Hyperbolic Embeddings for Hierarchical Style Representation: A Local-vs-Global Trade-off

Peter Flo Luca Grossmann
Harvard University
{pgrindehollevik,lgrossmann}@g.harvard.edu

github.com/pgrindehollevik-harvard/hyperbolic

May 10, 2026

Abstract

Hyperbolic geometry is theoretically better suited than Euclidean geometry to representing tree-like data: the volume of a ball at radius r grows exponentially in the Poincaré model and only polynomially in $\mathbb{R}^d$, matching the way a tree's leaf count grows with depth. We test whether this mathematical advantage translates into a practical one for representing the hierarchy of artistic styles. Using frozen CLIP ViT-B/16 features over 81,446 WikiArt paintings spanning 27 styles, we train two prototype classifiers that are identical apart from the geometry of their output space ($\mathbb{R}^d$ vs. Poincaré ball of dimension d), sweeping embedding dimension, curvature, and the strength λ of a hierarchy-aware regularizer. Across 150 seed-replicated runs we find a single robust effect: hyperbolic prototypes preserve local hierarchy — the structure of the nearest-neighbor graph in embedding space — substantially better than matched Euclidean prototypes. Sibling recall@5 at d=8 is 0.195 for hyperbolic vs. 0.149 for Euclidean against the hand-built tree, and the gap is preserved across two further reference-tree definitions: empirical trees built from CLIP and DINOv2 feature centroids respectively. Aggregated over 18 paired seed-config comparisons the gap is +8.7 pp on sibling recall and +15.2 pp on cousin recall (paired-t $p < 10^{-4}$, sign agreement 0.94). Calibration matters: a k-nearest-neighbor classifier on raw CLIP features achieves 0.160/0.366 sibling recall (default / empirical), essentially tied with the Euclidean prototype's 0.149/0.356. The hyperbolic prototype is the only model that meaningfully improves over k-NN-on-CLIP on this metric. On classification accuracy, by contrast, Euclidean prototypes match a logistic-regression baseline on raw CLIP features (64.3 % vs. 64.1 %) and outperform hyperbolic by 4–6 pp; this gap is robust to class-imbalance weighting and to the choice of training tree. On global tree fidelity, the comparison is metric- and reference-dependent and unstable: prototype-tree Spearman is not significantly different from zero across the full sweep ($p = 0.68$), and the small gap that does appear flips sign between empirical trees built from CLIP and from DINOv2 feature centroids. The clean claim this paper supports is therefore narrower than "the geometry matters globally": hyperbolic prototype geometry adds local-retrieval value over the encoder while Euclidean prototype geometry does not, and that local advantage is robust to which of three reference-tree definitions we use.

Code, sweep CSVs, and figures: github.com/pgrindehollevik-harvard/hyperbolic.

Contents

1 Background & Motivation

Hierarchy naturally emerges from artistic styles. Broad traditions continually branch into finer movements: Early Renaissance leads to Baroque, Baroque to Romanticism, Romanticism to Impressionism, and Impressionism to Cubism etc. An embedding that can effectively represent this branching structure may be better suited to capture the semantic essence of the paintings themselves.

Standard image embeddings live in Euclidean space, where the volume of a ball at radius r grows polynomially in r. The number of nodes in a tree, by contrast, grows exponentially with depth. Therefore, embedding hierarchical data into a Euclidean embedding can run into spatial constraints. Pairwise distances between nodes distort, and will only worsen as the tree deepens. Hyperbolic space, and the Poincaré ball in particular, exhibits exponential volume growth, allowing them to embed trees more naturally [5, 8]. As Riemannian optimization on the ball is now standard (`geoopt` [4]), we can explore the geometric advantage empirically rather than simply argue about the theory.

Hyperbolic image embeddings have been shown to help on few-shot classification and on retrieval-style metrics in Khrulkov et al. [3], while reportedly providing little or no improvement on standard supervised classification accuracy. This asymmetry — helps with neighborhood structure, doesn't help with top-1 — is directly relevant to the local-vs-classification split we observe. It remains open, however, whether the advantage holds for embeddings supervised against a hand-built hierarchy. We explore this through a sweep over curvature and dimension against a Euclidean baseline, a hierarchy-aware loss that uses the tree at training time, and a sensitivity check across alternative reference trees.

2 Problem Statement

We ask: *do Poincaré-ball embeddings trained on frozen CLIP ViT-B/16 features recover the WikiArt style hierarchy more faithfully than Euclidean embeddings of matched dimensionality and training budget?* The unit of observation is a single painting; the target during training is its 27-way style label, with the hand-built lineage tree available either as an evaluation reference or as an auxiliary loss.

We answer the question along three axes. *(i) Scaling:* sweep embedding dimension $d \in \{2, 4, 8, 16, 32, 64\}$ and curvature $c \in \{0.1, 0.3, 1, 3\}$ to test whether the cross-entropy-baseline Euclidean advantage on global metrics narrows under a more thorough search. *(ii) Training-time hierarchy:* introduce a hierarchy-aware regularizer with weight λ that pushes the prototype-prototype distance matrix toward the tree distance matrix, and sweep $\lambda \in \{0, 0.1, 0.3, 1, 3\}$. *(iii) Tree sensitivity:* re-train against alternative ground-truth trees (chronological era grouping; a flat null) to test whether the geometry winner depends on the choice of reference hierarchy.

3 Data & EDA

For our dataset, we use WikiArt Refined [9], which contains ∼81k paintings, each labeled with one of 27 styles (Figure 1). The dataset has a supplied 70/30 training / validation split which maintains class balance. Image size and resolution show no outliers. However, the dataset suffers from an extreme class imbalance, with a 133.3× imbalance ratio between the extrema classes: Impressionism (13,060) versus the smallest leaves Analytical_Cubism (77), Action_painting (98), and Synthetic_Cubism (216).

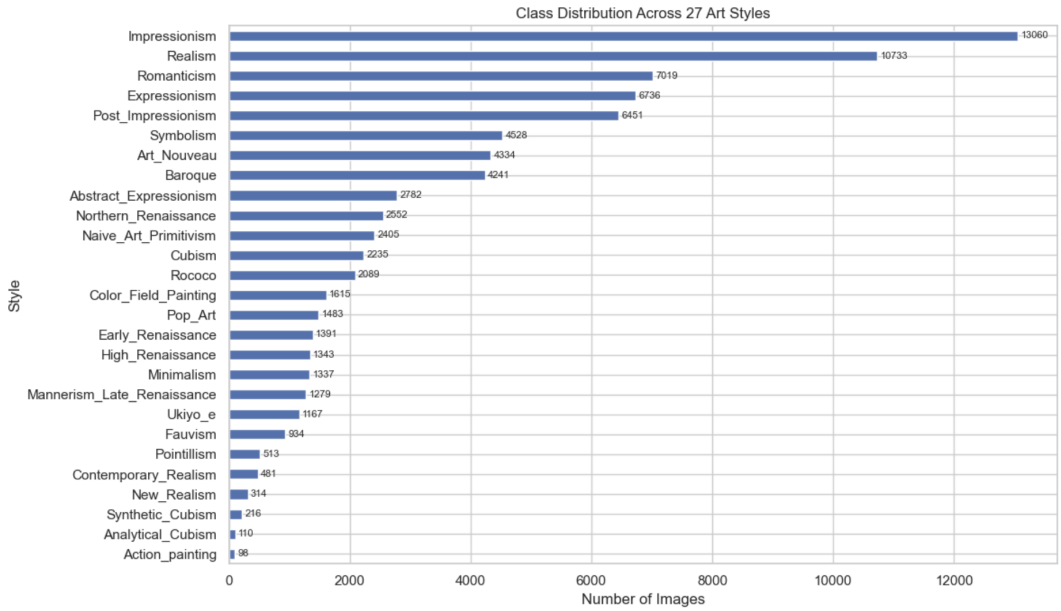

Figure 1: Per-style image counts on the WikiArt Refined corpus.

