), \qquad \sinh\theta_L = \frac{2t}{1 - t^2}. \tag{36}$$

Together with $\bar{\boldsymbol{x}}_L = \gamma(\boldsymbol{x}_P)\boldsymbol{x}_P$,

$$\alpha_L(\phi(\boldsymbol{x}_P))\gamma(\boldsymbol{x}_P) = \frac{2\operatorname{artanh}(t)}{2[2t/(1 - t^2)]}\frac{2}{1 - t^2} = \frac{\operatorname{artanh}(t)}{t} = \alpha_P(\boldsymbol{x}_P). \tag{37}$$

All expressions at $t = 0$ are interpreted by continuity. The two paths therefore produce identical common tangent coordinates:

$$\boldsymbol{v} = \log_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{x}_P) = \frac{1}{2}\left[\log_{\boldsymbol{o}_L}^{L,c}(\phi(\boldsymbol{x}_P))\right]_{\mathrm{sp}}. \tag{38}$$

D.4 Relevance equality in the controlled test

Under the shared encoder, downstream computation, and relevance procedure specified in Section 3.2, both paths receive the same $\boldsymbol{R}_{\boldsymbol{v}}$. LRP-radial-all propagates this relevance unchanged through each spatial radial module. The direct path contains the Poincaré logarithmic map, while the Lorentz path contains the aligned logarithmic map and the spatial coordinate conversion. Hence,

$$\boldsymbol{R}_{\boldsymbol{x}_P}^{P} = \boldsymbol{R}_{\boldsymbol{v}}, \qquad \boldsymbol{R}_{\boldsymbol{x}_P}^{L} = \boldsymbol{R}_{\bar{\boldsymbol{x}}_L} = \boldsymbol{R}_{\boldsymbol{v}}. \tag{39}$$

The Lorentz time coordinate is treated as a dependent geometric quantity within these modules. Identical relevance at $\boldsymbol{x}_P$, followed by the same deterministic propagation through the shared encoder, gives

$$\boldsymbol{R}_{\boldsymbol{a}}^{\mathrm{direct}} = \boldsymbol{R}_{\boldsymbol{a}}^{\mathrm{via}\,L}. \tag{40}$$

LRP-radial-all therefore satisfies GRI for this specified logarithmic-map test.

E Exponential Maps and the Tests

E.1 Poincaré exponential map

Let $r_z = \|\boldsymbol{z}\|_2$. The origin exponential map has the radial form

$$\boldsymbol{y}_P = \exp_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{z}) = \beta_P(\boldsymbol{z})\boldsymbol{z}, \qquad \beta_P(\boldsymbol{z}) = \frac{\tanh(\sqrt{c}\,r_z)}{\sqrt{c}\,r_z}. \tag{41}$$

The continuous value at the origin is $\beta_P(\boldsymbol{0}) = 1$.

E.2 LORENTZ EXPONENTIAL MAP AND ALIGNMENT

Under the tangent alignment in Eq. 33, the common coordinates $\boldsymbol{z}$ correspond to the ambient Lorentz tangent vector $(0, 2\boldsymbol{z})$. For an origin tangent vector $\boldsymbol{u} = (0, \bar{\boldsymbol{u}})$, with $\|\boldsymbol{u}\|_L = \|\bar{\boldsymbol{u}}\|_2$, the exponential map is

$$\exp_{\boldsymbol{o}_L}^{L,c}(\boldsymbol{u}) = \cosh(\sqrt{c}\,\|\boldsymbol{u}\|_L)\boldsymbol{o}_L + \frac{\sinh(\sqrt{c}\,\|\boldsymbol{u}\|_L)}{\sqrt{c}\,\|\boldsymbol{u}\|_L}\boldsymbol{u}. \tag{42}$$

Substituting $\boldsymbol{u} = (0, 2\boldsymbol{z})$ gives

$$\boldsymbol{y}_L = \exp_{\boldsymbol{o}_L}^{L,c}((0, 2\boldsymbol{z})) = \left(\frac{\cosh(2\sqrt{c}\,r_z)}{\sqrt{c}}, \beta_L(\boldsymbol{z})\boldsymbol{z}\right), \qquad \beta_L(\boldsymbol{z}) = \frac{\sinh(2\sqrt{c}\,r_z)}{\sqrt{c}\,r_z}. \tag{43}$$

Thus, the spatial output $\bar{\boldsymbol{y}}_L = \beta_L(\boldsymbol{z})\boldsymbol{z}$ is radial, with $\beta_L(\boldsymbol{0}) = 2$.

To verify forward equivalence, set $t = \sqrt{c}\,r_z$. Taking the norm of Eq. 41 gives $\|\boldsymbol{y}_P\|_2 = \tanh(t)/\sqrt{c}$. Substitution into Eq. 14 then yields

$$\begin{aligned}\phi(\boldsymbol{y}_P) &= \left(\frac{1+\tanh^2 t}{\sqrt{c}(1-\tanh^2 t)}, \frac{2\tanh t}{t(1-\tanh^2 t)}\boldsymbol{z}\right) \\ &= \left(\frac{\cosh(2t)}{\sqrt{c}}, \frac{\sinh(2t)}{t}\boldsymbol{z}\right) = \boldsymbol{y}_L,\end{aligned} \tag{44}$$

where the expressions at $t = 0$ are interpreted by continuity. The two exponential maps therefore produce corresponding geometric points when their tangent inputs are aligned.

