NEW HYPERBOLIC ENTAILMENT CONES BEYOND THE CAPACITY–FEASIBILITY TRADE-OFF

Anonymous authors
Paper under double-blind review

ABSTRACT

Entailment cones in hyperbolic space (Ganea et al., 2018b) represent the partial order of a hierarchy as containment between cones, and serve as a foundation of recent hyperbolic vision-language models. However, the cones' aperture cannot be defined everywhere; there is necessarily an infeasible ball region around the origin. The size of the infeasible ball trades off against the capacity (total aperture) of the cones, and at both ends the performance of the entailment cones degrades significantly. We first formalize this capacity–feasibility trade-off and show that this trade-off extends to any cones satisfying transitivity, i.e., the trade-off is an unavoidable problem in the original formulation. To overcome this trade-off, we then introduce (ε, δ)-transitivity, which reformulates transitivity with two parameters: the radius of an infeasible region and an allowance for violations of transitivity. Building on this, we propose cones that can be defined everywhere except the origin, have at least the same capacity, and generalize the original entailment cones. On hierarchical word embeddings (WordNet, MCG, Hearst), the proposed method outperforms the original entailment cones in 10 of 12 settings, with gains of up to 27.6 points. On the hyperbolic vision-language model MERU, it preserves accuracy on zero-shot classification while yielding embeddings that better satisfy the entailment condition of the original formulation.

1 INTRODUCTION

Hierarchically structured data appear frequently in machine learning, such as is-a relations between words or concepts and containment relations between images and text. Hyperbolic space embeds hierarchical structures more faithfully than Euclidean space, as its volume grows exponentially with the radius, matching the exponential growth of the number of nodes with depth in trees (Sarkar, 2011; Nickel & Kiela, 2017). Since the great success of Poincaré embeddings (Nickel & Kiela, 2017), hyperbolic machine learning has been extended to a wide range of fields, including hierarchical embeddings (Nickel & Kiela, 2018; Sala et al., 2018), neural network architectures (Ganea et al., 2018a; Shimizu et al., 2021), graph learning (Dai et al., 2021; Sun et al., 2021), image representations (Khrulkov et al., 2020), probabilistic models (Nagano et al., 2019), and vision-language models (Desai et al., 2023; Pal et al., 2025), mainly on the Poincaré ball model or the Lorentz model.

What matters in representation learning of hierarchies is the preservation of the partial order. In a hierarchy, if x is an ancestor of y and y is an ancestor of z, then x must be an ancestor of z. This property is known as **transitivity**. *Entailment cones* (Ganea et al., 2018b) enable learning partial orders by assigning a cone to each point and representing the order as containment between cones: the cone of a descendant lies entirely inside the cone of its ancestor and never leaks out of it. For this containment to hold throughout the hierarchy, the apertures of the cones must be governed by a single non-increasing function of the radius. Consequently, cones near the origin have large apertures while cones far from the origin have small apertures, which matches the fact that the most abstract concepts entail many descendant concepts. To learn the abstract-to-concrete order between texts and images, the entailment cones have recently been applied to hyperbolic vision-language models (Desai et al., 2023; Ramasinghe et al., 2024; Pal et al., 2025; Yoshikawa & Matsubara, 2026).

Despite this contribution, the performance of the entailment cones depends strongly on the choice of the non-increasing function controlling the apertures. This choice determines how fast the apertures

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.

Table 1: F1 (%) when sweeping K on Hearst-25k (hypernym pairs extracted by Hearst patterns, §6.1), 50% non-basic edges, $d = 5$, 3 seeds. Bold: row best. Full sweeps: Appendix J.1.

K	10^{-4}	$3{\cdot}10^{-4}$	10^{-3}	$3{\cdot}10^{-3}$	10^{-2}	$3{\cdot}10^{-2}$	10^{-1}	$3{\cdot}10^{-1}$	1	3	10
$(\varepsilon, 0)$ cones	0.3	1.3	4.5	11.6	12.7	29.5	**52.7**	44.2	16.4	0.9	0.0
$(0, \delta_\infty)$ cones (**Ours**)	0.1	1.5	2.1	15.3	21.3	66.2	67.3	**68.0**	67.9	67.7	67.6

shrink with the radius, and in the setting of Ganea et al. (2018b) the function is set to the constant value K. However, as discussed by Ganea et al. (2018b), this constant choice necessarily produces an infeasible ball region around the origin, whose radius also grows with K. For small K, the infeasible ball region is small, but the apertures shrink rapidly with the radius, and it becomes difficult to contain descendant cones inside their ancestor cones. For large K, the apertures are large overall, but the infeasible ball region occupies a large part of the space, and the feasible region becomes critically small. Optimization thus becomes difficult at both ends of the trade-off. Table 1 shows this phenomenon empirically: the conventional entailment cones perform well only at intermediate values of K, and their performance drops rapidly as K is made either larger or smaller, down to 0.0 at $K = 10$. Although the trade-off between *capacity* (how wide the cones are) and *feasibility* (the region where the cones are defined) affects embedding quality, resolving this trade-off has remained open (Ganea et al., 2018b; Yu et al., 2024; Huynh et al., 2026).

In this paper, we show that this trade-off holds for *every* entailment cone that satisfies transitivity exactly, and that every such cone necessarily has an infeasible ball region near the origin. This trade-off arises from transitivity itself: transitivity forces the function controlling the aperture to be non-increasing, and the trade-off follows from this monotonicity.

We then ask whether entailment cones can avoid the capacity–feasibility trade-off, which crucially harms optimization. We introduce (ε, δ)**-transitivity**, a new framework for transitivity with two parameters: the radius ε of the infeasible region and an allowance δ for violations of transitivity. This formulation can be regarded as a natural generalization of the conventional exact transitivity, as $(\varepsilon, 0)$-transitive cones allow no violation but have an infeasible region, which matches the definition of Ganea et al. (2018b); we therefore call the entailment cones the $(\varepsilon, 0)$ cones hereafter. In contrast, we focus on $(0, \delta)$-transitive cones, which allow descendant cones to leak out of their ancestor cones and have no infeasible region. Setting $\varepsilon = 0$ fixes feasibility to the whole domain except the origin, so that the trade-off is no longer between capacity and feasibility but between capacity and the allowance for violations, whose effect on optimization is bounded. Under this setting, we propose one-parameter $(0, \delta_q)$-transitive cones, where the parameter q controls the tail of the violations. In our parametrization, the budget δ_q concentrates in the former infeasible region. In the limit $q \to \infty$, the proposed $(0, \delta_\infty)$ cones confine the violations of transitivity entirely to the former infeasible region and coincide with the $(\varepsilon, 0)$ cones outside it. Therefore, the $(0, \delta_\infty)$ cones move the capacity–feasibility trade-off of the $(\varepsilon, 0)$ cones to a trade-off between capacity and the violation of transitivity occurring only inside the infeasible region. Table 1 shows an example of this effect: the $(0, \delta_\infty)$ cones outperform the conventional $(\varepsilon, 0)$ cones for $K \geq 3 \cdot 10^{-3}$, have no degradation on the large-K side, and maintain 67.6 at $K = 10$, while their degradation on the small-K side is shared with the $(\varepsilon, 0)$ cones because the capacity tends to 0.

In experiments on hierarchical word embeddings (WordNet, MCG, Hearst), the proposed $(0, \delta_q)$ cones outperform the original $(\varepsilon, 0)$ cones in 10 of 12 settings, and the gain reaches 27.6 points (77.0 versus 49.4) on a Hearst-based dataset. Finally, replacing the entailment loss of the hyperbolic vision-language model MERU with the proposed $(0, \delta_q)$ cones preserves accuracy while yielding embeddings that better satisfy the entailment condition of the original formulation. Replacing one side of the trade-off with one whose effect on optimization is bounded thus yields better embeddings.

2 Background

2.1 Related work

Hyperbolic embeddings of hierarchies have been actively studied since Poincaré embeddings (Nickel & Kiela, 2017), with refinements based on the Lorentz model (Nickel & Kiela, 2018).

These methods learn hierarchies with distance-based losses, but distance is a symmetric quantity and cannot directly express the partial order between hypernyms and hyponyms. To learn this partial order explicitly, following order embeddings in Euclidean space (Vendrov et al., 2016), Ganea et al. (2018b) proposed entailment cones in hyperbolic space. The entailment cones are the subject of this paper, and their definition and properties are given in the next section.

Shadow Cones (Yu et al., 2024) generalize entailment cones by a different geometric construction based on light sources and shadows. Among them, the penumbral formulation has an infeasible region corresponding to the radius of the light source, and is therefore a direct subject of the capacity–feasibility trade-off shown in this paper. The umbral formulation, on the other hand, uses a point light source, an object of finite volume, and a hypercycle boundary. It therefore belongs to a geometric class different from the entailment cones, and our theorems do not apply to it directly. The aim of this paper is to characterize the limitation of the entailment cones, which are widely used on the Poincaré ball and Lorentz models, and to relax the limitation while preserving the conditions of the cones and the existing losses.

The representation of partial orders by cones has also been deployed in vision-language models. MERU (Desai et al., 2023) extends CLIP to hyperbolic space and learns the partial order of abstract-to-concrete relationships between texts and images based on entailment cones. Since then, hyperbolic vision-language models such as HyCoCLIP (Pal et al., 2025), PHyCLIP (Yoshikawa & Matsubara, 2026), and ARGENT (Huynh et al., 2026) have been actively studied. However, most of these methods use the entailment cone in their losses without change. ARGENT mitigates the collapse of cones near the origin by an adaptive entailment loss and a norm regularizer, but keeps the aperture and hence the infeasible region. The problem of the infeasible region is therefore inherited by all current methods that build directly on entailment cones. In this work we propose a method that resolves this problem by a drop-in replacement alone.

2.2 Hyperbolic space and the Poincaré ball model

Hyperbolic space Hyperbolic space $\mathbb{H}^n$ is a Riemannian manifold whose sectional curvature is a constant negative value $-c$ ($c > 0$) everywhere, and several mutually isometric coordinate representations (models) are known. In this paper we mainly work in the Poincaré ball model; the formulation on the Lorentz model is given in Appendix K. Throughout, a **geodesic** is a curve connecting two points with minimal length.

Poincaré ball model The Poincaré ball model is the open unit ball $\mathbb{D}^n = \{x \in \mathbb{R}^n : \|x\| < 1\}$ equipped with the metric

$$g_x \;=\; \Big(\frac{2}{1-\|x\|^2}\Big)^2 g^E \tag{1}$$

where g^E is the Euclidean metric. Let O denote the origin and $r := \|x\|$ the Euclidean norm, and write the geodesic distance from O as the **hyperbolic norm** $\|x\|_{\mathbb{D}} := d(O,x) = 2\,\mathrm{arctanh}(r)$. The identities

$$\sinh\|x\|_{\mathbb{D}} = \frac{2r}{1-r^2}, \qquad \cosh\|x\|_{\mathbb{D}} = \frac{1+r^2}{1-r^2} \tag{2}$$

hold. Geodesics in this model are straight lines through the origin (diameters) and circular arcs orthogonal to the unit sphere.

3 Entailment Cones and the Latent Trade-off

We inherit the entailment loss of Ganea et al. (2018b) and address only the entailment cones. Starting from their definition, we show that transitivity imposes a trade-off between the size of the aperture and feasibility, first for the entailment cones. We then show this limitation extends to arbitrary transitive cones, and quantify it with the capacity C. Since we deal with quantities defined by integrals, the aperture functions ψ considered below are restricted to integrable ones.

Entailment cones are defined as geometric shapes satisfying the four conditions of (i) axial symmetry, (ii) rotation invariance, (iii) continuity, and (iv) transitivity. From these conditions, a cone $\mathfrak{S}_x^{\psi(x)}$ at a point $x \in \mathbb{D}^n \setminus \{O\}$ is defined as the geodesic cone with half-aperture $\psi(x)$ around the

outward axis extending from the origin O in the direction of x. By rotation invariance, ψ depends only on the radius $r = \|x\|$, so we write $\psi(r)$ hereafter. Here, transitivity means that whenever y belongs to the cone of x, the cone of y must be contained in the cone of x. More precisely, $y \in \mathfrak{S}_x^{\psi(x)} \Rightarrow \mathfrak{S}_y^{\psi(y)} \subseteq \mathfrak{S}_x^{\psi(x)}$ holds. Transitivity yields the following theorem (Ganea et al., 2018b, Theorem 3).

Theorem 3.1. *If transitivity holds, the function h defined by*

$$h : (0,1) \cap Dom(\psi) \to \mathbb{R}_+, \quad h(r) := \frac{r}{1-r^2} \sin(\psi(r)) \tag{3}$$

must be non-increasing.

Here $Dom(\psi)$ denotes the domain of ψ; when the cones are specified through h, it is the maximal set $\{r \in (0,1) : h(r) \leq h_{\max}(r)\}$ on which $\psi = \arcsin(h/h_{\max})$ is defined. Figure 1 shows the geometric illustration of Theorem 3.1. The dashed circle, centered at the origin with radius $\operatorname{arcsinh}(2h)$, is tangent to the complete geodesic extending the boundary generatrix. If h is constant or decreasing, the cone of a child y on the boundary of the cone of x still fits inside the parent cone, whereas if h increases, the child cone leaks out from the parent cone and transitivity breaks.

Accordingly, the aperture is determined by the choice of a non-increasing h. Taking $h \equiv K$ constant as Ganea et al. (2018b) do, the aperture function is given in closed form as follows:

$$\psi_K(r) = \arcsin\left(K\frac{1-r^2}{r}\right). \tag{4}$$

We refer to the cones with this aperture as the entailment cones. However, for the argument of $\arcsin$ not to exceed 1, the radius must satisfy

$$r \geq \varepsilon := \frac{2K}{1+\sqrt{1+4K^2}}, \tag{5}$$

which restricts the domain of ψ_K and h to $[\varepsilon, 1)$. Here ε increases monotonically with K, and so does the aperture by (4). Therefore, K simultaneously determines the width of the cones and the radius of the infeasible region, and thus a wide cone and a small infeasible region cannot be obtained at the same time.

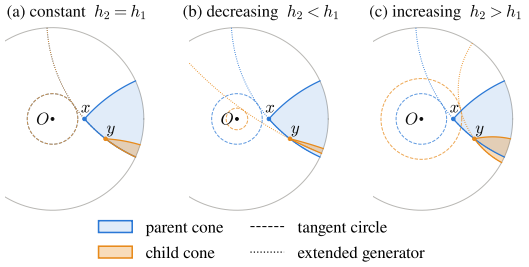

Figure 1: The relationship between the function h and the containment of cones.

It is known that this trade-off appears for the constant function $h \equiv K$. The essential point, however, is that the trade-off is not specific to the choice of a constant function but originates from exact transitivity and holds for any non-increasing h. Setting

$$h_{\max}(r) := \frac{r}{1-r^2}, \tag{6}$$

we have $h(r) \leq h_{\max}(r)$ for $r \in Dom(\psi)$, since $\psi \leq \pi/2$ (Ganea et al., 2018b, Lemma 2).

Proposition 3.2 (Generalization of the trade-off to arbitrary non-increasing h)**.** *Take any non-increasing function h not identically 0, and let $\varepsilon := \inf Dom(\psi)$. Then $\varepsilon > 0$, so the cones necessarily have a region near the origin where ψ cannot be defined. Moreover, defining the constant function $\tilde{h} \equiv K$ (with $K := h_{\max}(\varepsilon)$) with the same domain as h, we have $\psi(r) \leq \psi_K(r)$.*

The proof of Proposition 3.2 is given in Appendix A. Therefore, the aperture of any transitive cones with a given radius ε of the infeasible region is bounded from above at every point by the aperture ψ_K of the constant-function cones ($K = h_{\max}(\varepsilon)$).