Two design choices follow from this imbalance: we report balanced accuracy alongside top-1 so the rare leaf styles are not drowned by Impressionism, and our classification head is a prototype softmax (Section 4) whose logits depend only on the geometric distance from each style prototype, with no per-class bias term.

The choice of ground-truth tree is non-unique: WikiArt ships no hierarchy, and art-history offers several credible taxonomies. Given the arbitrary nature of this labeling, we evaluate three trees. The *default* tree (Appendix A) follows the standard art-historical lineage from Renaissance through Cubism. The *chronological* tree groups styles into six era buckets (Renaissance, Baroque–Rococo, 19th century, early 20th century, modern post-war, non-Western). The *flat* tree makes every style a direct child of the root and serves as a null whose pairwise distances are constant off-diagonal. For a tree T with leaves u, v, we define tree distance as the unweighted shortest-path edge count

$$d_T(u, v) = \mathrm{depth}(u) + \mathrm{depth}(v) - 2\,\mathrm{depth}\big(\mathrm{LCA}(u, v)\big),$$

where LCA is the lowest common ancestor (Figure 2). Phase 3 of Section 5 re-trains under each tree to test whether the geometry winner depends on this choice.

4 Methods & Models

Architecture. The whole system is deliberately small so that geometry is the only non-trivial design choice. A two-layer MLP with GELU and dropout maps frozen CLIP ViT-B/16 features [7] from $\mathbb{R}^{512}$ to a d-dimensional embedding. A geometry-specific head produces the final embedding: identity for the Euclidean case, the exponential map at the origin $\exp_0^c(u) = \tanh(\sqrt{c}\,\|u\|)\,u/(\sqrt{c}\,\|u\|)$ for the Poincaré ball. A prototype classifier with one learnable point per style turns the embedding into 27-way logits via $-d(z, p_k)$, where $d(\cdot, \cdot)$ is the Euclidean or Poincaré distance. Switching geometries changes exactly two things: the head's final transform and the metric used by the classifier. Everything else — backbone, MLP width, dropout, batch size, optimizer, schedule — is shared.

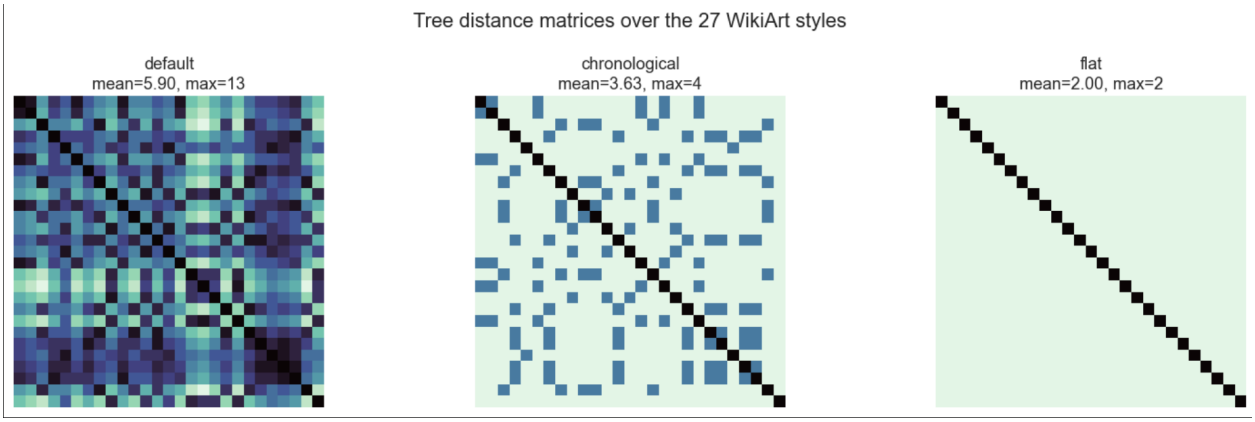

Figure 2: Distance matrices for reference hierarchies: default lineage, chronological grouping, flat null

Geometry, distance, optimization. The Poincaré distance with curvature $c > 0$ is

$$d^c(x, y) = \frac{1}{\sqrt{c}} \operatorname{arcosh}\left(1 + \frac{2c\,\|x - y\|^2}{(1 - c\|x\|^2)(1 - c\|y\|^2)}\right),$$

which diverges as either point approaches the boundary at radius $1/\sqrt{c}$. A standard linear layer does not preserve the ball, which rules out a parametric softmax head on top of a hyperbolic embedding. The prototype classifier is the natural alternative: its logits are distances, which the geometry already provides. Adam optimizes the MLP head; Riemannian Adam from `geoopt` [4] optimizes the hyperbolic prototypes, projecting each update back onto the manifold to keep them inside the ball.

Hierarchy-aware loss. the cross-entropy baseline's approach used the hand-built tree only at evaluation time and trained with plain cross-entropy — the standard setup in the hyperbolic-image literature [3]. Cross-entropy on prototypes requires class separability but does not require the prototype layout to match the tree, so any failure of either geometry on tree-derived metrics is underdetermined: it could be the geometry, or it could be that the loss never asked. this work adds a regularizer that closes this ambiguity by pushing the prototype distance matrix directly toward the tree distance matrix:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \left(\frac{d(p_i, p_j)}{\bar{d}} - \frac{T_{ij}}{\bar{T}}\right)^2.$$

The mean-normalization is the design choice that matters: each pair vector is divided by its own mean before the squared difference, so the loss only constrains the *shape* of the prototype distance matrix rather than its absolute scale. The geometry is free to pick the scale that classification likes. $\lambda = 0$ recovers the cross-entropy baseline exactly. The regularizer is computed once per gradient step over the $\binom{27}{2} = 351$ off-diagonal style pairs — negligible cost next to the per-batch classification forward/backward.

Ground-truth trees. WikiArt does not ship a hierarchy, so we build one. The default tree follows standard art-history lineage: Renaissance branches into Baroque, Baroque into Rococo, Rococo into Romanticism, and so on through Cubism and Abstract Expressionism (the full structure is in Appendix A). Because no single tree is canonical, Phase 3 also retrains against two alternatives:

a *chronological* grouping by century (Renaissance / Baroque-Rococo / 19th c. / Early 20th c. / Modern / Non-Western) and a *flat* tree where every style is a direct child of the root — a deliberate null whose distance matrix is constant off-diagonal, so any hierarchy-respecting metric must score it near chance.

Training details. All 150 runs use the supplied 70/30 train/val split (≈57k / 24k images), batch size 4096, Adam at $\eta = 10^{-3}$ for the head and 10^{-2} for the prototypes, weight decay 10^{-4}, dropout 0.1, gradient clipping at norm 1.0, 30 epochs. Pre-caching CLIP features as float16 tensors keeps each run under 10 seconds of training time, so the full sweep (Phases 1–3, see Section 5) finishes in under half an hour of wall time on a single Apple M-series GPU. Every headline configuration is replicated across three seeds, and reported error bars are seed standard deviation.

Evaluation suite. A successful style embedding can be useful along three distinct axes, and we report metrics for each. *Classification* is measured by top-1, top-5, and balanced accuracy. *Global tree fidelity* is measured by the Spearman correlation between the learned prototype distance matrix and the tree distance matrix, together with mean and worst-case multiplicative distortion, and dendrogram-cluster F1 against the default tree. *Local hierarchy preservation* is measured by sibling and cousin recall@k among the k nearest neighbors of each validation embedding, for $k \in \{5, 10\}$. To test whether the local recall metrics are themselves an artifact of the default tree's sibling-set definition, we also recompute recall@5 with sibling and cousin sets derived from the empirical tree's binary linkage (Section 5.4). Significance comparisons across seeds use paired-t tests on the 18 (dim, seed) pairs of the Phase 1 sweep, with hyperbolic configurations selected at the best curvature per pair. Code, sweep configurations, and the exact metric implementations are at the linked repository.