E.3 RELEVANCE RULES FOR THE SPATIAL RADIAL MODULES

Both $\boldsymbol{y}_P = \beta_P(\boldsymbol{z})\boldsymbol{z}$ and $\bar{\boldsymbol{y}}_L = \beta_L(\boldsymbol{z})\boldsymbol{z}$ have the scalar–signal structure of Eq. 7. We apply the same propagation rules as for the logarithmic maps, with $\boldsymbol{z}$ as the signal. The constant tangent alignment is included in β_L. LRP-radial-all and LRP-half are applicable here.

E.4 A CONTROLLED EXPONENTIAL-MAP TEST

The exponential-map test compares two paths from the same tangent coordinates $\boldsymbol{z}$ to the same Lorentz point. The direct path applies the aligned Lorentz exponential map, whereas the indirect path applies the Poincaré exponential map followed by the coordinate conversion:

$$\boldsymbol{y}_L = \exp_{\boldsymbol{o}_L}^{L,c}((0, 2\boldsymbol{z})) = \phi\big(\exp_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{z})\big)\,. \tag{45}$$

Both paths then use the same downstream computation $G(\boldsymbol{y}_L)$. With identical output relevance initialization and the same deterministic downstream propagation, they receive the same relevance $\boldsymbol{R}_{\boldsymbol{y}_L} = (R_{y_{L,0}}, \boldsymbol{R}_{\bar{\boldsymbol{y}}_L})$.

Shared treatment of the time coordinate. In both realizations, we express the time coordinate through the same spatial constraint,

$$y_{L,0} = \sqrt{c^{-1} + \|\bar{\boldsymbol{y}}_L\|_2^2}. \tag{46}$$

Any relevance assigned to this coordinate is propagated through the constraint using the same rule in both paths. For example, for $\bar{\boldsymbol{y}}_L \neq \boldsymbol{0}$, squared-coordinate redistribution gives

$$\widetilde{R}_{\bar{y}_{L,i}} = R_{\bar{y}_{L,i}} + \frac{\bar{y}_{L,i}^2}{\|\bar{\boldsymbol{y}}_L\|_2^2} R_{y_{L,0}}. \tag{47}$$

This convention conserves the combined spatial and time-coordinate relevance. At zero spatial input, a shared fallback must be specified. The equality argument below requires only that both paths use the same deterministic treatment of the time coordinate.

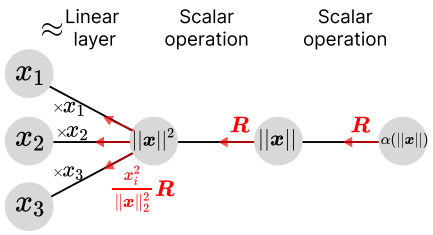


Figure 7: Redistribution of scalar-branch relevance from $\alpha(\|\boldsymbol{x}\|_2)$ to $\boldsymbol{x}$ through the squared norm.

Relevance equality under LRP-radial-all. After the shared time-coordinate propagation, both paths receive the same effective spatial relevance $\widetilde{\boldsymbol{R}}_{\bar{\boldsymbol{y}}_L}$. The direct spatial map is $\bar{\boldsymbol{y}}_L = \beta_L(\boldsymbol{z})\boldsymbol{z}$, while the indirect path contains two radial modules:

$$\boldsymbol{y}_P = \beta_P(\boldsymbol{z})\boldsymbol{z}, \qquad \bar{\boldsymbol{y}}_L = \gamma(\boldsymbol{y}_P)\boldsymbol{y}_P. \tag{48}$$

Applying LRP-radial-all to each spatial radial module gives

$$\boldsymbol{R}_{\boldsymbol{z}}^{\text{direct}} = \widetilde{\boldsymbol{R}}_{\bar{\boldsymbol{y}}_L}, \qquad \boldsymbol{R}_{\boldsymbol{z}}^{\text{via } P} = \boldsymbol{R}_{\boldsymbol{y}_P} = \widetilde{\boldsymbol{R}}_{\bar{\boldsymbol{y}}_L}. \tag{49}$$

Thus, the two paths yield identical relevance at the common input $\boldsymbol{z}$. If $\boldsymbol{z}$ is produced by a shared encoder, identical deterministic propagation through that encoder also gives identical relevance at the original input.

F LRP-HALF: CONSERVATION AND CONSISTENCY COUNTEREXAMPLES

This section analyzes the specified LRP-half baseline, which combines equal splitting at scalar–signal products with squared-norm redistribution of the scalar relevance. Although this rule conserves total relevance, it can violate GRI and zero-curvature consistency.

F.1 BRANCH SPLIT, SQUARED-NORM REDISTRIBUTION, AND CONSERVATION

Consider a radial module $\boldsymbol{y} = \alpha(\|\boldsymbol{x}\|_2)\boldsymbol{x}$ with $\boldsymbol{x} \neq \boldsymbol{0}$. LRP-half assigns half of the incoming relevance to the signal branch and half to the scalar factor:

$$R_{x_i}^{\text{sig}} = \frac{1}{2}R_{y_i}, \qquad R_\alpha = \frac{1}{2}\sum_j R_{y_j}. \tag{50}$$