Since the pointwise upper bound on ψ is a function of the radius r, it gives no single number with which to compare cones with different ε or to quantify the trade-off as a whole. Here we introduce the following scalar quantity, obtained by integrating out the radial direction, and call it the **capacity**:

$$C(\psi) := \int_{t_0}^{\infty} \sin\psi(t)\, dt, \qquad C(\psi_K) = \frac{2\varepsilon}{1-\varepsilon^2} \log\frac{1}{\varepsilon} \tag{7}$$

where $t_0 := 2\operatorname{arctanh}\big(\inf Dom(\psi)\big)$ is the hyperbolic norm of the infimum of the domain. Since $\sin\psi = h/h_{\max}$, $\sin\psi(t)$ can be regarded as the aperture used by the cone, normalized by the

maximum value $h_{\max}$ admissible at each radius. Thus $C(\psi)$ aggregates this normalized aperture along the radial direction. Since C is monotone in ψ, Proposition 3.2 gives $C(\psi) \leq C(\psi_K)$ for any transitive cones with the same infeasible region. Moreover, as $\varepsilon \to 0$ this upper bound also converges to 0. Hence, under transitivity, feasibility and capacity cannot be achieved together.

The cones attaining the upper bound of capacity within the same infeasible region are also determined.

Corollary 3.3 (Optimality of the constant function)**.** *Consider transitive cones with* $\inf Dom(\psi) = \varepsilon$ *for* $\varepsilon \in (0, 1)$*. The capacity is maximized by* $h \equiv h_{\max}(\varepsilon)$*, the entailment cones with* $K = h_{\max}(\varepsilon)$*.*

The proof is given in Appendix B. By Corollary 3.3, within the same infeasible region, the choice of Ganea et al. (2018b) is optimal in that it maximizes the capacity. However, since the most abstract concepts are placed near the origin in hyperbolic embeddings, an infeasible region near the origin directly hinders the representation of parents. Indeed, the update rule of the embeddings implemented by Ganea et al. (2018b) keeps projecting the embeddings to the outside of the infeasible region.

Furthermore, it follows from the above that transitive cones with aperture not identically 0 cannot be defined on the whole of $\mathbb{D}^n \setminus \{O\}$. The lower end of the domain would have to be positive, so such cones are limited to $\psi \equiv 0$ whose cones collapse to the half-line and cannot represent containment.

4 (ε, δ)-TRANSITIVITY

Since all the arguments in §3 are derived from transitivity, resolving this trade-off requires reconsidering the condition of transitivity itself. Moreover, from the previous discussion, the entailment cones only satisfy transitivity outside the ball region. In this section, to overcome the fundamental limitations, we introduce (ε, δ)**-transitivity**, which tolerates violations of transitivity.

Definition 4.1 ((ε, δ)-transitivity and budget)**.** Let $\varepsilon \in [0, 1)$ and $\delta : (\varepsilon, 1) \to \mathbb{R}_+$. Cones whose domain contains $(\varepsilon, 1)$ are (ε, δ)**-transitive** if $h + \delta$ is non-increasing. The function δ is then called the **budget**.

Canceling the increase of h with δ allows h to increase and thereby removes the constraints of the previous section that originate from exact transitivity. From the geometric perspective, (ε, δ)-transitivity can be interpreted as allowing the cone of a child to protrude from the cone of its parent within the budget (Figure 2). Also, setting $\delta \equiv 0$ in Definition 4.1 reduces the condition to the non-increasingness of h itself (Theorem 3.1). In this case the child cone never leaks out of the parent cone and the conventional exact nesting is recovered, so (ε, δ)-transitivity can be regarded as a natural generalization of the conventional formulation; accordingly, we call the entailment cones by Ganea et al. (2018b) the $(\varepsilon, 0)$ cones hereafter. Violations along chains of any length are bounded in terms of δ; this bound and the relation to the alternative of bounding violations directly by an angular tolerance are discussed in Appendices C and D.

Discussion The significance of this formulation lies in recasting transitivity as a three-way trade-off structure among feasibility, capacity, and the allowance for violations (budget δ). The $(\varepsilon, 0)$ cones force $\delta \equiv 0$, so feasibility and capacity fall into a direct trade-off, and the cones are consequently limited to $(\varepsilon, 0)$-transitive ones. In contrast, allowing $\delta > 0$ permits designing h as an increasing function, so that feasibility can be fixed to the whole domain except the origin ($\varepsilon = 0$) and the trade-off shifts to the relation between capacity and δ. Since this exchange depends on the shape of the budget function, we construct explicit $(0, \delta)$-transitive cones in §5 and evaluate their capacity and δ quantitatively.

5 MAXIMUM-ENTROPY CONES

In this section, we construct $(0, \delta)$-transitive cones in closed form. To obtain a closed-form aperture and budget, we first specify h and then construct δ as the minimal budget. Under $(0, \delta)$-transitivity, only $h + \delta$ must be non-increasing, so h itself may increase. Our construction requires three conditions: the non-increasingness of $h + \delta$, feasibility everywhere except the origin, and convergence to

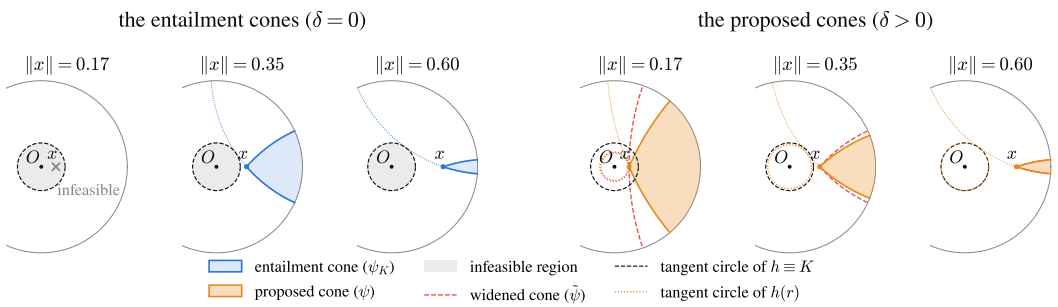

Figure 2: $(\varepsilon, 0)$ cones (left) and the proposed cones (right, the $(0, \delta_2)$ cones of §5) with $K = 0.3$. The tangent circle of h is the same at every point for $(\varepsilon, 0)$ cones, which are infeasible inside it, whereas it shrinks toward the origin for the proposed cones. The red dashed curve is the widened cone $\tilde{\psi}$ that bounds the violations of transitivity for $\|x\| \geq \varepsilon$.

the aperture of the $(\varepsilon, 0)$ cones as the radius grows. The following parametrization of h through a cumulative distribution satisfies all three by construction, with δ the minimal budget.

Theorem 5.1 (Cones from a cumulative distribution). *Let $s := h_{\max}(r)/K$ for $K > 0$. Here s is the radius rescaled so that $s = 1$ corresponds to $r = \varepsilon$. Let p be a non-increasing probability density on $[0, \infty)$ with $p(0) = 1$, and P its cumulative distribution function. Then the aperture $\psi = \arcsin(P(s)/s)$ determined by*

$$h := K\,P(s), \qquad \delta := K\big(1 - P(s)\big) \tag{8}$$

is defined everywhere except the origin, and since $h + \delta \equiv K$, these cones are $(0, \delta)$-transitive. The aperture of these cones converges to that of the $(\varepsilon, 0)$ cones as the radius grows. Indeed, $P(s) \to 1$ as $s \to \infty$, so $h \to K$ and the constructed aperture converges to that of the $(\varepsilon, 0)$ cones. Furthermore, the constructed δ is guaranteed to be minimal.

The proof is given in Appendix E. In particular, for $p = \mathbf{1}_{[0,1]}$, the aperture coincides with that of the $(\varepsilon, 0)$ cones on their feasible region $r \geq \varepsilon$. Under this construction, we determine p by the maximum-entropy principle.

Lemma 5.2 (Maximum-entropy density). *Let $1 \leq q < \infty$. Among the probability densities on $[0, \infty)$ with a fixed q-th moment, the one maximizing the differential entropy has the form $p(s) = c\,e^{-\lambda s^q}$, where $\lambda > 0$ is determined by the value of the moment and $c = \lambda^{1/q}/\Gamma(1 + 1/q)$. Among these densities, exactly one satisfies $p(0) = 1$, namely*

$$p_q(s) = \exp\big(-(\Gamma(1 + 1/q)\, s)^q\big). \tag{9}$$

The first part follows from Cover & Thomas (2006), and a detailed proof is given in Appendix F. Defining h through the cumulative distribution of this probability density yields cones that recover, in a limit, the extension of the $(\varepsilon, 0)$ cones to the whole domain. Indeed, as $q \to \infty$ the following corollary holds.

Corollary 5.3 (The limit $q \to \infty$). *As $q \to \infty$, p_q converges to the uniform density $\mathbf{1}_{[0,1]}$ on $s \neq 1$, and $P_q(s) \to \min(s, 1)$ on $s > 0$. Consequently, the aperture of the cones determined by p_q converges to $\psi_K(r)$ of the $(\varepsilon, 0)$ cones for $r > \varepsilon$, and to $\pi/2$ for $r \leq \varepsilon$.*

The proof is given in Appendix G. Therefore, the moment order q continuously connects the proposed cones with the $(\varepsilon, 0)$ cones joined with an aperture of $\pi/2$ on their infeasible region, and the proposed method is positioned as a one-parameter generalization of $(\varepsilon, 0)$ cones' aperture in the feasible region. In this paper, we consider $q = 1, 2, \infty$. The apertures determined by $q = 1, 2$ are

$$\psi(r) = \arcsin\left(\frac{1 - e^{-s}}{s}\right), \qquad \psi(r) = \arcsin\left(\frac{\operatorname{erf}\left(\frac{\sqrt{\pi}}{2} s\right)}{s}\right), \tag{10}$$

respectively.

In terms of Definition 4.1, we call them the $(0, \delta_1)$ cones and the $(0, \delta_2)$ cones, and the limit $q \to \infty$ of Corollary 5.3 the $(0, \delta_\infty)$ cones. The corresponding budgets are in closed form, $\delta_1 = Ke^{-s}$, $\delta_2 = K\,\mathrm{erfc}\big(\frac{\sqrt{\pi}}{2}s\big)$, and $\delta_\infty = K\max(0, 1-s)$, as shown in Figure 3. The budget is spent almost entirely inside the former infeasible region. Measured by $\int \delta\, ds$, 100% of the budget is spent inside ε for δ_∞, 87% for δ_2, and 63% for δ_1. δ_2 falls below $0.01K$ at about 2ε and δ_1 at about 4ε, so violations of transitivity are confined to a neighborhood of the region where the $(\varepsilon, 0)$ cones were infeasible. The gradient of the aperture near the origin is determined by the leading term of the Maclaurin expansion of $\delta_q - K(1-s)$, the first term in which the budgets differ (of order s^{q+1}, which gives $\psi_q = \pi/2 - O(s^{q/2})$): it is zero inside ε for the $(0, \delta_\infty)$ cones, finite for the $(0, \delta_2)$ cones, and infinite for the $(0, \delta_1)$ cones. The $(\varepsilon, 0)$ cones are infeasible inside ε, so their gradient is not even defined there.

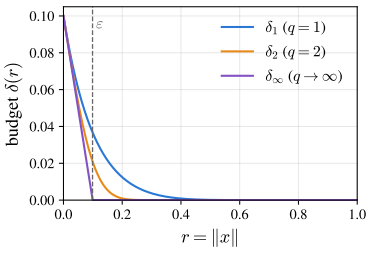

Figure 3: The budgets δ_1, δ_2, and δ_∞ along the radius for $K = 0.1$; the dashed line is ε.

The constructed cones avoid the conventional capacity–feasibility trade-off: they are feasible on the whole domain for every K, and their capacity is at least that of the $(\varepsilon, 0)$ cones with the same K (Appendix H). The trade-off instead moves to capacity–budget (Appendix I).

6 Experiments

We validate the proposed method on hierarchical word embedding benchmarks and on the hyperbolic vision-language model MERU by controlled comparisons that replace only the aperture formula. The experiments target the Poincaré ball and the Lorentz models, on which the $(\varepsilon, 0)$ cones are used directly. Shadow Cones change both the coordinate model and the conditions of the cones, which would be confounded with the relaxation of transitivity, so the comparison with them is given separately in Appendix L.

6.1 Setup for hierarchical word embeddings

Task We evaluate by binary classification of whether x is a hypernym of y for ordered pairs (x, y) in the transitive closure of the hierarchy. Following the protocol of Ganea et al. (2018b), training uses the basic edges plus a fixed percentage of non-basic edges (edges added by the transitive closure) as positive examples. Evaluation is F1 on held-out positive and negative pairs, where the containment criterion regards $f(x, y) = 0$ as a positive prediction (Appendix J).

Datasets We use three hierarchies: the nouns of WordNet (Fellbaum, 1998); the Microsoft Concept Graph (MCG) (Wu et al., 2012; Wang et al., 2015); and hypernym pairs extracted from a large corpus by Hearst patterns (Hearst) (Hearst, 1992; Le et al., 2019). The percentages of non-basic edges are $\{0, 10, 25, 50, 90\}\%$, plus a reconstruction setting that uses the entire closure for training. Because the closure of Hearst is too large for the settings of 25% or higher, the main table uses Hearst-25k, rebuilt from the top 25,000 rows by the same procedure (Appendix M). The original Hearst at $\{0, 1, 2, 5, 10\}\%$ is reported in Appendix Table 6.

Compared methods As reference methods we use Euclidean distance embeddings, Poincaré embeddings (Nickel & Kiela, 2017), and order embeddings (Vendrov et al., 2016). Against the Euclidean cone and the $(\varepsilon, 0)$ cones (the original formula ψ_K), we compare the proposed $(0, \delta_1)$ and $(0, \delta_2)$ cones, replacing only the aperture formula under identical loss, initialization, and projection, on both the Poincaré and the Lorentz model (Appendix K).

Training details Initialization, epochs, batch size, and margin are inherited from the original settings of each method (Appendix M). The learning rate is selected on validation for each pair of method and dataset. K is fixed to 0.1 during training, and at evaluation time K is selected on validation (Appendix J). Each condition uses 10 seeds and reports their average.

Table 2: F1 (%) under the containment criterion; colored numbers: difference from the $(\varepsilon, 0)$ cones of the same block. Bold: best in each column within each block, excluding the reference methods; green background : best over all rows.

	WordNet nouns				MCG				Hearst-25k			
non-basic edges (%)	25	50	90	recon.	25	50	90	recon.	25	50	90	recon.
Reference methods												
Poincaré embedding	38.3	40.1	45.3	54.8	23.4	23.0	23.5	24.0	23.0	23.1	23.7	24.0
order embedding	76.5	81.8	86.8	98.3	52.0	50.6	52.1	54.1	61.0	61.0	69.7	70.3
Protocol of Ganea et al. (Poincaré model)												
$(\varepsilon, 0)$ cones	83.9	90.1	**95.7**	**97.3**	49.6	46.6	53.4	65.4	49.9	49.4	67.3	78.3
$(0, \delta_2)$ cones	**86.9** +3.0	**91.6** +1.5	95.5 -0.2	95.4 -1.9	**59.3** +9.7	**64.9** +18.3	**75.4** +22.0	**76.0** +10.6	**70.0** +20.1	**77.0** +27.6	**81.9** +14.6	**83.4** +5.1
$(0, \delta_\infty)$ cones	84.7 +0.8	89.9 -0.2	94.0 -1.7	95.4 -1.9	52.8 +3.2	55.9 +9.3	63.0 +9.6	72.6 +7.2	60.2 +10.3	66.4 +17.0	77.3 +10.0	83.3 +5.0
Protocol of Ganea et al. (Lorentz model)												
$(\varepsilon, 0)$ cones	83.8	90.5	**95.5**	**97.4**	49.8	46.7	53.7	65.5	50.7	49.5	67.0	78.7
$(0, \delta_2)$ cones	**86.9** +3.1	**91.9** +1.4	95.4 -0.1	95.5 -1.9	**60.6** +10.8	**67.7** +21.0	**79.3** +25.6	**80.7** +15.2	**73.2** +22.5	**79.0** +29.5	**84.9** +17.9	**86.5** +7.8
$(0, \delta_\infty)$ cones	85.0 +1.2	90.2 -0.3	94.4 -1.1	95.3 -2.1	52.7 +2.9	55.8 +9.1	62.8 +9.1	72.5 +7.0	60.8 +10.1	66.4 +16.9	77.3 +10.3	83.3 +4.6

6.2 Results for hierarchical word embeddings

Effect of the aperture formula Table 2 reports F1 under the containment criterion. The main table covers 25% and above at $d = 5$; the full version with 0% and 10%, the Euclidean cone, and $d = 10$ is Appendix Table 4. Among the proposed cones, the main table shows the $(0, \delta_2)$ cones, whose gradient does not diverge as $r \to 0$, and, as an ablation, the $(0, \delta_\infty)$ cones.