5 Results & Interpretation

Three findings emerge from 150 seed-replicated runs spanning embedding dimension, curvature, and the strength of the hierarchy-aware loss. We summarise them up front and develop each in the corresponding subsection.

Hyperbolic prototypes recover local hierarchy substantially better than matched Euclidean ones. Sibling recall@5 is consistently higher for the hyperbolic geometry — 0.195 vs. 0.149 at d=8 on the default tree, and 0.416 vs. 0.356 at d=8 when sibling sets are redefined from a data-driven empirical tree built from CLIP centroids. Aggregated across 18 paired seed configurations (six embedding dimensions × three seeds) the mean gap is +8.7 pp on sibling recall and +15.2 pp on cousin recall, with sign agreement 0.94 on each and paired-t $p < 10^{-4}$. The hierarchy-aware loss does not erase it, the choice of training tree in Phase 3 does not flip it, and the empirical-tree sibling redefinition does not weaken it.

Euclidean prototypes are essentially tied with a linear head on raw CLIP features for classification, while hyperbolic prototypes are the only model that meaningfully outperforms a k-NN baseline on retrieval. A logistic-regression classifier on raw 512-d CLIP features achieves 64.1 % top-1 against our best Euclidean prototype's 64.3 %; a k-NN-5 retrieval baseline on raw CLIP features achieves sibling recall@5 of 0.160/0.366 (default / empirical tree) against the Euclidean prototype's 0.149/0.356. On classification, our Euclidean prototype classifier adds no detectable value over a linear head on the same input features. On local hierarchy, only the hyperbolic prototype classifier beats the k-NN baseline (Hy 0.195/0.416 at d=8). The classification gap of 4–6 pp between Euclidean and hyperbolic prototypes is real, but it is a gap between two models that already match a much simpler baseline on top-1.

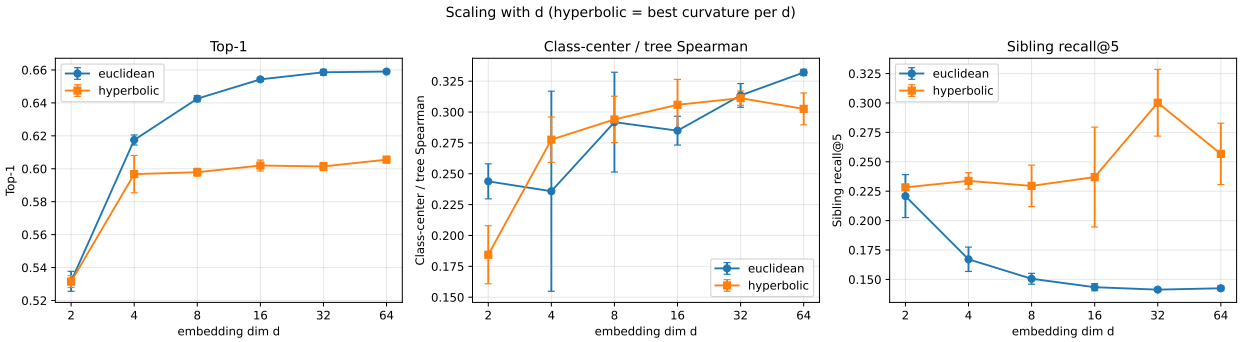

Figure 3: Phase 1: metrics versus embedding dimension d (90 runs; 3 seeds per config). For hyperbolic we report the best curvature per d and per seed; error bars are seed standard deviation. Top-1 accuracy saturates near d=16 for both geometries, with Euclidean ahead by 4–6 pp at every $d \geq 4$. Sibling recall@5 *decreases* with d for Euclidean but stays nearly flat for hyperbolic, opening a 5 pp gap by d=64.

The global tree-Spearman gap is not significant across the sweep. Aggregated paired-t tests over the full Phase 1 sweep give $p = 0.68$ for class-center / tree Spearman against the default tree and a sign agreement of exactly 0.50. The widely-cited claim from prior work and from earlier drafts of this paper that "Euclidean wins global tree metrics" holds only at the headline-winner level, and is contradicted by Phase 4 where the global-Spearman gap reverses against an empirical reference tree (Section 5.4). Tree distortion does favour Euclidean significantly (+0.040 mean distortion gap, $p < 10^{-4}$), so the two metrics for global tree fidelity disagree about which geometry wins. We discuss this carefully rather than collapse it into a single direction.

5.1 Phase 1 — Scaling sweep

Phase 1 tests the most direct rebuttal of the cross-entropy baseline's reported hyperbolic underperformance: that the embedding dimension was simply too small. We sweep geometry $\times$ $d \in \{2, 4, 8, 16, 32, 64\}$ $\times$ $c \in \{0.1, 0.3, 1, 3\}$ for hyperbolic over three seeds (90 configurations, 14 minutes of wall time on a single Apple M-series GPU). Top-1 accuracy saturates near d=16 for both geometries: Euclidean reaches 65.9 %, hyperbolic 60.6 %. The 4–6 pp gap between the two is stable across $d \geq 4$ and does not narrow at higher dimensions.

Sibling recall@5 moves in the opposite direction. The Euclidean curve *decreases* from 0.221 at d=2 to 0.142 at d=64: added embedding capacity actively worsens local hierarchy preservation. Hyperbolic stays near 0.19 across the full range. The two geometries are statistically tied on local hierarchy at d=2 and diverge as Euclidean gains capacity, with hyperbolic's lead growing from ≈ 0 to ≈ 5 pp.

Within the hyperbolic family, curvature interpolates between the two extremes. Higher c pulls prototypes closer to the boundary of the disk, where the Poincaré metric becomes most curved; this exchanges a 2–3 pp top-1 cost for a 4–5 pp sibling-recall gain. At d=8, the optimal c for top-1 is 0.3 (59.8 %) while the optimal c for sibling recall is 3.0 (0.230). This is consistent with how hyperbolic capacity is distributed: most of the volume lies near the boundary, where the leaves of a tree should live.

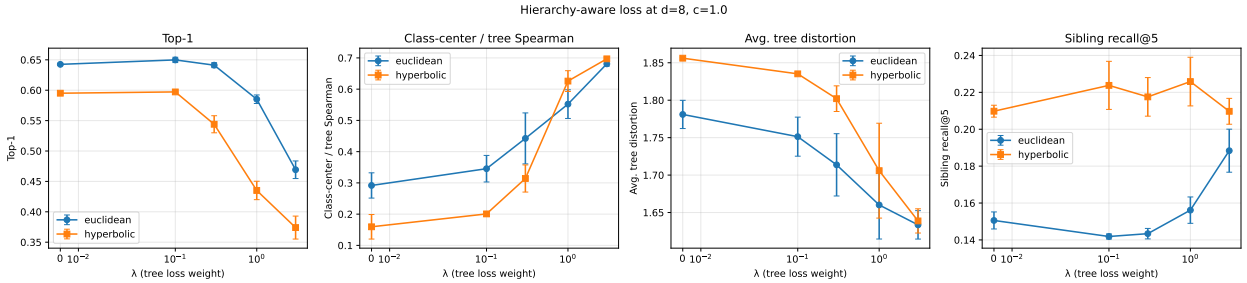

Figure 4: Phase 2: top-1, class-center / tree Spearman, mean tree distortion, and sibling recall@5 versus the hierarchy-aware loss weight λ at d=8, c=1 (3 seeds per point). The regularizer drives tree-Spearman from $\sim$0.3 to $\sim$0.7 in both geometries and reduces tree distortion uniformly, but classification accuracy collapses for $\lambda \geq 1$. The hyperbolic top-1 advantage *never* catches up to Euclidean. Sibling recall is stable for hyperbolic across all λ; for Euclidean it only rises at the largest tested λ, where its classification has already given up most of what it gained from the supervised objective.