Let $s = \|\boldsymbol{x}\|_2^2 = \sum_j x_j^2$ and $r = \sqrt{s}$. We treat the scalar transformations $s \mapsto r \mapsto \alpha(r)$ as relevance-preserving, so $R_s = R_r = R_\alpha$. Contribution-proportional redistribution through the sum, followed by relevance-preserving propagation through each scalar square, gives

$$R_{x_i}^{\text{rad}} = \frac{x_i^2}{\|\boldsymbol{x}\|_2^2}R_s. \tag{51}$$

Combining the two branches yields

$$R_{x_i} = \frac{1}{2}R_{y_i} + \frac{1}{2}\frac{x_i^2}{\|\boldsymbol{x}\|_2^2}\sum_j R_{y_j}. \tag{52}$$

Since the squared-coordinate weights sum to one, the rule conserves signed total relevance:

$$\sum_i R_{x_i} = \frac{1}{2}\sum_i R_{y_i} + \frac{1}{2}\sum_j R_{y_j} = \sum_j R_{y_j}. \tag{53}$$

At $\boldsymbol{x} = \boldsymbol{0}$, the redistribution weights are undefined and require a separate convention, such as identity propagation. The counterexamples below use nonzero inputs and are independent of this choice. Numerical stabilization is discussed in Appendix H.

F.2 FACTORIZATION DEPENDENCE AND THE GRI COUNTEREXAMPLE

Consider the aligned logarithmic-map paths in Eq. 4, with $\boldsymbol{x}_P \neq \mathbf{0}$ and identical incoming relevance $\boldsymbol{R}_{\boldsymbol{v}}$. The direct Poincaré path contains one radial module and gives

$$R^P_{x_{P,i}} = \frac{1}{2} R_{v_i} + \frac{1}{2} \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{v_j}. \tag{54}$$

The Lorentz path contains two radial modules: the spatial conversion $\bar{\boldsymbol{x}}_L = \gamma(\boldsymbol{x}_P)\boldsymbol{x}_P$ and the aligned logarithmic map. Because the conversion scales every coordinate by the same nonzero scalar,

$$\frac{\bar{x}^2_{L,i}}{\|\bar{\boldsymbol{x}}_L\|_2^2} = \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2}. \tag{55}$$

Propagation through the aligned Lorentz logarithmic map first gives

$$R_{\bar{x}_{L,i}} = \frac{1}{2} R_{v_i} + \frac{1}{2} \frac{\bar{x}^2_{L,i}}{\|\bar{\boldsymbol{x}}_L\|_2^2} \sum_j R_{v_j}, \qquad \sum_j R_{\bar{x}_{L,j}} = \sum_j R_{v_j}. \tag{56}$$

Applying the rule again through the spatial conversion yields

$$\begin{aligned} R^L_{x_{P,i}} &= \frac{1}{2} R_{\bar{x}_{L,i}} + \frac{1}{2} \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{\bar{x}_{L,j}} \\ &= \frac{1}{2} \left(\frac{1}{2} R_{v_i} + \frac{1}{2} \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{v_j} \right) + \frac{1}{2} \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{v_j} \\ &= \frac{1}{4} R_{v_i} + \frac{3}{4} \frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{v_j}. \end{aligned} \tag{57}$$

The constant tangent alignment is included in α_L and is not treated as an additional half-splitting node. Both paths conserve the same total relevance, but their coordinate-wise difference is

$$R^L_{x_{P,i}} - R^P_{x_{P,i}} = \frac{1}{4} \left(\frac{x^2_{P,i}}{\|\boldsymbol{x}_P\|_2^2} \sum_j R_{v_j} - R_{v_i} \right). \tag{58}$$

The allocations coincide if and only if $R_{v_i} = (x^2_{P,i}/\|\boldsymbol{x}_P\|_2^2) \sum_j R_{v_j}$ for every i.

An explicit counterexample at the original input. Choose the shared encoder to be the identity on the Poincaré ball, so that $\boldsymbol{a} = \boldsymbol{x}_P$. Let $c = 1$, $\boldsymbol{x}^*_P = (1/4, 1/4)^\top$, and $\boldsymbol{v}^* = \log^{P,1}_{\boldsymbol{o}_P}(\boldsymbol{x}^*_P)$. Use the shared linear head $G(\boldsymbol{v}) = v_1/v_1^*$, where v_1^* is fixed. At this input, the output score is one, and LRP-0 through the head with unit relevance initialization gives $\boldsymbol{R}_{\boldsymbol{v}} = (1, 0)^\top$. The two paths yield

$$\boldsymbol{R}^{\text{direct}}_{\boldsymbol{a}} = (3/4, 1/4)^\top, \qquad \boldsymbol{R}^{\text{via } L}_{\boldsymbol{a}} = (5/8, 3/8)^\top. \tag{59}$$

Both allocations sum to one but differ at the same original input. Since the aligned paths implement the same function throughout the common domain, this is a counterexample to GRI for the specified LRP-half baseline.

F.3 FAILURE OF ZERO-CURVATURE CONSISTENCY

For fixed inputs, the origin Poincaré logarithmic and exponential maps satisfy

$$\log^{P,c}_{\boldsymbol{o}_P}(\boldsymbol{x}) = \left(1 + \frac{c}{3}\|\boldsymbol{x}\|_2^2 + O(c^2)\right) \boldsymbol{x}, \tag{60}$$

$$\exp^{P,c}_{\boldsymbol{o}_P}(\boldsymbol{z}) = \left(1 - \frac{c}{3}\|\boldsymbol{z}\|_2^2 + O(c^2)\right) \boldsymbol{z}. \tag{61}$$

Both converge to the identity as $c \to 0$. However, for fixed nonzero input and incoming relevance, the LRP-half allocation in Eq. 52 is independent of curvature. Consequently,

$$\lim_{c\to 0}\left(R_{x_i} - R_{y_i}\right) = \frac{1}{2}\left(\frac{x_i^2}{\|\boldsymbol{x}\|_2^2}\sum_j R_{y_j} - R_{y_i}\right), \tag{62}$$

which is generally nonzero. Although the radial factor approaches one, the rule continues to redistribute half of the incoming relevance through the radial branch. The same argument applies to the exponential map with $\boldsymbol{z}$ as the input signal. Thus, the specified LRP-half baseline fails zero-curvature consistency for both Poincaré origin maps.