Under the protocol of Ganea et al. (2018b), the $(0, \delta_2)$ cones outperform the $(\varepsilon, 0)$ cones in 10 of the 12 cells of the main table (16 of 18 with 0% and 10%, Appendix Table 4). The exceptions are the 90% and reconstruction settings of nouns, which has few nodes with two or more parents, and the difference there is at most 1.9 points. The improvements are especially large on MCG and Hearst-25k, which have many nodes with two or more parents. On the 90% setting of MCG the $(0, \delta_2)$ cones attain 75.4 versus 53.4, and on the 50% setting of Hearst-25k 77.0 versus 49.4. The Lorentz-model block shows the same trends. On the Poincaré model, the $(0, \delta_1)$ and $(0, \delta_2)$ cones differ by at most 3.8 points in the main table (25% on MCG), far less than their improvements over the $(\varepsilon, 0)$ cones on MCG and Hearst-25k (Appendix Table 4). The $(0, \delta_2)$ cones exceed the $(0, \delta_\infty)$ cones in 11 of the 12 cells of the main table. Since the $(0, \delta_\infty)$ cones are the $(\varepsilon, 0)$ cones joined with the constant aperture $\pi/2$ on the infeasible region, and they alone exceed the $(\varepsilon, 0)$ cones by up to 17.0 points on MCG and Hearst-25k, removing the infeasible region already yields a large improvement, and $q = 1, 2$, which interpolate the limit smoothly, improve further. These outcomes follow from Figure 3: the $(0, \delta_\infty)$ cones have no budget outside ε and gain only from removing the infeasible region, the looser δ_1 helps on MCG with its many multiple-parent nodes, and the tighter δ_2 suits the tree-like nouns.

Sensitivity to K and violations of transitivity Sweeping K from 10^{-4} to 10 at 50% (Appendix J.1), the $(\varepsilon, 0)$ cones peak at $K = 0.1$ (89.6, 42.7, and 52.7 on nouns, MCG, and Hearst-25k) and fall to 8.2, 0.0, and 0.0 at $K = 10$, where the infeasible region covers 95% of the radius. The $(0, \delta_2)$ cones keep 89.7, 74.5, and 81.7 at $K = 10$, and the $(0, \delta_\infty)$ cones stay flat at a lower level. Large K is thus harmless once the infeasible region is removed, and smooth saturation adds the remaining improvement. Appendix M also measures the violations of transitivity in the relation predicted over all pairs. The $(0, \delta_2)$ cones violate transitivity in 0.6, 5.1, and 8.5% of their 2-paths, almost only for chains starting inside ε. Outside ε, the violations stay within the bound of Proposition C.1. The effect of the violations is thus limited, and the gain from feasibility on the whole domain outweighs it.

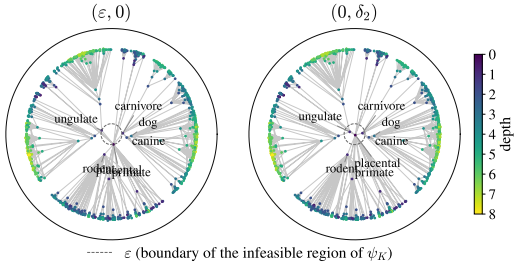

Figure 4: Embeddings of WordNet mammals trained in two dimensions. Node colors indicate the depth from the root, gray lines the basic edges.

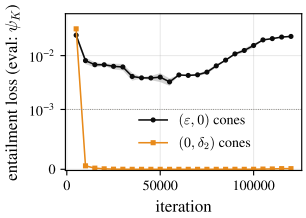

Figure 5: Validation entailment loss of MERU trained on each formula.

Table 3: ImageNet hierarchy metrics of MERU and mean normalized radius $\bar{s}$ of the validation text and image embeddings ($s < 1$: inside ε). Bold: best per seed, except $\bar{s}$.

		TIE↓	LCA↓	Jaccard↑	Hier. prec.↑	Hier. rec.↑	$\bar{s}_{\text{text}}$	$\bar{s}_{\text{image}}$
seed 1	$(\varepsilon, 0)$	3.877	2.339	0.768	0.840	0.841	0.97	1.99
	$(0, \delta_2)$	**3.762**	**2.282**	**0.775**	**0.845**	**0.846**	0.85	3.27
seed 2	$(\varepsilon, 0)$	3.870	2.327	0.768	0.840	0.840	0.97	1.99
	$(0, \delta_2)$	**3.775**	**2.299**	**0.774**	**0.845**	**0.846**	0.85	3.26
seed 3	$(\varepsilon, 0)$	3.829	**2.300**	0.771	**0.843**	0.842	0.97	1.99
	$(0, \delta_2)$	**3.809**	2.329	**0.774**	**0.843**	**0.847**	0.85	3.25

Qualitative evaluation Figure 4 visualizes embeddings of WordNet mammals trained in two dimensions as the $(\varepsilon, 0)$ cones and the $(0, \delta_2)$ cones. With the $(\varepsilon, 0)$ cones, the nodes closest to the root stay on the boundary of the infeasible region, because training keeps projecting the embeddings outside it, whereas with the $(0, \delta_2)$ cones they move inside toward the origin, where the cones are widest. The whole-domain feasibility of §5 is thus used in training, and the center of the ball, empty under the $(\varepsilon, 0)$ cones, is occupied by the top of the hierarchy (Appendix M).

6.3 Setup for hyperbolic vision-language models

We target MERU (Desai et al., 2023) (formulation on the Lorentz model: Appendix K), which, besides its contrastive loss, has a single entailment loss between image and text, so only the aperture formula changes in our comparison. We pretrain models whose entailment loss uses the $(\varepsilon, 0)$ cones or the $(0, \delta_2)$ cones, using a ViT-B/16 backbone and three seeds each on GRIT (Peng et al., 2024), with the same configuration otherwise (Appendix M). Since the implementation of MERU clips the argument of $\arcsin$ to 1, its $(\varepsilon, 0)$ cones are in fact the $(0, \delta_\infty)$ cones of Corollary 5.3, so the comparison isolates the effect of smooth saturation. Following HyCoCLIP, we evaluate zero-shot classification, retrieval, and the ImageNet hierarchy metrics (Appendix M), and re-evaluate the validation entailment loss of every checkpoint with ψ_K.

6.4 Results for hyperbolic vision-language models

Table 3 reports the ImageNet hierarchy metrics and the mean normalized radius $s = \sqrt{c}\,\|x\|/2K$ of (15), and Figure 5 the validation entailment loss; zero-shot classification and retrieval accuracies show no consistent trend across seeds (Appendix M). On the hierarchy metrics, the $(0, \delta_2)$ cones exceed the $(\varepsilon, 0)$ cones in 13 of the 15 cells. The differences are small, but the entailment loss of the original formula falls below 10^{-3} within about 10,000 iterations with the $(0, \delta_2)$ cones and stays there, whereas with the $(\varepsilon, 0)$ cones it rises in the latter half of training (final values 10^{-5} versus 2×10^{-2}). The radii explain this: with both cones the parents (text) settle at $s < 1$, where the clipped aperture of the $(\varepsilon, 0)$ cones, that of the $(0, \delta_\infty)$ cones, is the constant $\pi/2$ with vanishing gradient. The entailment loss can then hardly act, only the contrastive loss shapes the configuration, and the images stay at $\bar{s} \approx 2.0$, whereas the proposed cones keep a finite aperture and a non-vanishing gradient there, so the entailment loss stays active and pushes the images outward to $\bar{s} \approx 3.3$. In this sense the proposed cones may yield a better embedding configuration, although this does not translate into classification or retrieval accuracy, where the contrastive loss dominates.

7 Conclusion

Starting from transitivity, we analyzed the capacity–feasibility trade-off of entailment cones and introduced (ε, δ)-transitivity, a three-way structure among capacity, feasibility, and budget that generalizes conventional transitivity and yields $(0, \delta)$-transitive cones feasible on the whole domain except the origin. In experiments on hierarchical word embeddings (WordNet, MCG, Hearst), the proposed cones outperformed the $(\varepsilon, 0)$ cones in 10 of 12 settings, with a gain of up to 27.6 points. With MERU they preserved accuracy and kept the entailment loss active where the $(\varepsilon, 0)$ cones saturate, although their advantage depends on the task and the dataset. Task-dependent selection of q and an entailment loss whose improvement reaches the accuracy of MERU are left for future work.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. They helped develop the theoretical framework after its central idea, the reformulation of transitivity into (ε, δ)-transitivity, had been conceived by the authors without AI assistance; they helped refine the statements of mathematical claims, provided ingredients for proving them, and assisted in the writing of proofs; they proposed and refined hypotheses; they designed and gave feedback on the experiments; they implemented methods, including the experimental pipelines and the evaluation and figure scripts; they assisted with the translation of the manuscript into English; and they assisted in interpreting the results. We have not used generative AI tools to generate synthetic datasets, and the cleaning and reformatting of datasets and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting parts of the paper, editing the text for readability, identifying and summarizing related literature, formatting references, creating and modifying scientific figures, suggesting the structure of the paper, and brainstorming. We have reviewed all AI-assisted work. Every mathematical claim and proof was checked by the authors line by line and, where applicable, verified numerically; all AI-generated code was reviewed by the authors, and its outputs were cross-checked against independent implementations and published results; literature identified with AI tools was verified against the original sources; and all drafted or translated text was revised by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs of all theoretical claims are given in Appendices A to I, and the formulation of every aperture on the Lorentz model in Appendix K. The code of the experiments on hierarchical word embeddings is provided as supplementary material. It contains the implementations of the $(\varepsilon, 0)$, $(0, \delta_1)$, $(0, \delta_2)$, and $(0, \delta_\infty)$ cones on the Poincaré ball and on the Lorentz model, the reference methods, and the four formulations of Shadow Cones; the aperture formula is selected by a single configuration key, so that all compared cones share the same loss, initialization, and projection. The datasets are not redistributed, but scripts build the vocabularies, transitive closures, negative pairs, and train/validation/test splits from the public files of each hierarchy (the WordNet closures distributed with the implementation of Ganea et al. (2018b), the raw Microsoft Concept Graph, and the Hearst files released with Shadow Cones), and Hearst-25k is built from the same raw Hearst files by one of these scripts. One configuration file per table and figure specifies every experiment, task, and hyperparameter; a single command runs all jobs of a configuration, saves the result and the checkpoints of each job, and aggregates the results into the CSV files from which the tables are generated. Each configuration file lists one seed; the averages of the paper are obtained by extending the list to the 10 seeds of each condition (3 for the sweeps of K). The scripts of the two-dimensional visualization, the sweeps of K, the norm distributions, and the transitivity audit are included, and all experiments run on CPUs with one core per job (Appendix M). The datasets, splits, negative sampling, hyperparameters, and candidate grids of these experiments are also described in Appendix M and, for the protocol of Shadow Cones, in Appendix L. The experiments on MERU are not part of the supplementary code; they use the public implementation of MERU with the aperture formula of its entailment loss replaced, and their pretraining and evaluation settings, including the held-out validation split of GRIT, are given in Appendix M.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223, 2011.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Jindou Dai, Yuwei Wu, Zhi Gao, and Yunde Jia. A hyperbolic-to-hyperbolic graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 154–163, 2021.

Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2004.

Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018a.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1646–1655, 2018b.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 539–545, 1992.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Chuong Huynh, Hossein Souri, Abhinav Kumar, Vitali Petsiuk, Deen Dayal Mohan, and Suren Kumar. Argent: Adaptive hierarchical image-text representations. *arXiv preprint arXiv:2603.23311*, 2026.

Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6418–6428, 2020.

Aris Kosmopoulos, Ioannis Partalas, Éric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865, 2015.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 554–561, 2013.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Matt Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 3231–3241, 2019.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the International Conference on Machine Learning (ICML)*, volume 97, pp. 4693–4702, 2019.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 3779–3788, 2018.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Grounding multimodal large language models to the world. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27253–27262, 2024.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80, pp. 4460–4469, 2018.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Proceedings of the International Symposium on Graph Drawing*, pp. 355–366, 2011.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Jianing Sun, Zhaoyue Cheng, Saba Zuberi, Felipe Perez, and Maksims Volkovs. Hgcf: Hyperbolic graph convolution networks for collaborative filtering. In *Proceedings of the Web Conference 2021*, pp. 593–601, 2021.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zhongyuan Wang, Haixun Wang, Ji-Rong Wen, and Yanghua Xiao. An inference approach to basic level of categorization. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 653–662, 2015.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 481–492, 2012.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

Daiki Yoshikawa and Takashi Matsubara. PHyCLIP: ℓ_1-product of hyperbolic factors unifies hierarchy and compositionality in vision-language representation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Tao Yu, Toni J.B. Liu, Albert Tseng, and Christopher De Sa. Shadow cones: A generalized framework for partial order embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

A Proof of Proposition 3.2

Proof. We first show $\varepsilon > 0$. Suppose $\varepsilon = 0$. Then for any $r_1 \in Dom(\psi)$ we can take a sequence $r_n \le r_1$ in $Dom(\psi)$ converging to 0. By the non-increasingness of Theorem 3.1 and $h \le h_{\max}$,

$$h(r_1) \;\le\; h(r_n) \;\le\; h_{\max}(r_n) \;\to\; 0 \qquad (n \to \infty),$$

hence $h(r_1) = 0$. Since r_1 is arbitrary, $h \equiv 0$ so $\psi \equiv 0$, contradicting the assumption. Next we show that the domains coincide. Since the aperture is determined by $\psi(r) = \arcsin(h(r)/h_{\max}(r))$, ψ is defined where $h(r) \le h_{\max}(r)$ holds. As h is non-increasing and $h_{\max}$ is monotonically increasing, once this inequality holds at some point it holds at every point of larger radius, so $Dom(\psi)$, the maximal set on which it holds, is an interval with left end ε, which we write $[\varepsilon, 1)$ (whether ε itself belongs to it is immaterial below). On the other hand, by (5) the lower end of the domain of $\tilde{h} \equiv K$ is the solution of $h_{\max}(r) = K$, which equals ε since we set $K = h_{\max}(\varepsilon)$; hence the domain $[\varepsilon, 1)$ of $\tilde{h} \equiv K$ coincides with that of h. Finally we compare the apertures. Take any r in the domain and a sequence $r_n \le r$ in the domain converging to ε; by the non-increasingness and $h \le h_{\max}$,

$$h(r) \;\le\; h(r_n) \;\le\; h_{\max}(r_n) \;\to\; h_{\max}(\varepsilon) \;=\; K \qquad (n \to \infty).$$

Hence at every point

$$\sin\psi(r) \;=\; \frac{h(r)}{h_{\max}(r)} \;\le\; \frac{K}{h_{\max}(r)} \;=\; \sin\psi_K(r),$$

that is, $\psi(r) \le \psi_K(r)$. □

B Proof of Corollary 3.3

Proof. We first derive (7). By the change of variables $t = 2\,\mathrm{arctanh}(r)$ and $\sin\psi_K = K\frac{1-r^2}{r}$,

$$C(\psi_K) = \int_\varepsilon^1 \frac{2K}{r}\, dr = 2K \log\frac{1}{\varepsilon},$$

and solving (5) as $K = \varepsilon/(1-\varepsilon^2)$ and substituting gives (7). By the pointwise domination $\psi(r) \le \psi_K(r)$ of Proposition 3.2 and the coincidence of the domains, comparing the nonnegative integrands on the same interval shows that any cones satisfying the condition satisfy $C(\psi) \le C(\psi_K)$. Since $h \equiv K$ itself satisfies the condition and attains equality, the maximum is $C(\psi_K)$, which coincides with (7). □

C Upper Bound on the Violation of Chains

In this section we show that, for (ε, δ)-transitive cones, the endpoint violation induced by a chain of entailments of any length stays inside the cone widened by the budget.