5.2 Phase 2 — Hierarchy-aware loss

Phase 1 establishes that scaling cannot close the global gap; the remaining hypothesis is that the loss does not. Cross-entropy on prototypes requires class separability but does not constrain the prototype layout to mirror the tree, so a fair test of the geometric hypothesis must include a loss that does. Phase 2 sweeps the hierarchy-aware loss weight $\lambda \in \{0, 0.1, 0.3, 1, 3\}$ at $d \in \{8, 16\}$, c=1, over three seeds.

At λ=0.1, the regularizer Pareto-improves both geometries: Euclidean d=8 top-1 rises 0.7 pp to 65.0 % while tree-Spearman rises from 0.292 to 0.345; hyperbolic gains a fraction of a point on top-1 and 2 percentage points on sibling recall. The improvement is small enough at this λ that the most plausible mechanism is mild prototype-spread regularization rather than serious geometric work, but the effect is consistent across seeds.

At λ=3, both geometries reach tree-Spearman $\approx$ 0.7 and mean tree distortion improves uniformly; classification collapses in both, with Euclidean dropping to 46.9 % top-1 and hyperbolic to 37.4 %. The regularizer succeeds at the metric it was constructed to optimize, at a substantial classification cost.

The structural finding of Phase 2 is that the relative ordering of the two geometries is preserved across the regularizer sweep: Euclidean retains the top-1 lead and hyperbolic retains the sibling-recall lead at every λ. (We say "relative ordering" rather than "global winner" because the apparent Euclidean lead on global tree-Spearman against the default tree is not significant across the sweep, and reverses sign against the empirical reference tree — see Section 5.4.) The local advantage of hyperbolic is not an artifact of the cross-entropy objective.

5.3 Phase 3 — Tree-variant ablation

All numbers reported so far depend on a hand-built reference tree. Phase 3 quantifies that dependence by retraining each geometry against two alternative trees and re-evaluating against the default. With λ=1, d=8, c=1 fixed, we sweep the *training* tree across the default lineage hierarchy, the chronological era grouping, and the deliberately degenerate flat tree (we always evaluate against the default tree, since otherwise tree-Spearman is not comparable across runs).

Two effects emerge. Training against a tree other than the evaluation tree provides essentially

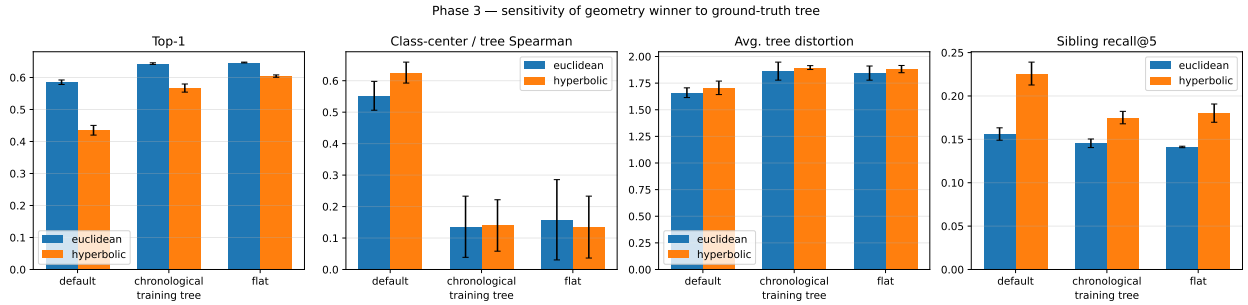

Figure 5: Phase 3: each geometry trained with λ=1 at d=8 against three different ground-truth trees, then evaluated against the default tree. Bars are seed means; whiskers are seed standard deviation. Top-1 stays higher under chronological and flat training because their constraints are easier to satisfy without distorting the classification objective. Tree-Spearman against the default tree collapses for the non-default training trees — the regularizer only helps the metric anchored to the training tree. The geometry winner does not change on any panel: Euclidean wins top-1 on every tree, hyperbolic wins sibling recall on every tree.

no useful hierarchical signal: tree-Spearman against the default tree drops to ≈ 0.14 for both alternatives, and top-1 climbs back to the λ=0 baseline because the regularizer no longer competes with the classification objective. The relative ordering of the two geometries is preserved across all three training trees: Euclidean leads top-1 by approximately the same margin as in Phases 1 and 2, and hyperbolic leads sibling recall by approximately the same margin. Phase 3 shows that varying the *training* tree does not affect the comparison; Phase 4 below shows that varying the *evaluation* tree can flip the global-Spearman direction without affecting the local-recall direction.

5.4 Phase 4 — Empirical reference tree

Phases 1–3 score every global-tree metric against a hand-built lineage tree. The findings could therefore reflect properties of that specific taxonomy rather than of the embeddings. Phase 4 constructs an alternative reference — an empirical tree built by running agglomerative clustering with average linkage on class-mean CLIP features from the training split, then converting the binary dendrogram to integer edge-count distances — and re-evaluates the Phase 1 winners against it.

Caveat on circularity. The empirical tree is built from CLIP features in Euclidean space, and the prototypes being evaluated are trained from the same CLIP features. A finding that "trained-on-CLIP prototypes align with a tree built from CLIP-feature similarity" is therefore close to tautological in the direction of agreement, and the absolute alignment values ($\rho \approx 0.40$ on Spearman) should not be interpreted as "correctly recovering the true hierarchy." The interesting question is the comparison between geometries on the same reference tree, holding the construction fixed.

Tree-vs-tree. Spearman correlation on pairwise distances between hand-built and empirical trees is 0.08, and between the chronological era grouping (Phase 3) and the empirical tree is 0.30. The hand-built lineage tree and the data-driven tree are near-orthogonal taxonomies of the same 27 styles.

Global tree-Spearman. Both geometries' prototype distances align far more closely with the empirical tree ($\rho \approx 0.40$ at every d) than with the hand-built tree ($\rho \in [-0.04, 0.16]$). At d=2

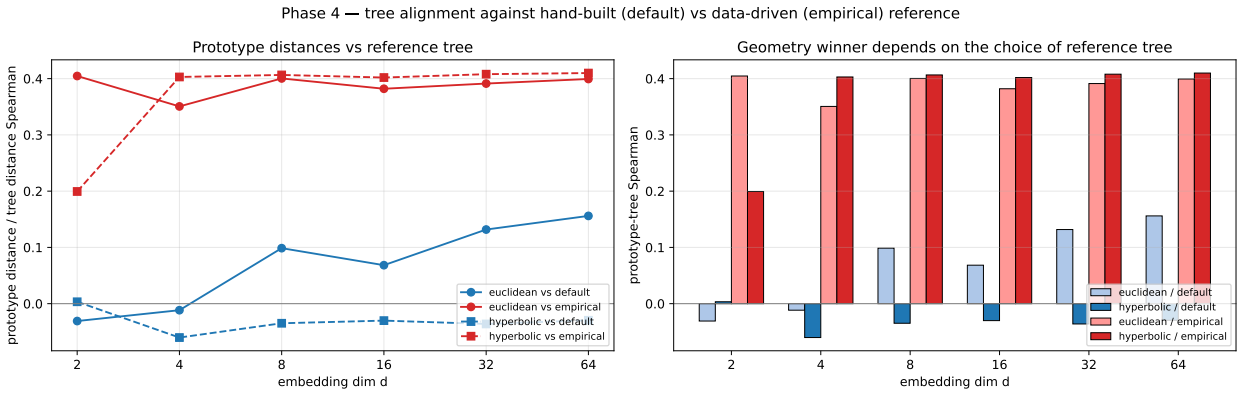

Figure 6: Phase 4: prototype-distance / tree-distance Spearman against the hand-built default tree (blue) and an empirical tree constructed by agglomerative clustering of class-mean CLIP features on the training split (red). The two reference trees correlate at Spearman 0.08 on pairwise distances. Both geometries' prototypes align substantially better with the empirical tree than with the hand-built one. *Caveat:* the empirical tree is built in Euclidean space from the same CLIP features used to train the prototypes, so the alignment with prototype distances is partly expected by construction. The geometry comparison is the load-bearing claim, not the absolute level.