G Bias, Möbius Addition, and Matched Lorentz Realization

G.1 Poincaré bias operation

Let $\boldsymbol{b}$ be a fixed tangent-space parameter and $\boldsymbol{b}_P = \exp_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{b})$. For $\boldsymbol{m}_P = \boldsymbol{W} \otimes_c \boldsymbol{x}_P$, define the bias operation as

$$\boldsymbol{y}_P = T_{\boldsymbol{b}_P}^P(\boldsymbol{m}_P) = \boldsymbol{m}_P \oplus_c \boldsymbol{b}_P. \tag{63}$$

Möbius addition can also be represented in the scalar-vector product form

$$\boldsymbol{m}_P \oplus_c \boldsymbol{b}_P = \alpha_c^{\oplus}(\boldsymbol{m}_P, \boldsymbol{b}_P)\boldsymbol{m}_P + \beta_c^{\oplus}(\boldsymbol{m}_P, \boldsymbol{b}_P)\boldsymbol{b}_P, \tag{64}$$

where

$$\begin{aligned}\alpha_c^{\oplus}(\boldsymbol{m}_P, \boldsymbol{b}_P) &= \frac{1 + 2c\langle \boldsymbol{m}_P, \boldsymbol{b}_P\rangle + c\|\boldsymbol{b}_P\|_2^2}{1 + 2c\langle \boldsymbol{m}_P, \boldsymbol{b}_P\rangle + c^2\|\boldsymbol{m}_P\|_2^2\|\boldsymbol{b}_P\|_2^2},\\ \beta_c^{\oplus}(\boldsymbol{m}_P, \boldsymbol{b}_P) &= \frac{1 - c\|\boldsymbol{m}_P\|_2^2}{1 + 2c\langle \boldsymbol{m}_P, \boldsymbol{b}_P\rangle + c^2\|\boldsymbol{m}_P\|_2^2\|\boldsymbol{b}_P\|_2^2}.\end{aligned} \tag{65}$$

Although $\boldsymbol{b}_P$ is fixed, both coefficients depend on the input $\boldsymbol{m}_P$. The bias contribution generally changes the output direction, so Möbius bias addition is not an origin-centered radial scaling of the signal.

G.2 Signal-only relevance convention

We explain the prediction in terms of input features rather than fixed model parameters. In this spirit, we suggest one way to propagate relevance conservatively:

$$\boldsymbol{R}_{\boldsymbol{b}_P} = \boldsymbol{0}, \qquad \boldsymbol{R}_{\boldsymbol{m}_P} = \boldsymbol{R}_{\boldsymbol{y}_P}. \tag{66}$$

Note it is a separate attribution convention and not a consequence of the radial factorization result.

G.3 Matched Lorentz realization

An exactly equivalent forward bias operation can be constructed through the Poincaré–Lorentz isometry ϕ. Set $\boldsymbol{m}_L = \phi(\boldsymbol{m}_P)$ and $\boldsymbol{b}_L = \phi(\boldsymbol{b}_P)$, and define

$$T_{\boldsymbol{b}_L}^L = \phi \circ T_{\boldsymbol{b}_P}^P \circ \phi^{-1}, \tag{67}$$

$$T_{\boldsymbol{b}_L}^L(\boldsymbol{m}_L) = \phi\big(\phi^{-1}(\boldsymbol{m}_L) \oplus_c \boldsymbol{b}_P\big). \tag{68}$$

The resulting computation is

$$\boldsymbol{m}_L \xrightarrow{\phi^{-1}} \boldsymbol{m}_P \xrightarrow{\oplus_c \boldsymbol{b}_P} \boldsymbol{y}_P \xrightarrow{\phi} \boldsymbol{y}_L. \tag{69}$$

By construction, $T_{\boldsymbol{b}_L}^L(\phi(\boldsymbol{m}_P)) = \phi(T_{\boldsymbol{b}_P}^P(\boldsymbol{m}_P))$, so the two realizations implement the same operation in different geometric representations.

H TANGENT-SPACE PROPAGATION AND NUMERICAL STABILIZATION

For a bias-free tangent-space linear transformation $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{v}$, the unstabilized LRP-0 rule is

$$R_{v_i} = \sum_j \frac{v_i W_{ji}}{z_j} R_{z_j}, \qquad z_j = \sum_k v_k W_{jk}. \tag{70}$$

Assuming $z_j \neq 0$ for every output receiving nonzero relevance, and omitting zero-relevance outputs, summation over the inputs gives exact conservation:

$$\sum_i R_{v_i} = \sum_j \frac{\sum_i v_i W_{ji}}{z_j} R_{z_j} = \sum_j R_{z_j}. \tag{71}$$

Numerical stabilization generally introduces a conservation residual. For LRP-ϵ with $\epsilon > 0$, replace the denominator by $\widetilde{z}_j = z_j + \epsilon \operatorname{sgn}_+(z_j)$, where $\operatorname{sgn}_+(z) = 1$ for $z \geq 0$ and -1 otherwise. The propagated total becomes

$$\sum_i R_{v_i} = \sum_j \frac{z_j}{z_j + \epsilon \operatorname{sgn}_+(z_j)} R_{z_j}. \tag{72}$$

Consequently, the signed conservation residual is

$$\sum_j R_{z_j} - \sum_i R_{v_i} = \sum_j \frac{\epsilon}{|z_j| + \epsilon} R_{z_j}. \tag{73}$$

Exact conservation is therefore not guaranteed unless the residual vanishes or is explicitly handled, for example as in NRM (Xiong et al., 2026).