Proposition C.1 (Upper bound on the violation of chains)**.** *Let the budget-augmented function and the aperture it determines be*

$$\tilde{h}(r) \;:=\; h(r) + \delta(r), \qquad \tilde{\psi}(r) \;:=\; \arcsin\frac{\tilde{h}(r)}{h_{\max}(r)}. \tag{11}$$

Suppose the cones determined by h are (ε, δ)-transitive and that $\tilde{h}(r) \le h_{\max}(r)$ at a point x with $\|x\| = r$. Then for any sequence of points $\{x, y_1, y_2, ..., y_{m-1}, z\}$ of any length with $f(x, y_1) = f(y_1, y_2) = \cdots = f(y_{m-1}, z) = 0$, the following hold.

- *(i) If $\delta \equiv 0$, that is, under exact transitivity (non-increasing h of Theorem 3.1), $z \in \mathfrak{S}_x^{\psi(r)}$. The endpoint of the chain does not leak out of the cone of x itself, and $f(x, z) = 0$.*
- *(ii) For general δ, $z \in \mathfrak{S}_x^{\tilde{\psi}(r)}$, and therefore*

$$f(x, z) \;\le\; \tilde{\psi}(r) - \psi(r). \tag{12}$$

Here $f(x, y)$ denotes the entailment loss, defined for two points x, y of the space excluding the origin by $f(x, y) := \max\big(0,\ \Xi(x, y) - \psi(x)\big)$, where $\Xi(x, y)$ is the angle at x between the outward axis and the geodesic xy (Ganea et al., 2018b). $\tilde{\psi}$ is the aperture of the cone determined by h widened by the budget, and (ii) states that the endpoint of the chain stays inside this cone.

Proof. We prove (ii) first and derive (i) as a special case. We begin by confirming that a point belonging to a cone is farther from the origin than the apex. The geodesic distance from the origin is convex along a geodesic, and its derivative at x equals the cosine $\cos \Xi(x, y)$ of the angle between the geodesic xy and the outward axis. Since the aperture is at most $\pi/2$, $y \in \mathfrak{S}_x^{\psi(x)}$ implies $\Xi(x, y) \leq \pi/2$, so this derivative is nonnegative. By convexity, the geodesic distance from the origin does not decrease beyond x. Therefore the radii of the points $x, y_1, \ldots, y_{m-1}, z$ are all at least r and lie in the domain $(\varepsilon, 1)$ of the cones.

We first treat the case where $h + \delta$ is a constant function. Writing its value as H, we have $H = h(r) + \delta(r) = \tilde{h}(r)$, and $H \leq h_{\max}(r)$ by the hypothesis of the proposition. Write the aperture obtained by substituting H for K in (4) as

$$\psi_H(r') \ := \ \arcsin \frac{H}{h_{\max}(r')}.$$

This is the aperture for the constant H, distinct from the cones determined by h. By (11), $\psi_H(r) = \tilde{\psi}(r)$, so it suffices to show $z \in \mathfrak{S}_x^{\psi_H(r)}$. At every point of the sequence (radius $r' \geq r$),

$$h(r') \ \leq \ h(r') + \delta(r') \ = \ H.$$

Moreover, since $h_{\max}$ is strictly increasing, $H \leq h_{\max}(r) \leq h_{\max}(r')$, so ψ_H is defined at every point of the sequence. From both, at every point

$$\sin \psi(r') \ = \ \frac{h(r')}{h_{\max}(r')} \ \leq \ \frac{H}{h_{\max}(r')} \ = \ \sin \psi_H(r'),$$

that is, $\psi(r') \leq \psi_H(r')$, and the cone at each point is contained in the cone of aperture ψ_H at the same point. Under this containment we use induction along the sequence. We have $y_1 \in \mathfrak{S}_x^{\psi(x)} \subseteq \mathfrak{S}_x^{\psi_H}$. Assuming $y_i \in \mathfrak{S}_x^{\psi_H}$, we have $y_{i+1} \in \mathfrak{S}_{y_i}^{\psi(y_i)} \subseteq \mathfrak{S}_{y_i}^{\psi_H}$, and since the cones with aperture ψ_H are the $(\varepsilon, 0)$ cones with $K = H$ and hence exactly transitive (Ganea et al., 2018b), $\mathfrak{S}_{y_i}^{\psi_H} \subseteq \mathfrak{S}_x^{\psi_H}$; therefore $y_{i+1} \in \mathfrak{S}_x^{\psi_H}$. Hence $z \in \mathfrak{S}_x^{\psi_H}$, and by the membership condition $\Xi(x, z) \leq \psi_H(r)$, so

$$f(x, z) \ = \ \max\big(0,\ \Xi(x, z) - \psi(r)\big) \ \leq \ \psi_H(r) - \psi(r),$$

which is (12).

In the general case where $h + \delta$ decreases, set again $H := \tilde{h}(r) = h(r) + \delta(r)$. By Definition 4.1 and $\delta \geq 0$, at every point of the sequence (radius $r' \geq r$)

$$h(r') \ \leq \ h(r') + \delta(r') \ \leq \ h(r) + \delta(r) \ = \ H.$$

Since the argument for the constant case used only $H = \tilde{h}(r) \leq h_{\max}(r)$ and $h \leq H$ at every point, the same argument applies verbatim and gives (12).

Finally we derive (i). If $\delta \equiv 0$, the hypothesis $\tilde{h} \leq h_{\max}$ always holds as $h \leq h_{\max}$, and $\tilde{h} = h$ gives $\tilde{\psi} = \psi$. Hence (ii) yields $z \in \mathfrak{S}_x^{\psi(r)}$ and $f(x, z) \leq 0$, that is, $f(x, z) = 0$. □

D Relation to Angular Tolerances

In this section we formalize why Definition 4.1 measures violations by the increase of h rather than by angles. A seemingly simpler alternative is to bound violations of transitivity directly by a constant angular tolerance. This tolerance, however, is not uniform in the hyperbolic radius: a fixed angular violation corresponds to an increase of h (a violation of the non-increasingness of Theorem 3.1) that grows exponentially with the geodesic distance from the origin. To keep violations bounded, the angular tolerance must decay at a rate matching the radius, and the requirement that a decaying tolerance bound the violations of chains eventually reduces to a constraint of the same form as

Definition 4.1. We show this below. Consider bounding violations by a constant angle $\varphi \in (0, \pi/2)$ independent of the radius. Allowing a violation of angle φ at a point x of radius r amounts to widening the cone of x to aperture $\psi(r) + \varphi$, and the corresponding increase of h, in the range $\psi(r) + \varphi \le \pi/2$, is

$$\Delta h(r) \;=\; h_{\max}(r)\big(\sin(\psi(r)+\varphi) - \sin\psi(r)\big) \;=\; 2\,h_{\max}(r)\cos\Big(\psi(r)+\frac{\varphi}{2}\Big)\sin\frac{\varphi}{2}.$$

Since $\cos(\psi + \varphi/2) \le \cos(\varphi/2)$, we have $\Delta h \le h_{\max}\sin\varphi$, with equality approached at points where the aperture is close to 0. For the $(\varepsilon, 0)$ cones and cones satisfying Definition 4.1 with a bounded budget, h is bounded, so $\sin\psi = h/h_{\max} \to 0$, and $\Delta h(t)$ diverges exponentially in the geodesic distance $t \to \infty$ from the origin:

$$\Delta h(t) \;\sim\; h_{\max}(t)\sin\varphi \;=\; \frac{\sinh t}{2}\sin\varphi.$$

That is, a constant angular tolerance is the same as allowing exponentially large h-violations at points far from the origin. Therefore, to keep the increase of h bounded, the angular tolerance must decay with the radius at roughly the rate $1/h_{\max}(t)$, that is, e^{-t}.

Suppose then that we adopt a radius-dependent decaying tolerance $\varphi : (\varepsilon, 1) \to [0, \pi/2]$ (with $\psi + \varphi \le \pi/2$ at every point). To guarantee that φ bounds the violations of chains of entailments of any length by the same mechanism as Proposition C.1, namely the nesting of the cones widened to aperture $\psi + \varphi$, the widened cones must satisfy transitivity. By Theorem 3.1, this is equivalent to the non-increasingness of the function of the widened cones,

$$h_{\max}\sin(\psi+\varphi) \;=\; h + \delta_\varphi, \qquad \delta_\varphi(r) \;:=\; h_{\max}(r)\big(\sin(\psi(r)+\varphi(r)) - \sin\psi(r)\big)$$

(sufficiency follows by applying Proposition C.1 (i) to the widened cones), which is nothing but the (ε, δ)-transitivity of Definition 4.1 with budget δ_φ. That is, a decaying tolerance measured in angles translates into a budget measured in h, and reduces, in the end, to the same constraint. Definition 4.1 skips this translation and measures the increase directly in h, the quantity whose non-increasingness characterizes exact transitivity by Theorem 3.1, and Proposition C.1 (ii) provides the reverse translation from the budget to the angular upper bound $\tilde\psi - \psi$. The identity $\sin\tilde\psi - \sin\psi = \delta/h_{\max}$ then builds in the factor $1/h_{\max}$, that is, the decay above, automatically.

E Proof of Theorem 5.1

Proof. Since p is non-increasing with $p(0) = 1$, we have $p \le 1$ on $[0, \infty)$, and integrating over $[0, s]$ gives $P(s) \le s$ for every $s > 0$. Hence the argument $P(s)/s$ of arcsin is at most 1, and the aperture is defined for every $s > 0$, that is, at every point except the origin. Next, $h + \delta = KP(s) + K(1 - P(s)) \equiv K$ is constant and hence non-increasing, and together with $\delta = K(1 - P(s)) \ge 0$, the condition of Definition 4.1 holds with $\varepsilon = 0$. For the convergence, $P(s) \to 1$ as $s \to \infty$ since p is a probability density, so $\sin\psi = P(s)/s$ and $\sin\psi_K = 1/s$ differ by $(1 - P(s))/s \to 0$, and $\psi - \psi_K \to 0$ as $r \to 1$. For the minimality of the budget, let $\delta' \ge 0$ be any budget for which the cones with the same h are $(0, \delta')$-transitive. Definition 4.1 gives $h(t) \le h(r) + \delta'(r)$ for every $t > r$; letting $t \to 1$, that is, $s \to \infty$, we have $h(t) \to K$, hence $\delta'(r) \ge K - h(r) = \delta(r)$ at every r. Finally, $p = \mathbf{1}_{[0,1]}$ is non-increasing with $p(0) = 1$, and $P(s) = \min(s, 1)$. By (5), $s \ge 1$ is equivalent to $r \ge \varepsilon$, and there $\psi = \arcsin(1/s) = \arcsin(K/h_{\max}(r))$, which coincides with $\psi_K(r)$ of (4). □

F Proof of Lemma 5.2

Proof. For $\lambda > 0$, set

$$p_\lambda(s) \;:=\; c_\lambda\, e^{-\lambda s^q}, \qquad c_\lambda \;:=\; \frac{\lambda^{1/q}}{\Gamma(1+1/q)}.$$

Since $\int_0^\infty e^{-\lambda s^q}\,ds = \lambda^{-1/q}\Gamma(1+1/q)$, p_λ is a probability density on $[0, \infty)$, and its q-th moment is

$$\int_0^\infty s^q\, p_\lambda(s)\,ds \;=\; c_\lambda\,\frac{\Gamma(1+1/q)}{q\,\lambda^{1+1/q}} \;=\; \frac{1}{q\lambda}.$$

We first show that, among the densities whose q-th moment equals $m > 0$, the maximizer of the differential entropy $H(p) := -\int p \log p \, ds$ is restricted to p_λ with $\lambda = 1/(qm)$. Let p be any density with $\int_0^\infty s^q p(s)\, ds = m$. By the Gibbs inequality $\int p \log(p/p_\lambda)\, ds \geq 0$,

$$H(p) \;\leq\; -\int_0^\infty p \log p_\lambda \, ds \;=\; -\log c_\lambda + \lambda m \;=\; H(p_\lambda),$$

where the last equality holds because the moment of p_λ itself is $1/(q\lambda) = m$. Equality in the Gibbs inequality holds only for $p = p_\lambda$ (almost everywhere), so the maximizer is uniquely p_λ. Since $\lambda \mapsto 1/(q\lambda)$ is a bijection onto $(0, \infty)$, the maximum-entropy densities obtained by varying the value of the moment are exactly $\{p_\lambda : \lambda > 0\}$.

Next we find the member of this set with $p(0) = 1$. $p_\lambda(0) = c_\lambda = \lambda^{1/q}/\Gamma(1 + 1/q)$ is strictly increasing in λ and a bijection onto $(0, \infty)$, so $p_\lambda(0) = 1$ holds for exactly one λ, namely $\lambda = \Gamma(1 + 1/q)^q$, and the corresponding density is

$$p_q(s) \;=\; \exp\bigl(-(\Gamma(1 + 1/q)\, s)^q\bigr).$$

Its q-th moment is $1/(q\lambda) = 1/(q\,\Gamma(1+1/q)^q)$, and specifying $p(0) = 1$ is equivalent to specifying the moment to this value. Note that p_q is strictly decreasing on $[0, \infty)$ with $p_q(0) = 1$, so it satisfies the hypotheses of Theorem 5.1. □

G Proof of Corollary 5.3

Proof. Set $c_q := \Gamma(1 + 1/q)$; by the continuity of Γ, $c_q \to \Gamma(1) = 1$ as $q \to \infty$. Fix $s < 1$. Since $c_q s \to s < 1$, for sufficiently large q we have $c_q s \leq (1 + s)/2 < 1$, and

$$(c_q s)^q \;\leq\; \Big(\frac{1+s}{2}\Big)^q \;\to\; 0 \qquad (q \to \infty),$$

so $p_q(s) = \exp(-(c_q s)^q) \to 1$. Similarly, for $s > 1$, for sufficiently large q we have $c_q s \geq (1 + s)/2 > 1$, hence $(c_q s)^q \to \infty$, that is, $p_q(s) \to 0$. Therefore p_q converges to $\mathbf{1}_{[0,1]}$ at every point $s \neq 1$. Since $0 \leq p_q \leq 1$, the bounded convergence theorem gives, for every $s > 0$,

$$P_q(s) \;=\; \int_0^s p_q(u)\, du \;\to\; \int_0^s \mathbf{1}_{[0,1]}(u)\, du \;=\; \min(s, 1).$$

Finally, for the aperture: for $s > 1$ ($r > \varepsilon$), $P_q(s)/s \to 1/s < 1$ and the continuity of $\arcsin$ give $\psi \to \arcsin(1/s) = \psi_K(r)$; for $s \leq 1$ ($r \leq \varepsilon$), $P_q(s)/s \to 1$ gives $\psi \to \pi/2$. □

H Capacity of the Constructed Cones

The capacity of the constructed cones has the following closed form.

Proposition H.1 (Capacity of the maximum-entropy cones)**.** *Using the expectation $\mathbb{E}_{p_q}$ over s with density p_q, the capacity of the cones determined by p_q is given by*

$$C(\psi) \;=\; 2K\, \mathbb{E}_{p_q}\Big[\operatorname{arcsinh} \frac{1}{2Ks}\Big]. \tag{13}$$

Proof. With the change of variables $t = \operatorname{arcsinh}(2Ks)$ from (2), we have $dt = 2K\, w(s)\, ds$ with $w(s) := (1 + 4K^2 s^2)^{-1/2}$, and together with $\sin\psi = P_q(s)/s$,

$$C(\psi) \;=\; 2K \int_0^\infty \frac{P_q(s)}{s}\, w(s)\, ds.$$

Noting $\frac{d}{ds} \operatorname{arcsinh} \frac{1}{2Ks} = -\frac{w(s)}{s}$, integration by parts gives

$$\int_0^\infty \frac{P_q(s)}{s}\, w(s)\, ds \;=\; \Big[-P_q(s) \operatorname{arcsinh} \frac{1}{2Ks}\Big]_0^\infty + \int_0^\infty p_q(s)\, \operatorname{arcsinh} \frac{1}{2Ks}\, ds.$$

The boundary term vanishes as $s \to \infty$ by $P_q \leq 1$ and $\operatorname{arcsinh} x \leq x$, and as $s \to 0$ by $P_q(s) \leq s$ and $\operatorname{arcsinh}(1/(2Ks)) \leq \log(1 + 1/(Ks))$. Hence (13) follows. □

Using the same change of variables, the comparison with the $(\varepsilon, 0)$ cones stated in §5 follows.