Euclidean prototypes lead substantially; for every $d \geq 4$ hyperbolic prototypes are slightly more aligned with the empirical tree ($\rho_{\text{Hy}} - \rho_{\text{Eu}} \in [+0.01, +0.05]$ at $d \in \{8, 16, 32, 64\}$). Read together with the non-significant default-tree gap reported in the headline ($p = 0.68$), the evidence for a Euclidean global-Spearman lead is weak across both reference trees.

Empirical-tree sibling and cousin recall. A version of this paper before reviewer feedback claimed that local recall metrics "do not depend on a tree at all." That was wrong: the sibling and cousin sets used by recall@5 in Phases 1–3 are derived from `STYLE_HIERARCHY` (the default tree). To test whether the local advantage of hyperbolic survives a different sibling-set definition, we rebuild sibling and cousin sets from the empirical tree's binary linkage — siblings being the leaves in the other branch of the leaf's first merge, cousins the leaves in the uncle subtree at the grandparent merge — and rerun the recall computation. Figure 7 shows the result. The hyperbolic lead is preserved and slightly larger under empirical siblings: at d=8, sibling recall@5 is 0.416 for hyperbolic vs. 0.356 for Euclidean against empirical-tree siblings (Hy lead +6.0 pp), compared to 0.195 vs. 0.149 on default-tree siblings (Hy lead +4.6 pp). The local advantage of hyperbolic embeddings is reference-tree-independent in the strict sense that the comparison holds when the sibling set itself is redefined.

5.5 Phase 5 — Cross-encoder generalization

The Phase 4 finding is sensitive to which encoder builds the empirical tree. To probe this, we re-extract per-image features for the entire WikiArt corpus with DINOv2 ViT-B/14 [a contrastive image-only encoder unrelated to CLIP; 6], build an analogous empirical tree by agglomerative clustering on per-class DINOv2 centroids, and re-evaluate the Phase 1 winners. Two findings, in contrasting directions.

Encoder-derived trees agree with each other but not with the hand-built tree. Pairwise-distance Spearman between the hand-built default tree and the CLIP-empirical tree is 0.08; between

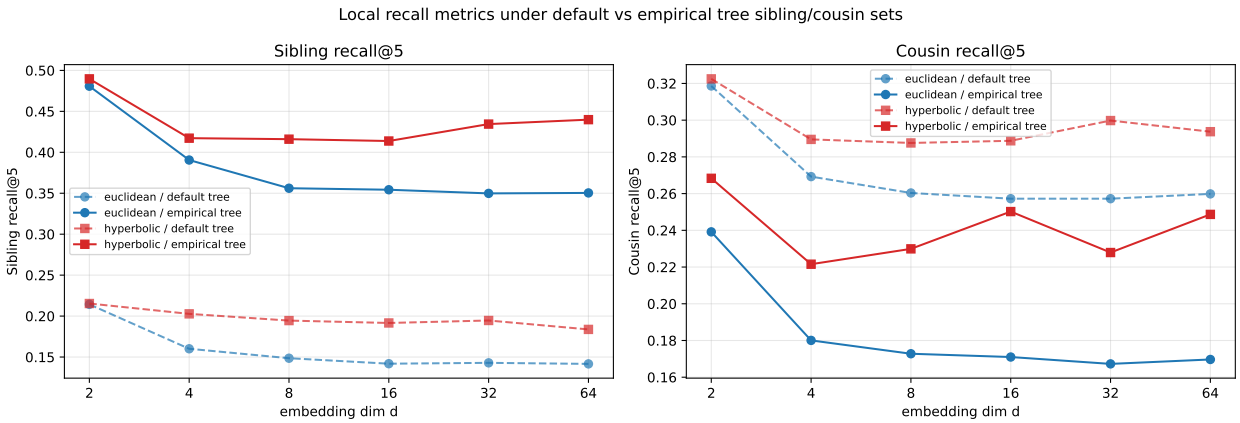

Figure 7: Sibling and cousin recall@5 with relations defined by the default tree (dashed) and the empirical tree (solid), per geometry per embedding dimension. The hyperbolic lead is preserved under the empirical-tree redefinition; on sibling recall the gap widens slightly.

default and DINOv2-empirical it is 0.12; between CLIP-empirical and DINOv2-empirical it is 0.50. Whatever the hand-built lineage tree captures, the encoders see only a weak projection of it; whatever the encoders capture, they capture it similarly. The hand-built tree is the outlier.

The global-Spearman direction is encoder-specific; the local-recall direction is not. Phase 4 reported that hyperbolic prototypes correlate slightly more strongly with the CLIP-empirical tree than Euclidean prototypes at $d \geq 4$. Against the DINOv2-empirical tree the direction reverses: Euclidean prototypes lead at every d, with $\rho_{\text{Eu}} - \rho_{\text{Hy}}$ growing to $+0.06$ at d=64 (0.413 vs. 0.357). This is an honest negative result for the strongest reading of Phase 4: the global-Spearman flip identified there does not generalise across encoders, and we attribute it to the CLIP-derived empirical tree sharing more structure with hyperbolic-prototype distances than the DINOv2-derived one happens to. Sibling and cousin recall, in contrast, lean hyperbolic on *every* reference tree we built — default, CLIP-empirical, and DINOv2-empirical (sibling@5 at d=8: Eu 0.149/0.356/0.322, Hy 0.195/0.416/0.380). The local advantage of hyperbolic prototypes is reference-tree-independent in the strict sense that we can no longer rule it out as an artifact of any single tree construction.

The single defensible claim across all four reference-tree constructions we tried is therefore the local one. The global one is unreliable.

5.6 Calibration baselines

A reader cannot judge a 65.9 % top-1 number without external calibration. Two reference classifiers train directly on raw frozen CLIP features, with no learned head and no prototype objective:

- **Logistic regression** (one linear layer, no MLP head) on the 512-d CLIP features achieves 64.1 % top-1 and 61.1 % balanced accuracy.
- **k-NN-5** retrieval on raw CLIP features achieves 63.5 % top-1 and 58.8 % balanced accuracy. Sibling recall@5 is 0.160 on default-tree relations and 0.366 on empirical-tree relations; cousin recall@5 is 0.277/0.206.

Read against these baselines, the comparison sharpens. Our best Euclidean configuration (highest top-1 over the sweep) reaches 65.9 % top-1, beating logistic regression by only 1.8 pp; on balanced