I GEOMETRIC REPRESENTATION INVARIANCE EXPERIMENT

Setup. We consider curvature $-c$ with $c = 1$ and choose the shared encoder to be the identity on the Poincaré ball, so that the original input is $\boldsymbol{a} = \boldsymbol{x}_P$. We evaluate at $\boldsymbol{x}_P = (0.2, 0.3, -0.1)^\top$. Its Lorentz representation is

$$\boldsymbol{x}_L = \phi(\boldsymbol{x}_P) = \left(\frac{1 + c\|\boldsymbol{x}_P\|_2^2}{\sqrt{c}(1 - c\|\boldsymbol{x}_P\|_2^2)}, \frac{2\boldsymbol{x}_P}{1 - c\|\boldsymbol{x}_P\|_2^2} \right). \tag{74}$$

The direct Poincaré path and the aligned Lorentz path reach the same tangent coordinates:

$$\boldsymbol{v} = \log_{\boldsymbol{o}_P}^{P,c}(\boldsymbol{x}_P) = \frac{1}{2} \left[\log_{\boldsymbol{o}_L}^{L,c}(\phi(\boldsymbol{x}_P)) \right]_{\mathrm{sp}}, \qquad \boldsymbol{o}_P = \mathbf{0}, \quad \boldsymbol{o}_L = (c^{-1/2}, \mathbf{0}). \tag{75}$$

Here $[\cdot]_{\mathrm{sp}}$ extracts the spatial components, and the factor $1/2$ converts the ambient Lorentz tangent vector to the common coordinates, as specified in Eq. 33. Both paths use the shared scalar head $G(\boldsymbol{v}) = v_1$ and implement the same function throughout the Poincaré ball. Initializing output relevance with the target score and applying LRP-0 through this head gives $\boldsymbol{R}_{\boldsymbol{v}} = (v_1, 0, 0)^\top$.

Relevance propagation. For a radial module $\boldsymbol{y} = \alpha(\|\boldsymbol{x}\|_2)\boldsymbol{x}$, LRP-radial-all propagates $\boldsymbol{R}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{y}}$, whereas the specified LRP-half baseline uses

$$R_{x_i} = \frac{1}{2} R_{y_i} + \frac{1}{2} \frac{x_i^2}{\|\boldsymbol{x}\|_2^2} \sum_j R_{y_j}, \qquad \boldsymbol{x} \neq \mathbf{0}. \tag{76}$$

Both rules conserve signed total relevance locally. The direct path contains one radial module, $\boldsymbol{v} = \alpha_P(\boldsymbol{x}_P)\boldsymbol{x}_P$. The Lorentz path contains two: the spatial conversion $\bar{\boldsymbol{x}}_L = \gamma(\boldsymbol{x}_P)\boldsymbol{x}_P$ and the aligned logarithmic map $\boldsymbol{v} = \alpha_L(\boldsymbol{x}_L)\bar{\boldsymbol{x}}_L$. We apply each rule consistently to these modules and compare the resulting relevance at the common original input $\boldsymbol{a} = \boldsymbol{x}_P$. The tangent alignment factor is included in α_L, rather than treated as an additional multiplication node. The Lorentz time coordinate is treated as a dependent geometric quantity within the radial modules.

Table 3: Relevance at the common original input $\boldsymbol{a} = \boldsymbol{x}_P$ for the two equivalent logarithmic-map paths. Values are rounded to six decimal places.

Rule	Path	R_{a_1}	R_{a_2}	R_{a_3}
LRP-radial-all	Direct	0.210205	0	0
	Via Lorentz	0.210205	0	0
LRP-half	Direct	0.135132	0.067566	0.007507
	Via Lorentz	0.097595	0.101349	0.011261

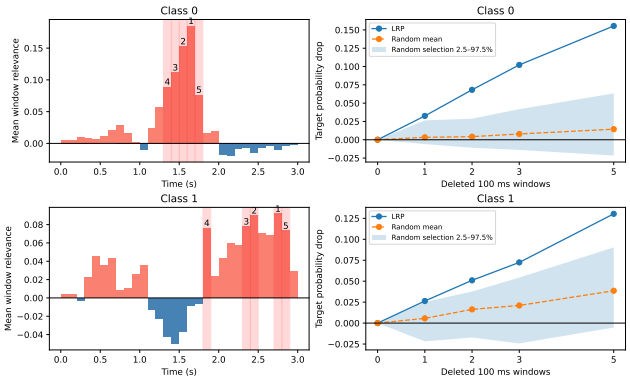


Figure 8: Temporal perturbation validation of LRP for low-load (top) and high-load (bottom) trials. Left: window relevance and the five highest-ranked windows. Right: mean target probability drops under LRP-guided versus random perturbation. Shading indicates the 2.5th-97.5th percentiles across 20 random selections.