Corollary H.2 (Comparison of capacities)**.** *The capacity of the cone determined by p_q is at least that of the $(\varepsilon, 0)$ cones with the same K, that is, $C(\psi) \geq C(\psi_K)$.*

Proof. As in the proof above, change variables by $t = \operatorname{arcsinh}(2Ks)$ and set $w(s) := (1 + 4K^2 s^2)^{-1/2}$. Since $\sin\psi = P_q(s)/s$, and the $(\varepsilon, 0)$ cones have $\sin\psi_K = 1/s$ for $s \geq 1$,

$$C(\psi) - C(\psi_K) \;=\; 2K\Big[\int_0^1 \frac{P_q(s)}{s}\, w(s)\, ds - \int_1^\infty \frac{1 - P_q(s)}{s}\, w(s)\, ds\Big].$$

Since w is decreasing, the first term is at least $w(1)\int_0^1 P_q(s)/s\, ds$ and the second is at most $w(1)\int_1^\infty (1 - P_q(s))/s\, ds$, so

$$C(\psi) - C(\psi_K) \;\geq\; 2K\, w(1) \int_0^\infty \frac{P_q(s) - \mathbf{1}_{s\geq 1}}{s}\, ds.$$

We evaluate the integral on the right. From $\int_0^\infty p_q\, du = 1$ we can write $P_q(s) - \mathbf{1}_{s\geq 1} = \int_0^\infty (\mathbf{1}_{u\leq s} - \mathbf{1}_{1\leq s})\, p_q(u)\, du$. The double integral of the absolute value of the integrand equals $\int_0^\infty p_q(u)\, |\log u|\, du < \infty$, so the order of integration can be exchanged; the inner integral is $\int_u^1 ds/s = \log(1/u)$ for $u < 1$ and $-\int_1^u ds/s = -\log u$ for $u > 1$, that is, $-\log u$ in both cases. Hence

$$\int_0^\infty \frac{P_q(s) - \mathbf{1}_{s\geq 1}}{s}\, ds \;=\; -\int_0^\infty p_q(u) \log u\, du.$$

It remains to show that this logarithmic mean is negative. Under the substitution $v = (\Gamma(1 + 1/q)\, u)^q$, v follows the gamma distribution with shape parameter $1/q$, whose logarithmic mean is $(\log\Gamma)'(1/q)$, so

$$\int_0^\infty p_q(u) \log u\, du \;=\; \frac{(\log\Gamma)'(1/q)}{q} - \log\Gamma\Big(1 + \frac{1}{q}\Big) \;=\; g\Big(\frac{1}{q}\Big),$$

where

$$g(x) \;:=\; x\,(\log\Gamma)'(x) - \log\Gamma(1 + x).$$

Using the logarithmic derivative $(\log\Gamma)'(1 + x) = (\log\Gamma)'(x) + 1/x$ of $\Gamma(1 + x) = x\Gamma(x)$ and its derivative,

$$g'(x) \;=\; (\log\Gamma)'(x) + x\,(\log\Gamma)''(x) - (\log\Gamma)'(1 + x) \;=\; x\,(\log\Gamma)''(1 + x),$$

which is positive by the strict convexity of $\log\Gamma$. Moreover $g(1) = (\log\Gamma)'(1) = -\gamma < 0$, where γ is Euler's constant, so $g < 0$ for $x = 1/q \in (0, 1]$. Therefore $C(\psi) - C(\psi_K) \geq -2K\, w(1)\, g(1/q) > 0$. □

I Properties of the Constructed Cones

Figure 6 shows the profiles of the function h, the budget, and the aperture of the constructed cones along the radius.

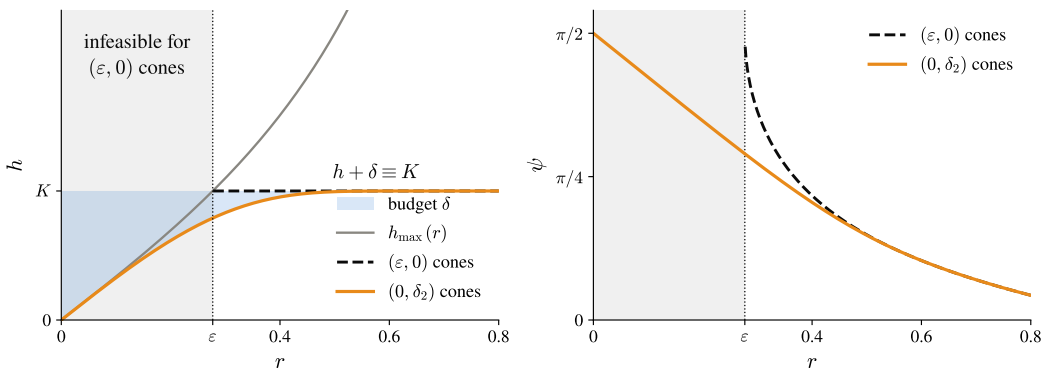

Figure 6: Profiles of the function h and of the aperture along the radius ($K = 0.3$, $q = 2$). Left: h. The $(\varepsilon, 0)$ cones ($h \equiv K$, black dashed) exceed $h_{\max}$ (gray) for $r < \varepsilon$, so the cones are infeasible there (shaded). The h of the proposed cones (orange) is increasing but keeps $h \leq h_{\max}$, and the budget $\delta = K - h$ (blue) cancels its increase so that $h + \delta \equiv K$. Right: the corresponding apertures.

Decay of the violations Since $h + \delta \equiv K$, by Proposition C.1 the violation of any chain of entailments of any length is bounded by $\psi_K - \psi$ for $r > \varepsilon$. By (2), $h_{\max} = \sinh(t)/2$, so $\sinh(t)(\psi_K - \psi)$ is asymptotically $2K(1 - P_q)$, twice the budget, and the bound decays to 0 exponentially in the geodesic distance t from the origin. Inside ε, the violation angle is bounded only by the trivial bound $\pi - \psi$; its empirical size is measured in Appendix M.

Containment on the common feasible domain From $\int_0^\infty p_q \, ds = 1$ we have $P_q \leq 1$, that is, $h \leq K$, and $\psi \leq \psi_K$ for $r > \varepsilon$, so on the common feasible domain $r > \varepsilon$ the cone at each point is contained in the $(\varepsilon, 0)$ cone with the same K. Therefore, for a parent x with $\|x\| > \varepsilon$, every entailment predicted by the proposed cones is also predicted by the $(\varepsilon, 0)$ cones with the same K; inside ε the $(\varepsilon, 0)$ cones are infeasible and no such comparison exists.

Trade-off between capacity and budget As stated at the end of §4, the trade-off does not disappear but moves to a capacity–budget trade-off. Its form is given by the following upper bound for arbitrary cones.

Proposition I.1 (Trade-off between capacity and budget)**.** *For any $(0, \delta)$-transitive cones, set $B := \limsup_{r \to 0} \delta(r)$. If $B < \infty$, then $h(r) \leq B - \delta(r) \leq B$ holds at every point, and*

$$C(\psi) \;\leq\; \operatorname{arcsinh}(2B) + C(\psi_B) \tag{14}$$

holds, where ψ_B is the aperture obtained by substituting B for K in (4)*.*

Proof. Take any r and let $r' < r$. By Definition 4.1 and the fact that the cones are defined at r' ($h(r') \leq h_{\max}(r')$),

$$h(r) + \delta(r) \;\leq\; h(r') + \delta(r') \;\leq\; h_{\max}(r') + \delta(r').$$

The left-hand side does not depend on r', so taking the limit inferior as $r' \to 0$ and using $h_{\max}(r') \to 0$,

$$h(r) + \delta(r) \;\leq\; \liminf_{r' \to 0} \big(h_{\max}(r') + \delta(r')\big) \;\leq\; \lim_{r' \to 0} h_{\max}(r') + \limsup_{r' \to 0} \delta(r') \;=\; B,$$

that is, $h(r) \leq B - \delta(r) \leq B$. Therefore $\sin\psi = h/h_{\max} \leq \min(1, B/h_{\max})$, and by (2), $B/h_{\max} \geq 1$ is equivalent to $t \leq \operatorname{arcsinh}(2B)$, so

$$C(\psi) \;\leq\; \int_0^{\operatorname{arcsinh}(2B)} dt + \int_{\operatorname{arcsinh}(2B)}^{\infty} \frac{2B}{\sinh t}\, dt \;=\; \operatorname{arcsinh}(2B) + C(\psi_B).$$

The second integral coincides with the computation of $C(\psi_K)$ in §3 with $K = B$. □

Under exact transitivity, the excluded radius ε determined the upper limit of the capacity through $h \leq h_{\max}(\varepsilon)$; under $(0, \delta)$-transitivity, the budget B at the origin determines the upper limit of the capacity through $h \leq B$. The upper bound (14) is attainable. Indeed, the limiting cones $h = \min(h_{\max}, K)$, $\delta = K - h$ of Corollary 5.3 as $q \to \infty$ are $(0, \delta)$-transitive since $h + \delta \equiv K$, have $B = K$ since $\delta \to K$ as $r \to 0$, and their aperture $\arcsin \min(1, K/h_{\max})$ satisfies the estimate $\sin \psi \leq \min(1, B/h_{\max})$ in the proof above with equality at every point. Hence these cones attain (14) with equality. For finite q, the capacity is strictly below the bound since $P_q(s) < \min(s, 1)$ for $s > 0$, and it converges to the bound as $q \to \infty$ by the pointwise convergence of Corollary 5.3 and the dominated convergence theorem, with the integrable envelope $\min(1, K/h_{\max})$ given by $P_q(s) \leq \min(s, 1)$.

J Containment and Threshold Criteria

There are two criteria for evaluating F1 for cones. The containment criterion regards the loss $f(x, y)$ being 0, that is, y being contained in the cone of x, as a prediction of entailment. The threshold criterion applies a threshold τ selected on validation to the loss and regards $f(x, y) \leq \tau$ as a prediction of entailment.

The criterion $f(x, y) \leq \tau$ is equivalent to one that widens the aperture by τ at every point. The widened aperture $\psi + \tau$ coincides with no ψ_K, and since an aperture of at least τ remains even at points far from the origin, the capacity in the sense of §3 is unbounded. The difference from the proposed cones is not whether violations of transitivity occur but how much is allowed: by Appendix D, a uniform angular tolerance τ corresponds to an allowance of $h_{\max} \sin \tau$ in h, which grows exponentially with the geodesic distance from the origin, whereas the budget δ of the proposed cones is bounded by K and decays exponentially by (8). Selecting K on validation is model selection within the cones for which Theorem 3.1 or Proposition C.1 holds, while the threshold τ is a calibration outside them that corresponds to no (ε, δ)-transitive cones with a bounded budget.

This difference appears prominently in the region of small K. On WordNet nouns at $d = 5$ (Figure 7), the containment criterion of the $(\varepsilon, 0)$ cones falls below 2 for $K \leq 3 \cdot 10^{-4}$, while the threshold criterion stays at 80.4. For small K, the aperture falls below the margin and the trained cones contain almost nothing. Even so, applying a threshold to the angle yields a pair-classification accuracy above 80. The predictions in this case are not predictions of containment, but the result of thresholding the ordering of exterior angles produced by training. For these reasons, the main text uses the containment criterion as the primary metric, and the values of the threshold criterion are reported alongside it below.

J.1 Sweep of K for all datasets

Figures 7 and 8 sweep K under the protocol of Ganea et al. for the three datasets, the four aperture formulas, and both criteria, at a percentage of non-basic edges of 50%. No post-hoc selection of K is applied: every point is evaluated at the K used for training. The corresponding sweep under the protocol of Shadow Cones is given in Appendix L.

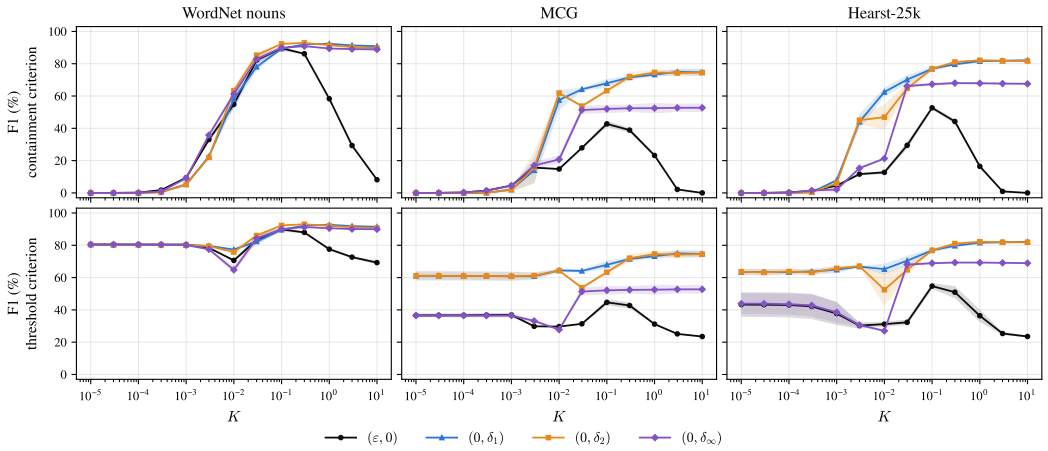

Figure 7: Sweep of K under the protocol of Ganea et al. at $d = 5$ (percentage of non-basic edges 50%). The top row is the containment criterion and the bottom row the threshold criterion; the lines are the means over 3 seeds and the bands their standard deviations.

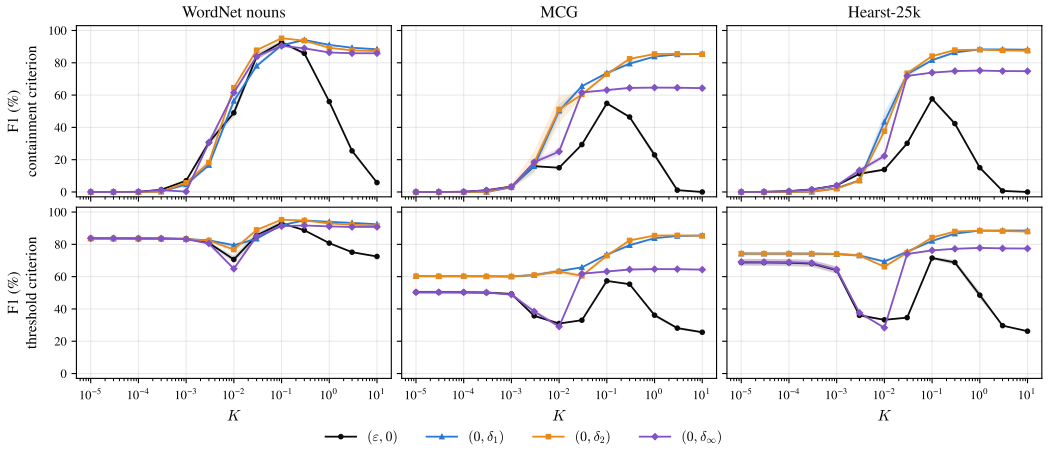

Figure 8: Sweep of K under the protocol of Ganea et al. at $d = 10$. The conventions are those of Figure 7.

J.2 FULL VERSION OF TABLE 2

Table 4: Full version of F1 (%) under the containment criterion. Same models and criterion as Table 2. For each method the upper row is $d = 5$ and the lower row $d = 10$; each value is the mean and standard deviation ($\pm$) over 10 seeds. Bold marks the column best within each block (protocol or model) across both $d = 5, 10$; green background marks the column best across all rows including reference methods. Colored numbers after each value: difference from the $(\varepsilon, 0)$ cones of the same block.