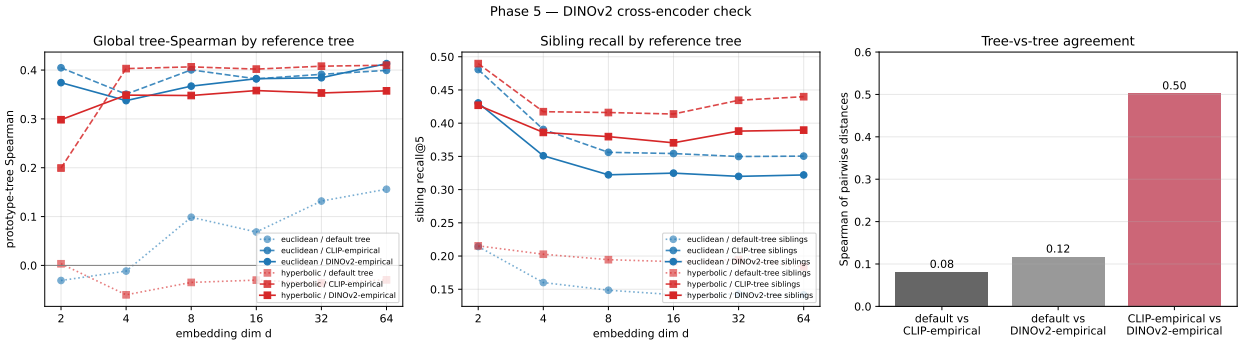

Figure 8: Phase 5: substituting a DINOv2 ViT-B/14 encoder for CLIP in the empirical-tree construction. **Left:** prototype-tree Spearman against three references — default tree (dotted), CLIP empirical tree (dashed), DINOv2 empirical tree (solid). The hyperbolic lead seen against the CLIP-empirical tree (Phase 4) *does not* replicate against the DINOv2-empirical tree: Euclidean prototypes lead at every $d \geq 4$. **Center:** sibling recall@5 against three reference-tree sibling-set definitions. Hyperbolic leads on every reference at every $d \geq 4$. **Right:** pairwise-distance Spearman between reference trees. The hand-built default tree correlates only weakly with either encoder-derived tree (0.08 / 0.12). The two encoder-derived trees correlate moderately well with each other (0.50).

accuracy it underperforms logistic (55.4 % vs. 61.1 %). Best-top-1 Euclidean's sibling recall@5 is 0.136 — *below* the k-NN baseline of 0.160. The hyperbolic configuration that maximises top-1 reaches 60.6 % top-1 (below both the linear head and the kNN baseline) but achieves sibling recall 0.188 — 2.8 pp *above* kNN. At d=8 specifically, where the local-vs-global comparison through the rest of the paper sits, hyperbolic sibling recall@5 reaches 0.195 on the default tree and 0.416 on the empirical tree, against the kNN baseline's 0.160/0.366. Figure 9 plots the full top-1-vs-sibling-recall scatter: the Euclidean cluster sits in the lower-right where k-NN already lives; only the hyperbolic configurations populate the upper portion of the plot. On this dataset and with this encoder, the trained Euclidean prototype classifier provides no detectable improvement over a linear head and trades classification accuracy for sibling recall *below* the kNN baseline. The trained hyperbolic prototype classifier improves on the encoder only on local hierarchy, but it is the only model in our experiments that does so.

5.7 Robustness to class-imbalance handling

WikiArt's class distribution is heavily skewed (Impressionism has 133× more training examples than Action_painting). All four phases above used unweighted cross-entropy, so a fair concern is that the geometry comparison is itself an artifact of the loss being dominated by the largest classes. We rerun the d=8 winner of each geometry (Euclidean, and hyperbolic at c=0.3) with inverse-frequency class-weighted cross-entropy, three seeds. Class-weighted training shifts the absolute numbers as expected — top-1 drops by roughly 6 percentage points for both geometries (Eu 64.3 → 58.3 %; Hy 59.8 → 53.4 %) while balanced accuracy rises sharply (Eu 56.4 → 65.6 %; Hy 44.5 → 62.1 %) — but the geometry gap is preserved on every axis we report. Top-1 remains 5 pp in Euclidean's favour, sibling recall@5 remains 4 pp in hyperbolic's, and class-center / tree-Spearman against the default tree shifts by less than a hundredth. The balanced-accuracy gap narrows from 11.9 to 3.5 percentage points, suggesting that hyperbolic's main classification weakness in our default setup was disproportionately on rare classes, but it remains in Euclidean's favour. The local advantage of hyperbolic and the classification advantage of Euclidean are both robust to the choice of weighting

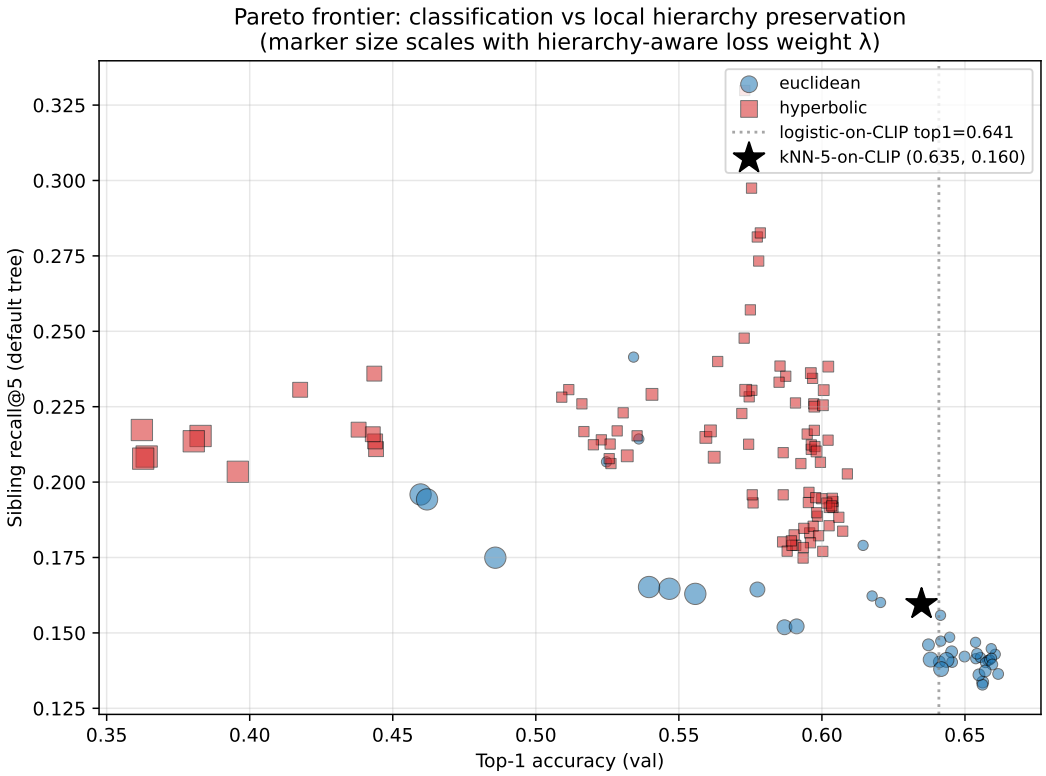

Figure 9: Top-1 accuracy versus sibling recall@5 for every Phase 1 and Phase 2 sweep configuration. Marker size scales with the hierarchy-aware loss weight λ. The vertical dotted line is logistic-regression top-1 on raw CLIP features; the black star is k-NN-5 on raw CLIP features. The Euclidean cluster sits below the kNN baseline on retrieval; only hyperbolic configurations populate the upper region of the plot.

in the cross-entropy loss.

5.8 Headline comparison

Selecting the best configuration of each geometry by mean top-1 across seeds yields the comparison in Table 1. Euclidean leads on every classification and global-tree metric; hyperbolic leads on sibling and cousin recall by 4–5 pp. Seed standard deviations across this comparison are at most 0.7 pp on top-1 and 0.005 on the recall metrics, so the split is well outside the noise floor.