Results. Table 3 shows identical input relevance for LRP-radial-all across the two paths. In contrast, LRP-half produces different relevances, although both allocations sum to the same target score v_1. This numerical example illustrates the counterexample derived in Appendix F: local conservation does not guarantee GRI, because equivalent radial factorizations can yield different relevance allocations at the same original input.

J Validate sEEG relevant window explanation by perturbation

To assess whether the relevant time windows in the LRP heatmaps identify input regions relevant to the network's predictions, we conduct a window perturbation test. We compute trial-wise relevance for the target class logit and averaged the maps within each class. The average input time series (3 seconds long) is divided into non-overlapping 100-ms windows, ranked by their signed relevance summed across channels and time. We perturb the top $k \in \{1, 2, 3, 5\}$ windows in each original trial by replacing the signal within these windows, across all channels, with each channel's original temporal mean. Randomly selected windows serve as a control, averaged over 20 repetitions. The results are summarized in Figure 8. LRP-guided perturbation produces larger mean reductions in target probabilities than random perturbation for both classes. These results indicate that the temporal relevance estimated by LRP is aligned with the model's predictive behavior and can effectively identify time periods that are important for the classification decision. This further supports the practical utility of the relevance heatmaps.

K Model Configuration for CIFAR-10

Model architecture. We use a lightweight Lorentz convolutional network inspired by HyperbolicCV (Bdeir et al., 2024), with explicit origin-centered exponential and logarithmic maps for evaluating relevance propagation through radial transformations. We fix the sectional curvature to $-c$ with $c = 1$ and represent each hyperbolic feature by its spatial coordinates $\boldsymbol{s} \in \mathbb{R}^d$, with the time coordinate implicitly given by $t(\boldsymbol{s}) = \sqrt{c^{-1} + \|\boldsymbol{s}\|_2^2}$. Let $\boldsymbol{o}_L = (c^{-1/2}, \mathbf{0})$. For convenience, we

denote the spatial components of the origin exponential and logarithmic maps by

$$\exp^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}(\boldsymbol{v}) := \left[\exp^{L,c}_{\boldsymbol{o}_L}((0,\boldsymbol{v}))\right]_{\mathrm{sp}} = \frac{\sinh(\sqrt{c}\,\|\boldsymbol{v}\|_2)}{\sqrt{c}\,\|\boldsymbol{v}\|_2}\boldsymbol{v},$$
$$\log^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}(\boldsymbol{s}) := \left[\log^{L,c}_{\boldsymbol{o}_L}((t(\boldsymbol{s}),\boldsymbol{s}))\right]_{\mathrm{sp}} = \frac{\operatorname{arsinh}(\sqrt{c}\,\|\boldsymbol{s}\|_2)}{\sqrt{c}\,\|\boldsymbol{s}\|_2}\boldsymbol{s}, \tag{77}$$

where both scalar factors are defined as one at zero. Here $\boldsymbol{v}$ denotes the spatial component of the ambient Lorentz tangent vector $(0,\boldsymbol{v})$, rather than the aligned Poincaré tangent coordinates used in the GRI test. These maps operate independently at each spatial location across the feature channels. The normalized RGB input $\boldsymbol{x}$ is first mapped to spatial Lorentz coordinates as $\boldsymbol{s}^{(0)} = \exp^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}(0.25\boldsymbol{x})$. We use the factor of 0.25 to moderate the initial tangent-space radius and the resulting hyperbolic coordinate magnitudes.

The backbone contains six 3×3 convolutional layers with channel widths $(32, 32, 64, 64, 128, 128)$, strides $(1, 1, 2, 1, 2, 1)$, and padding of one. For the flattened spatial patch $\boldsymbol{p}_u^{(\ell-1)}$ at location u, the ℓth convolution computes

$$\tau_u^{(\ell)} = \sqrt{c^{-1} + \|\boldsymbol{p}_u^{(\ell-1)}\|_2^2}, \qquad \boldsymbol{z}_u^{(\ell)} = \boldsymbol{W}_s^{(\ell)}\boldsymbol{p}_u^{(\ell-1)} + \boldsymbol{w}_t^{(\ell)}\tau_u^{(\ell)} + \boldsymbol{b}^{(\ell)}, \tag{78}$$

where $\boldsymbol{W}_s^{(\ell)}$ and $\boldsymbol{w}_t^{(\ell)}$ are learnable spatial and time-coordinate weights, respectively. Each convolution is followed by a tangent-space activation,

$$\boldsymbol{s}_u^{(\ell)} = \exp^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}\Big(\mathrm{ReLU}\Big(\log^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}(\boldsymbol{z}_u^{(\ell)})\Big)\Big). \tag{79}$$

The resulting feature resolutions are 32×32, 16×16, and 8×8, with two convolutional layers at each resolution. After the final layer, we average the spatial coordinates over the 8×8 feature grid, apply the logarithmic map, and obtain the ten class logits through a linear classifier:

$$\bar{\boldsymbol{s}} = \frac{1}{64}\sum_{u=1}^{64}\boldsymbol{s}_u^{(6)}, \qquad f(\boldsymbol{x}) = \boldsymbol{W}_{\mathrm{cls}}\log^{L,c}_{\boldsymbol{o}_L,\mathrm{sp}}(\bar{\boldsymbol{s}}) + \boldsymbol{b}_{\mathrm{cls}}. \tag{80}$$

Thus, pooling is an arithmetic mean of spatial coordinates rather than a Lorentz centroid, and classification is performed in the tangent space. Note that this architecture is a custom lightweight variant rather than the L-ResNet18 in HyperbolicCV.