WordNet nouns	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	22.7±0.5	35.0±1.2	38.3±0.5	40.1±0.3	45.3±0.4	54.8±0.5
	10	25.8±0.7	36.0±0.5	38.8±0.3	40.4±0.3	45.1±0.2	54.4±0.2
order embedding	5	35.4±1.1	69.6±1.1	76.5±1.0	81.8±0.7	86.8±1.1	98.3±0.1
	10	43.1±1.0	69.7±0.8	78.3±1.2	83.7±0.7	88.3±0.8	99.8±0.0
Protocol of Ganea et al.							
Euclidean cone	5	11.0±1.6	55.7±2.6	68.5±1.2	73.3±1.2	75.8±1.5	79.2±1.5
	10	8.7±1.1	62.2±1.3	72.2±1.1	74.7±1.0	77.0±1.0	80.3±0.7
$(\varepsilon, 0)$ cones	5	17.7±0.9	73.5±3.8	83.9±3.1	90.1±1.7	95.7±0.6	97.3±0.4
	10	15.9±0.5	80.0±1.4	87.7±1.7	93.2±1.3	97.0±0.2	**98.3**±0.2
$(0, \delta_1)$ cones	5	30.7±0.6 +13.0	76.4±2.8 +2.9	83.6±1.5 -0.3	88.7±1.0 -1.4	93.7±0.9 -2.0	94.5±0.5 -2.8
	10	32.7±0.9 +16.8	82.5±0.6 +2.5	87.6±0.9 -0.1	91.1±0.6 -2.1	95.6±0.3 -1.4	96.4±0.3 -1.9
$(0, \delta_2)$ cones	5	32.4±0.6 +14.7	79.0±3.1 +5.5	86.9±1.8 +3.0	91.6±0.8 +1.5	95.5±0.7 -0.2	95.4±0.3 -1.9
	10	**35.0**±0.8 +19.1	**85.8**±0.8 +5.8	**92.5**±0.7 +4.8	**95.2**±0.3 +2.0	**97.4**±0.1 +0.4	97.4±0.2 -0.9
$(0, \delta_\infty)$ cones	5	26.4±0.6 +8.7	76.2±3.7 +2.7	84.7±1.7 +0.8	89.9±1.3 -0.2	94.0±1.1 -1.7	95.4±0.7 -1.9
	10	27.3±0.7 +11.4	80.8±0.9 +0.8	86.9±0.8 -0.8	90.9±0.7 -2.3	96.6±0.6 -0.4	97.8±0.6 -0.5
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	17.8±0.5	74.1±4.0	83.8±3.0	90.5±0.9	95.5±0.6	97.4±0.3
	10	15.7±1.2	78.1±2.2	88.1±1.8	93.5±0.6	97.0±0.3	**98.4**±0.2
$(0, \delta_1)$ cones	5	30.9±0.7 +13.1	76.3±2.7 +2.2	83.7±1.4 -0.1	88.6±1.2 -1.9	93.7±0.9 -1.8	94.9±0.5 -2.5
	10	32.6±0.8 +16.9	82.5±0.8 +4.4	87.9±0.7 -0.2	91.0±0.6 -2.5	95.7±0.4 -1.3	96.5±0.3 -1.9
$(0, \delta_2)$ cones	5	32.4±0.5 +14.6	78.8±3.3 +4.7	86.9±1.7 +3.1	91.9±1.0 +1.4	95.4±0.5 -0.1	95.5±0.6 -1.9
	10	**35.0**±0.9 +19.3	**85.6**±0.9 +7.5	**92.5**±0.5 +4.4	**95.1**±0.3 +1.6	**97.3**±0.2 +0.3	97.5±0.2 -0.9
$(0, \delta_\infty)$ cones	5	26.4±0.6 +8.6	76.3±3.9 +2.2	85.0±1.4 +1.2	90.2±1.0 -0.3	94.4±1.0 -1.1	95.3±0.6 -2.1
	10	27.3±0.6 +11.6	81.2±0.8 +3.1	86.8±0.8 -1.3	91.0±0.6 -2.5	96.6±0.6 -0.4	97.9±0.5 -0.5

MCG	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	16.8±0.1	24.0±0.2	23.4±0.1	23.0±0.1	23.5±0.2	24.0±0.2
	10	16.7±0.0	24.5±0.1	23.9±0.1	23.5±0.1	23.9±0.1	24.7±0.3
order embedding	5	28.6±0.9	54.3±0.5	52.0±1.3	50.6±1.8	52.1±1.5	54.1±2.1
	10	34.3±0.6	66.0±0.6	71.4±0.9	73.5±1.8	76.5±3.4	82.0±3.9
Protocol of Ganea et al.							
Euclidean cone	5	6.5±0.6	53.6±0.5	59.0±0.8	61.1±0.7	60.5±0.3	59.3±0.6
	10	4.9±0.6	54.6±1.0	58.2±0.4	59.3±0.5	58.0±0.6	57.3±0.5
$(\varepsilon, 0)$ cones	5	5.5±0.4	52.9±1.3	49.6±1.6	46.6±4.3	53.4±4.9	65.4±4.2
	10	3.8±0.3	52.4±1.0	51.5±2.0	54.3±1.4	65.2±2.3	71.6±3.7
$(0, \delta_1)$ cones	5	18.1±0.9 +12.6	60.3±0.8 +7.4	63.1±0.8 +13.5	68.1±2.3 +21.5	77.0±1.5 +23.6	79.0±1.3 +13.6
	10	16.1±0.8 +12.3	**62.7**±0.7 +10.3	**68.2**±1.0 +16.7	**73.2**±1.5 +18.9	79.6±1.3 +14.4	81.8±1.5 +10.2
$(0, \delta_2)$ cones	5	**22.2**±1.2 +16.7	57.6±0.7 +4.7	59.3±1.6 +9.7	64.9±1.6 +18.3	75.4±1.5 +22.0	76.0±1.4 +10.6
	10	22.1±0.7 +18.3	61.8±0.7 +9.4	66.0±0.9 +14.5	72.8±1.1 +18.5	**80.0**±0.9 +14.8	81.9±1.4 +10.3
$(0, \delta_\infty)$ cones	5	11.4±0.5 +5.9	50.2±1.5 -2.7	52.8±1.1 +3.2	55.9±3.5 +9.3	63.0±3.5 +9.6	72.6±3.4 +7.2
	10	9.5±0.4 +5.7	53.5±1.1 +1.1	56.7±1.7 +5.2	64.1±1.1 +9.8	77.0±0.9 +11.8	**83.9**±0.8 +12.3
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	5.5±0.6	53.1±1.1	49.8±1.4	46.7±4.5	53.7±4.9	65.5±4.2
	10	3.8±0.4	52.3±0.8	51.0±1.8	54.2±1.7	64.5±2.2	71.5±3.4
$(0, \delta_1)$ cones	5	18.1±1.0 +12.6	61.5±0.8 +8.4	64.6±1.1 +14.8	71.4±1.8 +24.7	81.4±0.7 +27.7	82.7±0.8 +17.2
	10	16.3±0.8 +12.5	**63.8**±0.5 +11.5	**70.1**±0.8 +19.1	**76.4**±1.0 +22.2	**85.2**±0.4 +20.7	**87.0**±0.3 +15.5
$(0, \delta_2)$ cones	5	**22.1**±1.1 +16.6	59.3±0.6 +6.2	60.6±1.3 +10.8	67.7±2.4 +21.0	79.3±0.8 +25.6	80.7±1.2 +15.2
	10	**22.1**±0.9 +18.3	62.7±0.7 +10.4	67.5±0.6 +16.5	74.5±1.0 +20.3	84.1±0.6 +19.6	86.0±0.5 +14.5
$(0, \delta_\infty)$ cones	5	11.3±0.4 +5.8	50.2±1.5 -2.9	52.7±1.1 +2.9	55.8±3.4 +9.1	62.8±3.8 +9.1	72.5±3.3 +7.0
	10	9.5±0.4 +5.7	53.6±1.0 +1.3	56.7±1.9 +5.7	64.1±1.2 +9.9	77.0±1.1 +12.5	83.9±0.9 +12.4

Hearst-25k	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	17.5±0.1	23.1±0.1	23.0±0.1	23.1±0.1	23.7±0.1	24.0±0.3
	10	17.4±0.1	23.4±0.1	23.2±0.0	23.3±0.0	23.8±0.1	24.2±0.1
order embedding	5	25.8±0.6	61.0±1.0	61.0±1.2	61.0±2.3	69.7±2.7	70.3±3.3
	10	29.5±0.8	72.9±1.0	78.4±0.7	81.9±1.1	89.9±1.3	92.1±1.9
Protocol of Ganea et al.							
Euclidean cone	5	1.2±0.3	67.3±0.3	70.0±0.2	71.2±0.4	70.9±0.3	70.5±0.5
	10	0.7±0.0	66.9±0.4	69.3±0.5	70.7±0.3	71.0±0.2	70.3±0.3
$(\varepsilon, 0)$ cones	5	1.2±0.2	56.7±1.9	49.9±1.6	49.4±9.3	67.3±3.3	78.3±8.5
	10	1.0±0.1	54.1±1.6	50.7±1.7	56.8±1.3	70.6±1.1	85.6±1.7
$(0, \delta_1)$ cones	5	13.1±0.4 +11.9	67.1±1.1 +10.4	70.6±1.0 +20.7	76.3±1.3 +26.9	82.0±1.5 +14.7	83.4±1.1 +5.1
	10	11.1±0.5 +10.1	**69.8**±1.5 +15.7	75.7±0.6 +25.0	82.4±0.6 +25.6	85.9±1.0 +15.3	87.3±1.5 +1.7
$(0, \delta_2)$ cones	5	**16.4**±0.3 +15.2	66.5±1.2 +9.8	70.0±1.5 +20.1	77.0±1.0 +27.6	81.9±1.4 +14.6	83.4±1.7 +5.1
	10	15.2±0.6 +14.2	68.2±1.9 +14.1	**76.1**±1.1 +25.4	**84.9**±0.7 +28.1	**88.0**±0.9 +17.4	89.5±1.1 +3.9
$(0, \delta_\infty)$ cones	5	2.9±0.3 +1.7	59.7±1.7 +3.0	60.2±2.2 +10.3	66.4±2.7 +17.0	77.3±1.4 +10.0	83.3±2.8 +5.0
	10	2.5±0.2 +1.5	60.1±2.8 +6.0	64.0±2.2 +13.3	73.5±1.0 +16.7	84.9±0.3 +14.3	**91.0**±0.4 +5.4
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	1.2±0.1	56.9±1.7	50.7±1.5	49.5±9.1	67.0±3.5	78.7±8.0
	10	1.0±0.1	54.6±1.4	50.6±1.6	57.2±1.6	70.1±2.1	85.5±1.5
$(0, \delta_1)$ cones	5	13.0±0.4 +11.8	68.8±1.0 +11.9	72.6±1.7 +21.9	78.7±1.0 +29.2	85.3±0.9 +18.3	86.6±1.1 +7.9
	10	11.1±0.4 +10.1	**70.0**±1.3 +15.4	76.9±1.2 +26.3	84.5±0.6 +27.3	89.5±0.4 +19.4	90.6±0.8 +5.1
$(0, \delta_2)$ cones	5	**16.4**±0.5 +15.2	67.9±1.3 +11.0	73.2±1.3 +22.5	79.0±1.1 +29.5	84.9±1.2 +17.9	86.5±1.3 +7.8
	10	15.2±0.6 +14.2	69.3±1.4 +14.7	**77.6**±0.9 +27.0	**85.5**±0.6 +28.3	**90.8**±0.5 +20.7	**91.9**±0.7 +6.4
$(0, \delta_\infty)$ cones	5	2.9±0.3 +1.7	60.0±1.7 +3.1	60.8±2.2 +10.1	66.4±3.0 +16.9	77.3±1.4 +10.3	83.3±2.6 +4.6
	10	2.5±0.2 +1.5	60.7±2.5 +6.1	64.4±2.1 +13.8	73.6±1.0 +16.4	85.0±0.3 +14.9	91.0±0.4 +5.5

Table 5: Full version of F1 (%) under the threshold criterion. The same models as Table 2 evaluated with the threshold criterion. For each method the upper row is $d = 5$ and the lower row $d = 10$; each value is the mean and standard deviation (±) over 10 seeds. Bold marks the column best within each block (protocol or model) across both $d = 5, 10$; green background marks the column best across all rows including reference methods. Colored numbers after each value: difference from the $(\varepsilon, 0)$ cones of the same block.

WordNet nouns	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	22.7±0.5	35.0±1.2	38.3±0.5	40.1±0.3	45.3±0.4	54.8±0.5
	10	25.8±0.7	36.0±0.5	38.8±0.3	40.4±0.3	45.1±0.2	54.4±0.2
order embedding	5	35.4±1.1	69.6±1.1	76.5±1.0	81.8±0.7	86.8±1.1	98.3±0.1
	10	43.1±1.0	69.7±0.8	78.3±1.2	83.7±0.7	88.3±0.8	99.8±0.0
Protocol of Ganea et al.							
Euclidean cone	5	28.1±0.4	56.6±2.5	69.5±1.3	74.0±1.1	76.5±1.4	80.2±1.5
	10	30.0±0.3	62.6±1.4	72.8±1.1	75.3±1.0	77.7±1.1	81.3±0.9
$(\varepsilon, 0)$ cones	5	29.5±0.3	75.4±3.5	84.6±2.8	90.3±1.7	95.7±0.6	97.5±0.4
	10	31.1±0.6	82.8±1.2	89.7±1.2	93.7±1.0	97.1±0.3	**98.4**±0.2
$(0, \delta_1)$ cones	5	31.3±0.4 +1.8	77.3±2.9 +1.9	84.1±1.5 -0.5	89.1±1.0 -1.2	93.8±0.9 -1.9	94.6±0.5 -2.9
	10	33.1±0.8 +2.0	83.2±0.6 +0.4	88.6±1.0 -1.1	92.1±0.3 -1.6	95.8±0.2 -1.3	96.6±0.2 -1.8
$(0, \delta_2)$ cones	5	32.4±0.5 +2.9	79.0±3.2 +3.6	86.9±1.8 +2.3	91.7±0.8 +1.4	95.5±0.7 -0.2	95.4±0.3 -2.1
	10	**34.6**±0.8 +3.5	**86.5**±0.7 +3.7	**92.9**±0.5 +3.2	**95.2**±0.2 +1.5	**97.5**±0.2 +0.4	97.4±0.2 -1.0
$(0, \delta_\infty)$ cones	5	29.4±0.3 -0.1	76.3±3.7 +0.9	85.0±1.7 +0.4	90.1±1.2 -0.2	94.1±1.1 -1.6	95.5±0.7 -2.0
	10	31.2±0.6 +0.1	81.4±0.7 -1.4	87.8±0.7 -1.9	91.9±0.7 -1.8	96.8±0.5 -0.3	97.9±0.6 -0.5
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	29.5±0.3	75.8±3.9	84.6±2.6	90.7±0.9	95.6±0.6	97.5±0.3
	10	31.1±0.6	81.9±2.5	89.8±1.2	93.8±0.6	97.1±0.3	**98.5**±0.2
$(0, \delta_1)$ cones	5	31.6±0.6 +2.1	77.3±2.9 +1.5	84.3±1.5 -0.3	88.9±1.1 -1.8	93.8±0.9 -1.8	95.0±0.5 -2.5
	10	33.1±0.7 +2.0	83.1±0.9 +1.2	88.9±0.7 -0.9	92.0±0.7 -1.8	95.8±0.4 -1.3	96.6±0.4 -1.9
$(0, \delta_2)$ cones	5	32.3±0.4 +2.8	78.8±3.2 +3.0	86.9±1.8 +2.3	91.8±1.0 +1.1	95.4±0.5 -0.2	95.5±0.6 -2.0
	10	**34.8**±0.9 +3.7	**86.5**±0.9 +4.6	**92.8**±0.5 +3.0	**95.2**±0.3 +1.4	**97.4**±0.2 +0.3	97.5±0.2 -1.0
$(0, \delta_\infty)$ cones	5	29.4±0.3 -0.1	76.4±3.9 +0.6	85.2±1.5 +0.6	90.4±1.0 -0.3	94.5±1.0 -1.1	95.4±0.6 -2.1
	10	31.2±0.6 +0.1	81.7±0.8 -0.2	87.9±0.6 -1.9	91.8±0.5 -2.0	96.8±0.5 -0.3	98.0±0.5 -0.5