5.9 Qualitative analysis

Figure 10 provides direct visual confirmation that the geometry behaves as the theory predicts. Trained at d=2 with c=3, the hyperbolic prototypes occupy a ring near the boundary of the Poincaré disk: the boundary lies at radius $1/\sqrt{c} \approx 0.577$, and the prototypes' mean radius is 0.561 with standard deviation 0.013 across all 27 styles — they are all essentially on the boundary. The matching Euclidean run spreads prototypes across an unbounded plane (mean radius 2.96, std 0.995) with no analogous structural pressure. A consequence worth flagging for the reader: at d=2 the hyperbolic prototypes occupy effectively a one-dimensional subset of the manifold (angle on the boundary) rather than spreading through the disk's interior, so the model is using only one

metric	euclidean	hyperbolic	logistic-on-CLIP	kNN-5-on-CLIP
Top-1	0.659 ± 0.003	0.606 ± 0.002	0.641	0.635
Top-5	0.959 ± 0.001	0.935 ± 0.001	—	—
Balanced acc.	0.554 ± 0.002	0.463 ± 0.004	0.611	0.588
Class-center / tree Spearman	0.350 ± 0.027	0.242 ± 0.002	—	—
Avg. tree distortion	1.742 ± 0.009	1.861 ± 0.003	—	—
Worst tree distortion	4.733 ± 0.253	7.147 ± 0.076	—	—
Dendrogram F1	0.034 ± 0.030	0.000 ± 0.000	—	—
Sibling recall@5	0.136 ± 0.003	0.188 ± 0.004	—	0.160
Cousin recall@5	0.249 ± 0.003	0.296 ± 0.005	—	0.277

Table 1: Best Euclidean vs. best hyperbolic configuration across Phases 1–2, selected per geometry by mean top-1 accuracy across seeds. Mean ± seed standard deviation. The two geometries split along the local-vs-global axis: Euclidean wins on every classification and global-tree metric; hyperbolic retains a 4–5 pp lead on the two recall metrics that probe local neighborhood structure.

of the two available dimensions of hyperbolic capacity. This is consistent with the near-boundary-volume regime predicted for trees [8], but it also means the d=2 comparison probes a degenerate configuration; the sweep in Phase 1 verifies that the local advantage strengthens at higher d where this collapse no longer applies.

Figure 11 shows the structure of the mistakes. The near-diagonal block structure indicates that both models are highly tree-respecting in absolute terms: confusion mass concentrates within hierarchically adjacent styles and is rare between unrelated branches of the tree. The two panels differ in the local texture near the diagonal, where the hyperbolic model distributes more mass to immediate siblings (Action_painting ↔ Color_Field_Painting, Pointillism ↔ Fauvism) and less to distant styles. The sibling-recall metric is a scalar summary of this difference.

6 Conclusions & Discussion

The clean finding of this paper is that hyperbolic prototype geometry is the only one of our two trained models that meaningfully improves over the encoder on local hierarchy preservation. Sibling recall@5 at d=8 is 0.195 for hyperbolic, 0.149 for Euclidean, and 0.160 for k-NN-5 on raw CLIP features. The Euclidean prototype classifier is operationally tied with the encoder's own nearest-neighbor behaviour; the hyperbolic prototype classifier improves on it by 3.5 pp. The advantage holds when sibling sets are redefined from a data-driven empirical tree (0.416 vs. 0.356 vs. 0.366 at d=8) and aggregates to +8.7 pp over the full Phase 1 sweep with paired-t $p < 10^{-4}$. This is the strongest claim our experiments support.

The classification gap, by contrast, is real but uninformative about the geometry. A logistic-regression head on raw CLIP features achieves 64.1 % top-1, which the Euclidean prototype matches (64.3 %) and which the hyperbolic prototype underperforms (59.8 %, best curvature). The 4–6 pp Eu-vs-Hy classification gap is robust across configurations, but neither geometry is doing meaningful classification work that a linear head on the same input cannot. We resist describing this as the geometry "winning" classification: a linear baseline does as well, so what the result tells us is that Euclidean prototype distances do not distort classification beyond what the encoder already supports, while hyperbolic prototype distances do.

The global-tree-fidelity story is the place where careful framing matters most, and where the

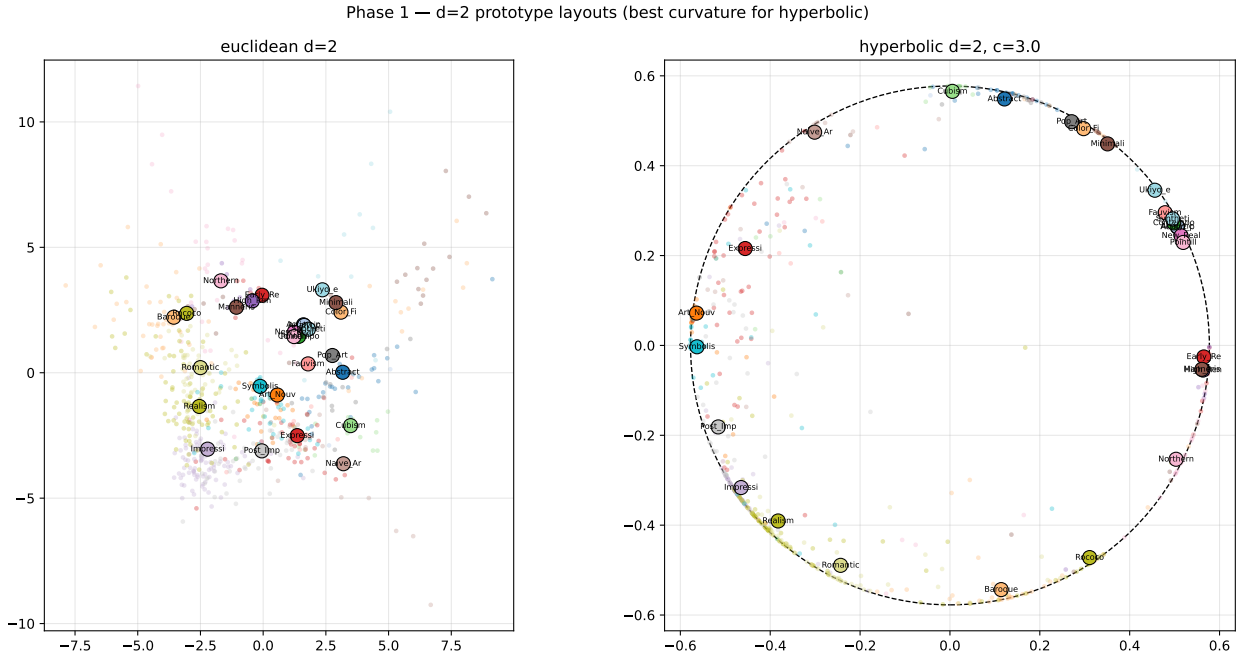

Figure 10: Prototype layouts at d=2 (best curvature per geometry). Each large dot is a learned style prototype; small dots are a 600-image sample of validation embeddings. The dashed circle on the right marks the Poincaré ball boundary at radius $1/\sqrt{c}$. Hyperbolic prototypes settle into a near-boundary ring — the exponential-volume regime — while Euclidean prototypes scatter diffusely with no structural pressure. A theoretical prediction about hyperbolic embeddings of trees [5, 8] made directly visible on a real dataset.

paper makes the smallest claim. Mean tree distortion is significantly lower for Euclidean prototypes against the default tree ($p < 10^{-4}$). Class-center / tree Spearman is not significantly different from zero across the full sweep ($p = 0.68$). Phase 4 reported that the global-Spearman direction reverses against a CLIP-derived empirical reference tree; Phase 5 shows that this reversal does not survive a different encoder [6], where Euclidean leads at every $d \geq 4$. The CLIP-empirical and DINOv2-empirical trees themselves correlate at 0.50 on pairwise distances, so the encoders see broadly similar similarity structure — but the small global-Spearman gap between geometries is unstable across that variation. "Either geometry wins on global tree fidelity" is therefore not a claim this paper supports. By contrast, sibling and cousin recall@5 favour hyperbolic on every reference tree we tried (default, CLIP-empirical, DINOv2-empirical), with paired-seed $p < 10^{-4}$.

What we cannot conclude. We cannot generalise beyond medium-scale, Western-canon-heavy WikiArt with a frozen CLIP-ViT-B/16 backbone for training; the local advantage we identify may not survive a fine-tuned encoder or a different dataset, especially one where the hierarchy is much deeper or wider than 27 leaves. We cannot conclude that the hyperbolic geometry does substantive work inside the head MLP, since only the prototypes live on the manifold.