Training configuration. We split the 50000 CIFAR-10 training images into 45000 training and 5000 validation samples, and reserve the official 10,000-image test set for final evaluation. Training augmentation consists of random 32×32 crops with four-pixel padding and random horizontal flips. Inputs are normalized using channel-wise means $(0.4914, 0.4822, 0.4465)$ and standard deviations $(0.2470, 0.2435, 0.2616)$. We optimize cross-entropy loss for 200 epochs using AdamW with batch size 128, initial learning rate 10^{-3}, and weight decay 10^{-4}. The learning rate follows a cosine schedule to zero, and the global gradient norm is clipped to 5. The curvature and input scaling factor remain fixed throughout training. The checkpoint with the highest validation accuracy (85.80%) is selected for test evaluation (test accuracy is 85.70%) and attribution experiments.

Relevance propagation through Lorentz convolutions. We treat each convolution as an affine mapping of the augmented patch coordinates $(\boldsymbol{p}_u, \tau_u)$ and first redistribute output relevance to the spatial and time branches using the $z^{\mathcal{B}}$ rule in the first layer and the γ-rule in subsequent layers. For a γ-rule layer, suppressing the layer index, the relevance assigned to the patch time coordinate is

$$R_{\tau_u} = \sum_j \frac{\tau_u\,\rho_\gamma(w_{t,j})}{\mathrm{stab}_\epsilon(\sum_i p_{u,i}\rho_\gamma(W_{s,ji}) + \tau_u\rho_\gamma(w_{t,j}) + b_j)}R_{u,j}, \tag{81}$$

where $\rho_\gamma(w) = w + \gamma\max(w,0)$ and $\mathrm{stab}_\epsilon(z) = z + \epsilon\,\mathrm{sign}_+(z)$, with $\mathrm{sign}_+(0) = 1$. Since τ_u is determined by the spatial patch, we subsequently redistribute its relevance using squared-coordinate proportions:

$$R_{u,i}^{\mathrm{time}} = \frac{p_{u,i}^2}{\sum_m p_{u,m}^2}R_{\tau_u}, \qquad R_{u,i} = R_{u,i}^{\mathrm{space}} + R_{u,i}^{\mathrm{time}}. \tag{82}$$

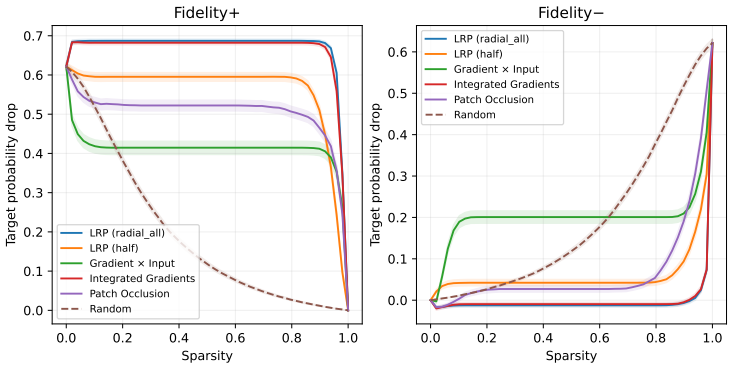


Figure 9: Fidelity-Sparsity curve for MNIST experiment with 95% confidence interval. For Fidelity+ higher is better and for Fidelity- lower is better. Lower sparsity corresponds to a larger selected pixel set, which is removed for Fidelity+ and retained for Fidelity-.

For a zero patch, where $\sum_m p_{u,m}^2 = 0$, we set $R_{u,i}^{\text{time}} = 0$ for every coordinate and record R_{τ_u} as unassigned relevance. The recorded zero-patch contribution, together with residuals from affine biases and numerical stabilization, are summed up as the total relevance conservation residual.

The first-layer $z^{\mathcal{B}}$ rule instead uses bounded contributions for both spatial and time coordinates, followed by the same squared-proportion redistribution. Contributions from overlapping patches are summed at each input coordinate. The fixed curvature term receives no relevance. This time-coordinate rule is shared by LRP-radial-all and LRP-half, which differ only in their treatment of the exponential and logarithmic radial maps.

We use a fixed $\gamma = 0.25$ in convolutional layers 2–6 and $\epsilon = 10^{-6}$ for denominator stabilization throughout relevance propagation, including the ϵ-rule for classifier and average pooling. The first convolution uses the $z^{\mathcal{B}}$ rule without γ modification, for which we derive a fixed bounding box from the valid RGB range $[0, 1]$, accounting for normalization, input scaling, and the exponential map. Time-coordinate bounds are computed from these spatial bounds using the Lorentz constraint. The same box is used for all images. These settings are identical for LRP-radial-all and LRP-half.

Baseline configurations. Integrated Gradients uses a zero baseline in normalized input space, corresponding to the RGB normalization mean, and integration over 128 steps. Attributions are summed across RGB channels to obtain pixel scores. The random baseline averages three independent pixel permutations per image.

Perturbations replace all three channels of selected pixels with the mean-color baseline. We evaluate sparsity with 5% increment steps. Evaluation uses 512 test images sampled uniformly without replacement. We estimate 95% percentile bootstrap confidence intervals using 10,000 image-level resamples with replacement, sharing resampling indices across methods and metrics.