MCG	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	16.8±0.1	24.0±0.2	23.4±0.1	23.0±0.1	23.5±0.2	24.0±0.2
	10	16.7±0.0	24.5±0.1	23.9±0.1	23.5±0.1	23.9±0.1	24.7±0.3
order embedding	5	28.6±0.9	54.3±0.5	52.0±1.3	50.6±1.8	52.1±1.5	54.1±2.1
	10	34.3±0.6	66.0±0.6	71.4±0.9	73.5±1.8	76.5±3.4	82.0±3.9
Protocol of Ganea et al.							
Euclidean cone	5	22.2±0.6	53.6±0.5	59.0±0.8	61.1±0.7	60.5±0.3	59.3±0.6
	10	23.1±0.6	56.8±0.3	60.9±0.6	59.8±0.3	58.7±0.5	57.9±0.3
$(\varepsilon, 0)$ cones	5	22.9±0.4	54.6±1.3	51.6±1.3	48.1±4.2	53.6±5.1	65.8±4.4
	10	23.8±0.3	55.7±1.0	54.9±1.5	57.2±1.1	67.2±2.4	72.9±3.9
$(0, \delta_1)$ cones	5	26.1±0.3 +3.2	60.6±0.8 +6.0	63.1±0.8 +11.5	68.1±2.3 +20.0	77.0±1.5 +23.4	79.0±1.3 +13.2
	10	26.6±0.3 +2.8	**63.4**±0.8 +7.7	**68.2**±0.9 +13.3	**73.2**±1.5 +16.0	79.6±1.3 +12.4	81.8±1.5 +8.9
$(0, \delta_2)$ cones	5	26.5±0.3 +3.6	57.6±0.7 +3.0	59.3±1.6 +7.7	64.9±1.6 +16.8	75.4±1.5 +21.8	76.0±1.4 +10.2
	10	**27.0**±0.3 +3.2	61.8±0.6 +6.1	66.0±0.9 +11.1	72.8±1.1 +15.6	**80.0**±0.9 +12.8	81.9±1.4 +9.0
$(0, \delta_\infty)$ cones	5	22.9±0.4 +0.0	49.7±1.6 -4.9	52.6±1.1 +1.0	56.0±3.5 +7.9	63.0±3.6 +9.4	72.6±3.4 +6.8
	10	23.8±0.3 +0.0	53.0±1.0 -2.7	56.6±1.8 +1.7	64.2±1.2 +7.0	77.2±0.9 +10.0	**84.0**±0.8 +11.1
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	22.9±0.4	54.7±1.2	51.8±1.3	48.2±4.5	54.0±5.2	65.9±4.3
	10	23.8±0.3	55.8±0.9	54.8±1.5	56.8±1.6	66.8±2.6	72.7±3.6
$(0, \delta_1)$ cones	5	26.2±0.3 +3.3	61.8±0.8 +7.1	64.6±1.1 +12.8	71.4±1.8 +23.2	81.4±0.7 +27.4	82.7±0.8 +16.8
	10	26.7±0.3 +2.9	**64.5**±0.6 +8.7	**70.1**±0.8 +15.3	**76.4**±1.0 +19.6	**85.2**±0.4 +18.4	**87.0**±0.3 +14.3
$(0, \delta_2)$ cones	5	26.5±0.3 +3.6	59.3±0.6 +4.6	60.6±1.3 +8.8	67.7±2.4 +19.5	79.3±0.8 +25.3	80.7±1.2 +14.8
	10	**27.0**±0.2 +3.2	62.7±0.7 +6.9	67.5±0.6 +12.7	74.5±1.0 +17.7	84.1±0.6 +17.3	86.0±0.5 +13.3
$(0, \delta_\infty)$ cones	5	22.9±0.4 +0.0	49.8±1.6 -4.9	52.6±1.1 +0.8	55.8±3.4 +7.6	62.8±3.8 +8.8	72.5±3.3 +6.6
	10	23.8±0.3 +0.0	53.0±1.0 -2.8	56.6±1.9 +1.8	64.1±1.1 +7.3	77.2±1.1 +10.4	84.0±0.9 +11.3

Hearst-25k	d	non-basic edges (%) 0	10	25	50	90	recon.
Reference methods							
Poincaré embedding	5	17.5±0.1	23.1±0.1	23.0±0.1	23.1±0.1	23.7±0.1	24.0±0.3
	10	17.4±0.1	23.4±0.1	23.2±0.0	23.3±0.0	23.8±0.1	24.2±0.1
order embedding	5	25.8±0.6	61.0±1.0	61.0±1.2	61.0±2.3	69.7±2.7	70.3±3.3
	10	29.5±0.8	72.9±1.0	78.4±0.7	81.9±1.1	89.9±1.3	92.1±1.9
Protocol of Ganea et al.							
Euclidean cone	5	19.2±0.4	71.0±1.5	70.2±0.3	73.0±0.2	70.9±0.3	70.5±0.5
	10	20.2±0.4	**77.2**±0.2	**79.9**±0.1	78.2±0.2	77.2±0.1	77.4±0.1
$(\varepsilon, 0)$ cones	5	17.9±0.3	59.9±1.9	57.2±2.6	51.7±8.9	71.6±3.5	80.9±6.6
	10	19.2±0.4	59.8±1.7	61.2±2.3	70.4±2.4	76.4±0.7	87.1±1.8
$(0, \delta_1)$ cones	5	24.0±0.2 +6.1	67.7±1.1 +7.8	71.2±1.0 +14.0	76.3±1.3 +24.6	82.0±1.5 +10.4	83.4±1.2 +2.5
	10	**24.4**±0.2 +5.2	70.3±1.4 +10.5	76.5±0.7 +15.3	82.8±0.5 +12.4	86.0±0.9 +9.6	87.3±1.4 +0.2
$(0, \delta_2)$ cones	5	24.1±0.2 +6.2	66.5±1.2 +6.6	70.0±1.5 +12.8	77.0±1.0 +25.3	81.9±1.4 +10.3	83.4±1.7 +2.5
	10	**24.4**±0.2 +5.2	68.3±1.8 +8.5	76.3±1.0 +15.1	**84.9**±0.7 +14.5	**88.0**±0.9 +11.6	89.5±1.1 +2.4
$(0, \delta_\infty)$ cones	5	17.9±0.3 +0.0	59.7±1.7 -0.2	60.9±2.4 +3.7	67.7±3.4 +16.0	78.3±1.3 +6.7	83.7±2.7 +2.8
	10	19.3±0.4 +0.1	60.1±2.8 +0.3	64.9±2.5 +3.7	75.8±1.2 +5.4	85.9±0.3 +9.5	**91.3**±0.4 +4.2
Lorentz model (protocol of Ganea et al.)							
$(\varepsilon, 0)$ cones	5	17.9±0.2	60.1±1.8	58.1±2.4	51.9±8.9	71.4±4.0	80.8±6.7
	10	19.2±0.4	60.2±1.5	61.4±2.2	70.9±2.3	76.7±1.0	87.0±1.6
$(0, \delta_1)$ cones	5	24.0±0.1 +6.1	69.4±1.1 +9.3	73.3±1.7 +15.2	78.7±1.0 +26.8	85.3±0.9 +13.9	86.6±1.1 +5.8
	10	24.4±0.2 +5.2	**70.6**±1.3 +10.4	**77.7**±1.2 +16.3	84.8±0.5 +13.9	89.5±0.4 +12.8	90.6±0.8 +3.6
$(0, \delta_2)$ cones	5	24.0±0.1 +6.1	67.9±1.3 +7.8	73.2±1.3 +15.1	79.0±1.1 +27.1	84.9±1.2 +13.5	86.5±1.3 +5.7
	10	**24.5**±0.2 +5.3	69.4±1.4 +9.2	**77.7**±0.8 +16.3	**85.5**±0.6 +14.6	**90.8**±0.5 +14.1	**91.9**±0.7 +4.9
$(0, \delta_\infty)$ cones	5	18.0±0.3 +0.1	60.0±1.7 -0.1	61.5±2.5 +3.4	67.7±3.6 +15.8	78.4±1.3 +7.0	83.7±2.4 +2.9
	10	19.3±0.4 +0.1	60.7±2.5 +0.5	65.4±2.4 +4.0	75.9±1.1 +5.0	86.0±0.3 +9.3	91.3±0.3 +4.3

Table 6: F1 (%) under the containment criterion on the original Hearst dataset (percentages 0–10%). Since its transitive closure is large (6.8M edges) and the settings of 25% or higher and reconstruction cannot be run, the main table uses the rebuilt Hearst-25k with the cutoff halved, and the original is reported here. For each method the upper row is $d = 5$ and the lower row $d = 10$; each value is the mean and standard deviation ($\pm$) over 10 seeds. The rules for bold, the green background, and the colored numbers are as in Table 4. Values under the protocol of Shadow Cones are given in Appendix L.

Hearst	d	non-basic edges (%)				
		0	1	2	5	10
Reference methods						
Poincaré embedding	5	17.2±0.0	21.3±0.0	21.8±0.0	21.9±0.1	21.8±0.1
	10	17.1±0.0	21.4±0.1	22.1±0.0	22.3±0.0	22.1±0.1
order embedding	5	22.6±0.5	45.3±0.6	48.8±0.8	49.3±0.5	47.2±0.8
	10	26.9±0.4	44.9±0.3	55.3±0.4	63.7±0.6	66.0±0.8
Protocol of Ganea et al.						
Euclidean cone	5	1.2±0.3	50.7±0.8	59.1±0.4	63.1±0.3	63.9±0.5
	10	0.7±0.1	51.2±0.7	57.8±0.7	61.1±0.4	61.9±0.2
$(\varepsilon, 0)$ cones	5	1.2±0.1	40.2±3.2	52.0±1.3	46.8±2.1	55.6±1.5
	10	0.8±0.1	38.1±1.6	46.5±1.2	43.9±1.3	50.9±1.8
$(0, \delta_1)$ cones	5	11.6±0.3 +10.4	48.8±2.3 +8.6	59.1±0.9 +7.1	64.1±1.0 +17.3	71.5±0.5 +15.9
	10	8.8±0.1 +8.0	**52.9**±0.5 +14.8	**60.0**±0.7 +13.5	**67.0**±0.6 +23.1	**73.4**±0.7 +22.5
$(0, \delta_2)$ cones	5	**15.2**±0.5 +14.0	48.0±2.2 +7.8	57.4±0.9 +5.4	60.8±1.4 +14.0	69.0±0.7 +13.4
	10	13.5±0.1 +12.7	51.4±0.5 +13.3	57.8±0.7 +11.3	64.0±0.6 +20.1	72.7±0.6 +21.8
$(0, \delta_\infty)$ cones	5	3.0±0.1 +1.8	44.8±2.3 +4.6	53.4±0.9 +1.4	53.0±2.6 +6.2	61.5±1.6 +5.9
	10	2.4±0.1 +1.6	47.4±0.6 +9.3	52.4±1.1 +5.9	55.5±0.8 +11.6	61.4±1.9 +10.5
Lorentz model (protocol of Ganea et al.)						
$(\varepsilon, 0)$ cones	5	1.2±0.1	40.0±3.5	52.0±1.2	47.3±2.2	55.6±1.6
	10	0.8±0.1	37.8±1.4	46.5±1.1	43.9±1.3	51.2±2.1
$(0, \delta_1)$ cones	5	11.6±0.3 +10.4	49.3±2.2 +9.3	60.2±0.8 +8.2	64.8±1.0 +17.5	72.4±0.6 +16.8
	10	8.8±0.1 +8.0	**53.4**±0.5 +15.6	**60.5**±0.9 +14.0	**67.3**±0.3 +23.4	**73.9**±0.7 +22.7
$(0, \delta_2)$ cones	5	**15.2**±0.5 +14.0	48.5±2.3 +8.5	58.4±0.9 +6.4	62.3±1.4 +15.0	70.2±0.7 +14.6
	10	13.5±0.2 +12.7	52.0±0.5 +14.2	58.8±0.8 +12.3	64.6±0.4 +20.7	72.9±0.7 +21.7
$(0, \delta_\infty)$ cones	5	3.0±0.1 +1.8	44.9±2.3 +4.9	53.5±1.0 +1.5	53.4±2.6 +6.1	62.0±1.6 +6.4
	10	2.4±0.1 +1.6	47.3±0.6 +9.5	52.6±1.1 +6.1	55.9±0.8 +12.0	61.5±1.9 +10.3

K Formulation on the Lorentz Model

Lorentz model The Lorentz (hyperboloid) model of hyperbolic space with curvature $-c$ ($c > 0$) is the hyperboloid

$$\mathcal{L}_c^n := \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; \langle x, x \rangle_{\mathcal{L}} = -1/c,\ x_0 > 0 \right\}$$

in $\mathbb{R}^{n+1}$ equipped with the Minkowski inner product $\langle x, y \rangle_{\mathcal{L}} := -x_0 y_0 + \sum_{i=1}^{n} x_i y_i$. Hereafter we split points into time and space components as $x = (x_0, x_s)$ ($x_0 \in \mathbb{R}$, $x_s \in \mathbb{R}^n$, $x_0 = \sqrt{1/c + \|x_s\|^2}$). The geodesic distance between two points is $d(x, y) = \frac{1}{\sqrt{c}} \operatorname{arccosh}\big(-c \langle x, y \rangle_{\mathcal{L}}\big)$, and the point corresponding to the origin is $O = (1/\sqrt{c}, \mathbf{0})$. For $c = 1$ the model is isometric to the Poincaré ball model of the main text, via the correspondence $x \mapsto x_s/(1 + x_0)$. For the geodesic distance $t = d(O, x)$ from the origin we have $\sqrt{c}\,\|x_s\| = \sinh(\sqrt{c}\,t)$; for $c = 1$, $\|x_s\| = \sinh t = 2h_{\max}(r)$, that is, the norm of the space component corresponds to $h_{\max}$ of the main text.

Aperture and loss Following MERU (Desai et al., 2023), for arbitrary curvature c define the normalized coordinate

$$s(x) := \frac{\sqrt{c}\,\|x_s\|}{2K} = \frac{\sinh(\sqrt{c}\,t)}{2K} \tag{15}$$

(which coincides with $s = h_{\max}(r)/K$ of §5 for $c = 1$); then the apertures of the $(\varepsilon, 0)$, $(0, \delta_1)$, and $(0, \delta_2)$ cones are all functions of s only:

$$\psi_K(x) = \arcsin\left(\frac{1}{s}\right), \qquad \psi_1(x) = \arcsin\left(\frac{1 - e^{-s}}{s}\right), \qquad \psi_2(x) = \arcsin\left(\frac{\operatorname{erf}\left(\frac{\sqrt{\pi}}{2} s\right)}{s}\right). \tag{16}$$

The original ψ_K is infeasible for $s < 1$, that is, inside the ball of radius $t_0 = \operatorname{arcsinh}(2K)/\sqrt{c}$ centered at the origin, and implementations must clip the argument of $\arcsin$ to at most 1, which coincides with the limit $q \to \infty$ of Corollary 5.3 (the implementation used here clips to $1 - 10^{-3}$, so the constant aperture is $87.4°$), whereas the $(0, \delta_1)$ and $(0, \delta_2)$ cones are defined on the whole

domain except the origin. The exterior angle is given by

$$\Xi(x,y) \;=\; \arccos\left(\frac{y_0 + x_0\, c\,\langle x,y\rangle_{\mathcal{L}}}{\|x_s\|\sqrt{\big(c\,\langle x,y\rangle_{\mathcal{L}}\big)^2 - 1}}\right) \tag{17}$$

(Desai et al., 2023), and the loss is $f(x,y) = \max\big(0,\ \Xi(x,y) - \psi(x)\big)$, as in the main text.

Treatment of the curvature in the experiments In the hierarchical word embedding experiments (the Lorentz rows of Table 4), c is fixed to 1 and the comparison is isometric to the Poincaré-ball formulation of the main text. In the MERU experiments, on the other hand, c is a learnable parameter following the original setup, and (15)–(17) are evaluated at the current value of c during training.