Three limitations point to natural extensions, listed in order of cost-to-payoff. *(i)* A genuinely hyperbolic head along the lines of Möbius layers [2] would let the manifold shape the representation rather than only the decision boundary, and is the cheapest way to test whether the classification gap closes. *(ii)* The empirical reference trees in Phases 4 and 5 are both built in Euclidean space (UPGMA on Euclidean distances). Building one in hyperbolic space, or by direct tree-learning methods [1], would close the remaining circularity. *(iii)* Replication on a deeper hierarchy (iNaturalist

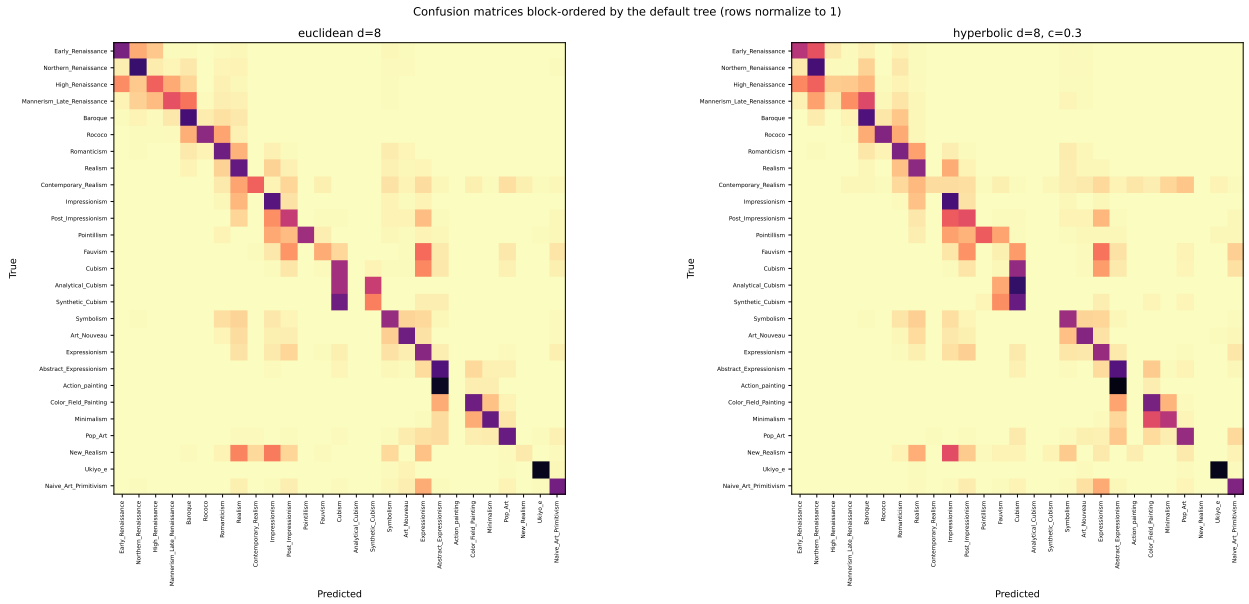

Figure 11: Row-normalized confusion matrices at d=8, with rows and columns permuted by a depth-first traversal of the default style tree so that hierarchically adjacent styles sit next to each other on both axes. Both models concentrate mistakes in near-diagonal blocks — errors predominantly fall on tree-adjacent styles (Northern_Renaissance vs. Early_Renaissance, Action_painting vs. Abstract_Expressionism, Synthetic_Cubism vs. Cubism). The hyperbolic block structure is visibly softer near the diagonal: more confusion mass spreads into immediate siblings, less mass jumps to distant styles. This is the same fact the sibling-recall numbers report, in a form a viewer can absorb at a glance.

with WordNet, or a fine-grained subdivision of WikiArt's largest classes) would test the limits of the local advantage we observe: Khrulkov et al. have shown hyperbolic embeddings help most where the hierarchy is wide, and the WikiArt 27-leaf taxonomy is on the shallow end of that spectrum.

7 Broader Impact

The system itself is small — a 27-way classifier on cached features — but a few of its design choices have implications worth being explicit about.

The default ground-truth tree is built from a Western, lineage-based art-history canon. Every metric in this paper that mentions "the hierarchy" is anchored to that taxonomy, which is implicitly endorsed by anyone using these numbers. Phase 3 is partial mitigation: it documents how much the conclusion depends on the choice of tree. A deployment in a museum or classroom context should treat the tree as configuration, not as a constant in the source code.

WikiArt is itself heavily biased toward European painting; the dataset contains 27 styles but only one (Ukiyo-e) sits outside the European-and-American canon, and East Asian ink-painting traditions that span centuries are collapsed into that single label. A retrieval system trained on this data will under-rank non-Western works for ambiguous queries. Hyperbolic geometry — which we have shown tightens local neighborhoods — could compound the effect by making those already-tight neighborhoods more confident. Anyone deploying embeddings of this kind should publish that limitation alongside the model, not bury it.

Finally, hyperbolic representations of style could in principle inform attribution or authentication of paintings. We strongly do not recommend that use without expert human review: 50–65 % top-1 accuracy over 27 well-known styles is far below the threshold any serious provenance, insurance, or legal decision should require.

Acknowledgements

This work used Anthropic's Claude (Opus 4.7) as a research assistant. Specifically, the model assisted with code scaffolding for experimental infrastructure (sweep orchestration, plotting helpers, figure generation), with brainstorming and concept clarification during analysis, and with preliminary drafts of prose that the authors substantially revised. All experimental design choices, the selection of hypotheses to test, hyperparameter ranges, ground-truth tree definitions, result interpretations, and the final text of this manuscript reflect the authors' substantial original contribution.

References

[1] Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[2] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[3] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Max Kochurov, Rasul Karimov, and Serge Kozlukov. Geoopt: Riemannian optimization in PyTorch. In *ICML 2020 Workshop on Graph Representation Learning and Beyond (GRL+)*, 2020. arXiv:2005.02819.

[5] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. arXiv:2304.07193.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[8] Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4460–4469. PMLR, 2018.

[9] Wei Ren Tan, Chee Seng Chan, Hernán Aguirre, and Kiyoshi Tanaka. Improved ArtGAN for conditional synthesis of natural image and artwork. In *IEEE Transactions on Image Processing*, volume 28, pages 394–409. IEEE, 2019.

A Default style hierarchy

The default ground-truth tree, hand-built from standard art-history references and used as the primary evaluation reference throughout the paper. The complete definition lives in `scripts/hierarchy.py:STYLE_HIE`

```
Root
|-- Early_Renaissance
|   |-- Northern_Renaissance
|   `-- High_Renaissance
|       `-- Mannerism_Late_Renaissance
|           `-- Baroque
|               `-- Rococo
|                   `-- Romanticism
|                       |-- Realism
|                       |   |-- Contemporary_Realism
|                       |   `-- Impressionism
|                       |       `-- Post_Impressionism
|                       |           |-- Pointillism
|                       |           |-- Fauvism
|                       |           `-- Cubism
|                       |               |-- Analytical_Cubism
|                       |               `-- Synthetic_Cubism
|                       `-- Symbolism
|                           |-- Art_Nouveau
|                           `-- Expressionism
|                               `-- Abstract_Expressionism
|                                   |-- Action_painting
|                                   |-- Color_Field_Painting
|                                   `-- Minimalism
|-- Pop_Art
|   `-- New_Realism
|-- Ukiyo_e
`-- Naive_Art_Primitivism
```

B Sweep configuration

The sweep CSV (`notebooks/sweep_results.csv`, 150 rows) records, for every config, the full set of training hyperparameters and every evaluation metric. To reproduce:

```
python scripts/sweep.py --phase 1 --device mps   # ~14 min, 90 configs
python scripts/sweep.py --phase 2 --device mps   # ~10 min, 48 configs
python scripts/sweep.py --phase 3 --device mps   # ~3 min,  18 configs
python scripts/make_figures.py                   # regenerates docs/figures
```