L Additional Figures for Experiments

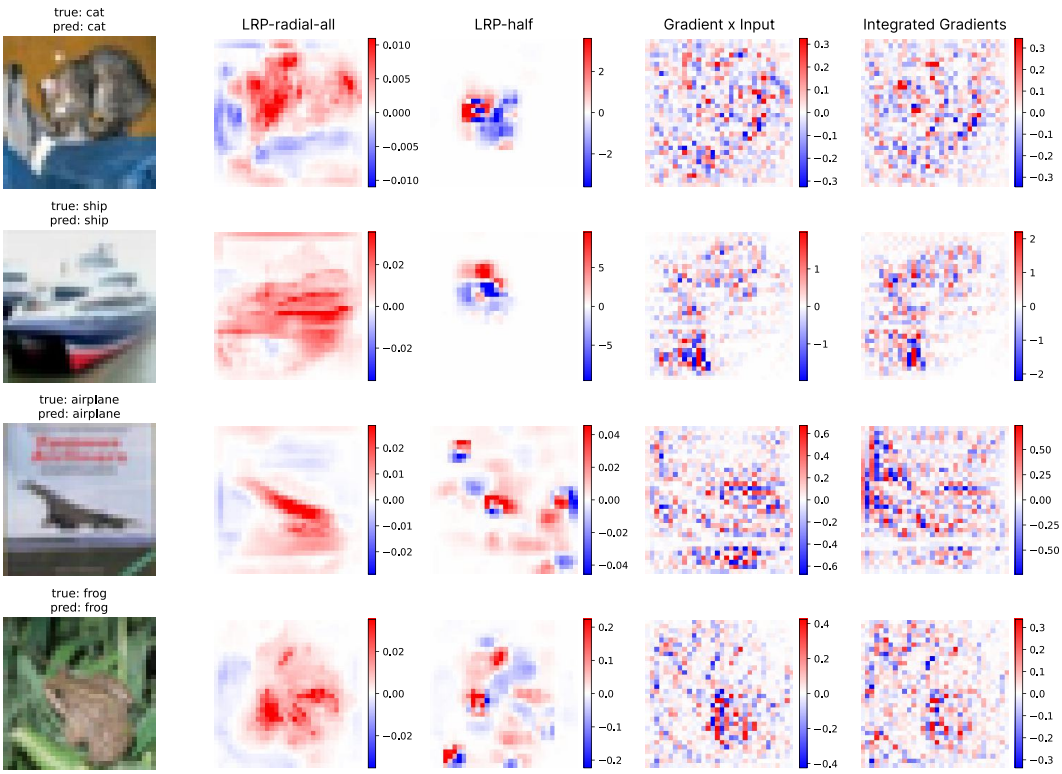
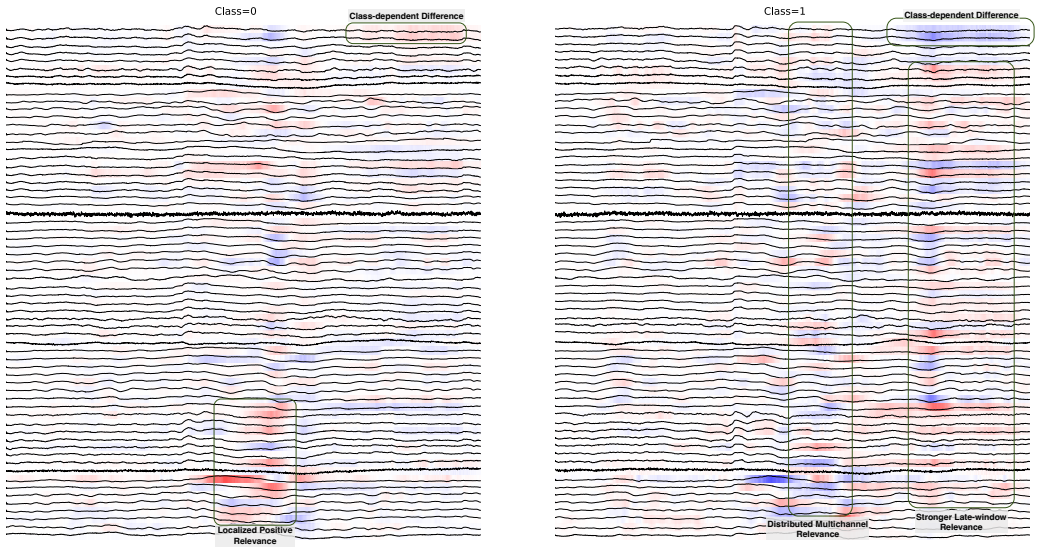


Figure 10: Qualitative attribution comparison on four CIFAR-10 examples. Columns show the input, heatmaps for LRP-radial-all, LRP-half, Gradient×Input, and Integrated Gradients. RGB-channel attributions are summed, with red and blue indicating positive and negative values, respectively.



(a) Low working-memory load class.

(b) High working-memory load class.

Figure 11: Averaged signals and relevance heatmaps for class 0 (low working-memory load) and class 1 (high working-memory load). Boxes highlight localized relevance, distributed multichannel relevance, stronger late-window relevance, and class-dependent attribution differences. Each time series stands for a sensor. Time range from left to right is 3-second with 3000 points. Red and blue indicate positive and negative relevance for the explained class score, respectively.