L Comparison with Shadow Cones

This section presents the comparison with Shadow Cones (Yu et al., 2024), separated from the main table (Table 2). Shadow Cones have two formulations on the Poincaré ball model (umbral, the umbra of a point source, and penumbral, the penumbra of a spherical source) and two on the half-space model (umbral and penumbral). Among them, the penumbral formulation on the Poincaré ball has the same geometry as the $(\varepsilon, 0)$ cones (the radius of the light source corresponds to ε), so we also compare the penumbral $(0, \delta_1)$, $(0, \delta_2)$, and $(0, \delta_\infty)$ cones, whose aperture is replaced accordingly, under the same training protocol as penumbral. The two half-space formulations and umbral change the coordinate model or the conditions of the cones, so, as stated in §6, they are excluded from the main comparison and shown only here for reference. The protocol of Shadow Cones has no initialization phase; the embeddings are initialized uniformly at random near the origin and, for the original penumbral and umbral formulations, projected to the outside of the sphere of the light source before training starts. The cones with replaced apertures have no infeasible region, so this projection is not applied to them. The loss is the contrastive loss of the original work ($\gamma_2 = 0$, $\gamma_3 = 10^{-4}$), with batch size 16, 10 negatives per positive pair drawn uniformly, and 400 epochs on nouns and 500 epochs on MCG and Hearst-25k. The first 20 epochs are a burn-in with the learning rate divided by 100. The margin γ_1 is 10^{-4} at 0% and 10^{-2} otherwise (10^{-3} for the two formulations on the Poincaré ball). The optimizer is Riemannian SGD on the half-space model and Riemannian Adam on the Poincaré ball and Lorentz models, the curvature is 1, the training radius of umbral is 0.05, and the height of the light source of the half-space penumbral is 20. The learning rate is selected by validation F1 for each pair of method and dataset from the candidates $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}\}$. Evaluation is performed every 20 epochs, and the model at the epoch with the best validation F1 is used. At evaluation time, the radius of umbral is selected from $\{0.01, 0.05, 0.06, 0.1, 0.2, 0.3\}$, the height of the half-space penumbral from $\{2, 5, 10, 20, 21, 25\}$, and K of the cones with replaced apertures from the same candidates as in the protocol of Ganea et al., all by validation F1 (Appendix J). Training is run in a single process rather than the 8-process asynchronous parallelism of the original work (we confirmed with the public implementation of the original work that the two give the same accuracy). In addition, the public implementation uses a squared length as a length in the computation of the aperture of umbral on the Poincaré ball and cannot run in this setting, so we corrected this part. Each condition uses 10 seeds, and reported values are their averages.

Tables 7 and 8 report F1 under the containment criterion, including all percentages and the Lorentz model. The half-space umbral formulation attains the best value in many settings, and the formulations on the Poincaré ball do not reach it. This is the result of changing both the coordinate model and the conditions of the cones, and is a different arena from the drop-in replacement of the aperture of the $(\varepsilon, 0)$ cones on the Poincaré and Lorentz models that this paper targets. Restricted to penumbral on the Poincaré ball, replacing the aperture is effective on data with many nodes that have two or more parents. The penumbral $(0, \delta_2)$ cones outperform the penumbral $(\varepsilon, 0)$ cones at 25% or higher on MCG (94.7 versus 92.2 at 90% for $d = 10$), and are at least as good at 25% or higher on Hearst-25k for $d = 10$. On nouns, in contrast, the penumbral $(\varepsilon, 0)$ cones win at 25% or higher, with a difference of at most 4.6 points. Note, however, that the best accuracy on nouns is attained by the $(0, \delta_2)$ cones under the protocol of Ganea et al. (Table 2), suggesting that the protocol of Shadow Cones may not be suited to nouns.

Sweep of K Figures 9 and 10 sweep K under the protocol of Shadow Cones, in the same format as Figures 7 and 8. The cones compared here are those on the Poincaré ball, whose penumbral formulation has the geometry of the $(\varepsilon, 0)$ cones.

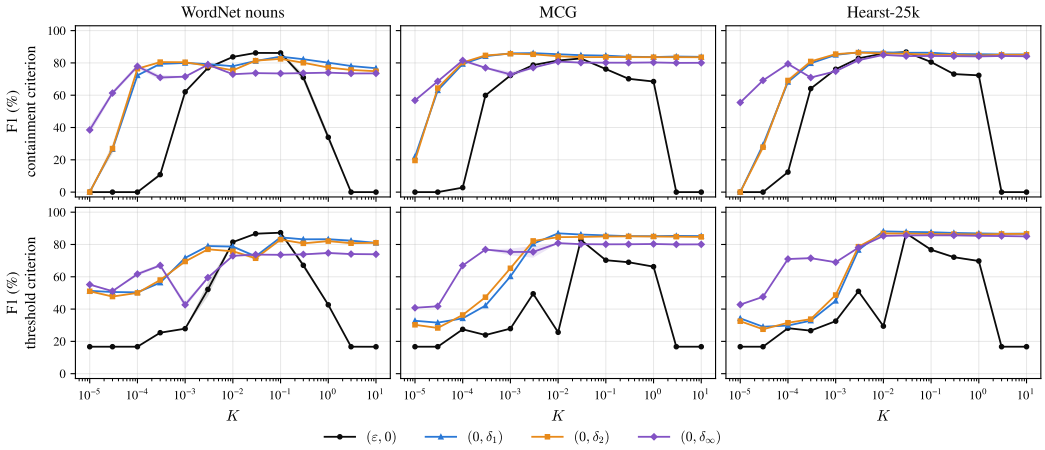

Figure 9: Sweep of K under the protocol of Shadow Cones at $d = 5$ (percentage of non-basic edges 50%). The top row is the containment criterion and the bottom row the threshold criterion; the lines are the means over 3 seeds and the bands their standard deviations.

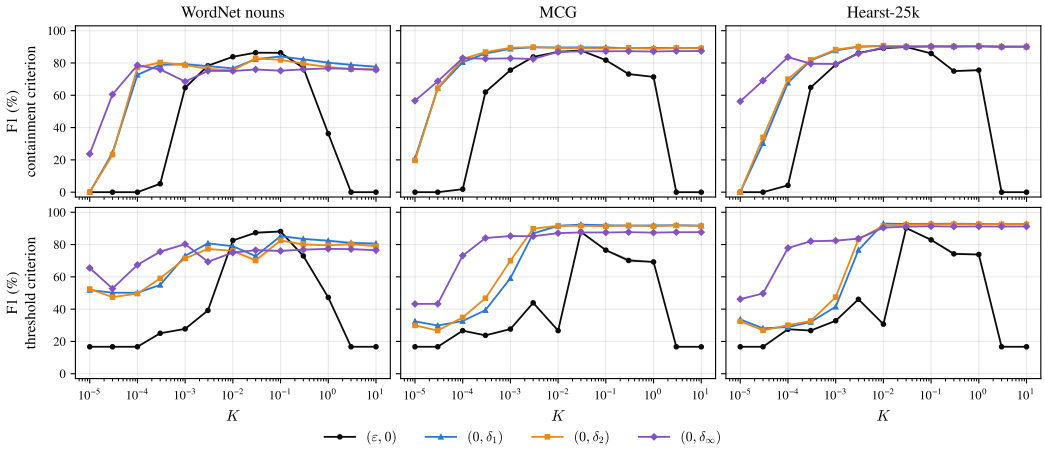

Figure 10: Sweep of K under the protocol of Shadow Cones at $d = 10$. The conventions are those of Figure 9.

The Shadow Cones protocol and the proposed method Under the protocol of Shadow Cones, the gain from replacing the aperture is smaller than under the protocol of Ganea et al., and on nouns the penumbral $(\varepsilon, 0)$ cones win. Two aspects of the Shadow Cones training protocol differ from the assumptions under which the proposed aperture is designed. First, the initialization presupposes projection to the outside of the infeasible ball. Since the original formula cannot define cones near the origin, the initial configuration must be pushed out to the sphere of the light source. This operation is unnecessary for cones without an infeasible ball, but it also acts as a regularizer that keeps embeddings away from the neighborhood of the origin where the aperture is close to $\pi/2$. When the proposed method is applied to this protocol, embeddings may remain in the region near the origin where the aperture saturates at $\pi/2$. Second, the loss is a cross-entropy contrastive loss and does not penalize configurations in which the loss is 0 for all pairs of words. In such a configuration, the difference between the losses f of positive and negative pairs vanishes and so do the gradients, so

training stalls in saturation. Indeed, on the 10% setting of Hearst the contrastive loss stalls near its lower bound $2 \log 11 \approx 4.80$, and F1 stops changing at 62. The max-margin objective used by the protocol of Ganea et al. pushes the f of negative pairs up to the margin, so this configuration has high loss and the same phenomenon does not occur. These observations suggest that the benefit of replacing the aperture in the Poincaré-ball formulation is sensitive to whether the training protocol itself is designed around the existence of an infeasible ball.

Table 7: Full version of F1 (%) under the containment criterion for the protocol of Shadow Cones (the two Poincaré-ball formulations, the two half-space formulations, penumbral with replaced apertures, and penumbral on the Lorentz model). For each method the upper row is $d = 5$ and the lower row $d = 10$; each value is the mean and standard deviation ($\pm$) over 10 seeds. Bold marks the column best within each block; green background marks the column best across all rows. Colored numbers after each value: difference from penumbral $(\varepsilon, 0)$ of the same block.

WordNet nouns	d	non-basic edges (%) 0	10	25	50	90	recon.
Protocol of Shadow Cones (incl. half-space)							
umbral (half-space)	5	45.0±1.2	91.0±1.4	93.8±0.9	94.0±1.7	95.9±1.2	98.3±0.1
	10	45.6±1.2	**93.4**±1.9	**95.7**±1.1	**96.4**±0.9	**98.0**±1.0	98.8±0.0
penumbral (half-space)	5	48.3±3.6	81.6±1.1	86.6±0.6	90.0±1.0	95.0±0.4	98.3±0.0
	10	**53.3**±2.9	79.8±0.9	85.4±0.6	89.0±1.1	93.8±0.4	**98.9**±0.0
umbral (ball)	5	34.9±0.7	78.5±0.5	79.7±0.8	79.5±0.5	85.7±0.6	89.1±0.4
	10	42.3±1.4	80.4±0.6	82.2±0.5	83.1±0.6	89.5±0.6	92.2±0.3
penumbral $(\varepsilon, 0)$	5	34.5±0.9	73.8±0.4	81.2±0.5	85.8±0.2	91.7±0.3	95.9±0.2
	10	40.5±0.7	75.5±0.6	81.4±0.3	85.7±0.4	92.4±0.2	97.5±0.1
penumbral $(0, \delta_1)$	5	34.1±1.0 -0.4	77.7±0.5 +3.9	80.1±0.4 -1.1	83.8±0.6 -2.0	89.0±0.3 -2.7	94.1±0.2 -1.8
	10	36.6±0.9 -3.9	79.6±0.5 +4.1	80.6±0.5 -0.8	83.9±0.5 -1.8	89.8±0.5 -2.6	96.2±0.2 -1.3
penumbral $(0, \delta_2)$	5	35.1±1.3 +0.6	73.5±0.5 -0.3	78.3±0.8 -2.9	82.6±0.7 -3.2	88.7±0.3 -3.0	94.4±0.2 -1.5
	10	39.1±1.0 -1.4	73.9±0.5 -1.6	78.2±0.4 -3.2	81.7±0.7 -4.0	87.8±0.6 -4.6	96.4±0.1 -1.1
penumbral $(0, \delta_\infty)$	5	31.6±0.8 -2.9	61.3±1.1 -12.5	69.5±0.9 -11.7	73.7±0.9 -12.1	79.5±0.4 -12.2	92.5±0.1 -3.4
	10	37.0±0.8 -3.5	65.5±0.3 -10.0	72.6±0.6 -8.8	76.5±0.9 -9.2	79.8±0.5 -12.6	95.0±0.0 -2.5
Lorentz model (protocol of Shadow Cones)							
penumbral $(\varepsilon, 0)$	5	34.8±0.9	73.3±0.7	**81.1**±0.2	**86.1**±0.3	91.6±0.3	95.9±0.2
	10	**40.5**±0.7	75.3±0.6	**81.1**±0.3	85.7±0.3	**92.3**±0.3	**97.5**±0.1
penumbral $(0, \delta_1)$	5	34.4±0.9 -0.4	77.6±0.5 +4.3	80.1±0.7 -1.0	83.5±0.4 -2.6	89.0±0.6 -2.6	94.0±0.2 -1.9
	10	36.3±0.8 -4.2	**79.8**±1.1 +4.5	80.7±0.7 -0.4	83.7±0.7 -2.0	89.7±0.5 -2.6	96.1±0.2 -1.4
penumbral $(0, \delta_2)$	5	34.8±0.7 +0.0	73.5±0.5 +0.2	78.5±0.4 -2.6	82.4±0.3 -3.7	88.7±0.4 -2.9	94.5±0.2 -1.4
	10	38.9±0.7 -1.6	74.0±0.6 -1.3	78.1±0.5 -3.0	82.2±0.5 -3.5	87.8±0.6 -4.5	96.4±0.1 -1.1
penumbral $(0, \delta_\infty)$	5	32.0±0.9 -2.8	61.7±1.1 -11.6	69.7±0.8 -11.4	73.8±0.7 -12.3	79.7±0.7 -11.9	92.5±0.1 -3.4
	10	37.6±0.6 -2.9	65.4±0.4 -9.9	72.4±0.6 -8.7	76.4±0.5 -9.3	80.1±0.5 -12.2	95.0±0.0 -2.5

MCG	d	non-basic edges (%) 0	10	25	50	90	recon.
Protocol of Shadow Cones (incl. half-space)							
umbral (half-space)	5	38.9±0.3	81.4±0.3	84.6±0.3	87.9±0.4	92.0±0.1	93.0±0.1
	10	**42.0**±0.2	**85.9**±0.3	87.2±0.3	91.0±0.2	94.9±0.2	96.4±0.1
penumbral (half-space)	5	36.7±0.4	80.0±0.1	84.6±0.3	88.0±0.2	92.1±0.2	93.2±0.1
	10	40.5±0.3	80.8±0.2	**87.4**±0.3	91.0±0.4	95.2±0.1	96.6±0.1
umbral (ball)	5	25.4±0.5	35.3±0.2	36.4±0.2	39.8±0.1	46.1±0.2	48.8±0.3
	10	26.7±0.4	40.4±0.1	40.9±0.2	44.1±0.1	50.2±0.1	52.8±0.2
penumbral $(\varepsilon, 0)$	5	32.0±2.7	76.2±0.3	79.9±0.3	82.6±0.1	87.3±0.2	89.7±0.2
	10	35.7±1.5	80.4±0.3	84.8±0.2	87.6±0.2	92.2±0.1	95.1±0.1
penumbral $(0, \delta_1)$	5	29.0±0.8 -3.0	75.6±0.7 -0.6	81.8±0.5 +1.9	85.6±0.2 +3.0	89.8±0.2 +2.5	91.3±0.2 +1.6
	10	34.2±0.8 -1.5	77.7±0.2 -2.7	87.3±0.4 +2.5	**91.9**±0.3 +4.3	**95.5**±0.2 +3.3	**96.7**±0.1 +1.6
penumbral $(0, \delta_2)$	5	24.4±1.2 -7.6	73.1±0.6 -3.1	80.0±0.3 +0.1	84.1±0.4 +1.5	89.2±0.3 +1.9	90.6±0.2 +0.9
	10	29.9±0.6 -5.8	76.8±0.3 -3.6	85.0±0.3 +0.2	89.5±0.2 +1.9	94.7±0.1 +2.5	96.6±0.0 +1.5
penumbral $(0, \delta_\infty)$	5	23.4±1.3 -8.6	74.0±0.3 -2.2	77.0±0.2 -2.9	80.2±0.2 -2.4	85.5±0.2 -1.8	88.3±0.2 -1.4
	10	26.4±1.0 -9.3	75.6±0.3 -4.8	83.1±0.2 -1.7	87.3±0.2 -0.3	92.5±0.2 +0.3	96.1±0.0 +1.0
Lorentz model (protocol of Shadow Cones)							
penumbral $(\varepsilon, 0)$	5	30.9±1.0	76.1±0.2	79.7±0.2	82.6±0.2	87.3±0.1	89.7±0.2
	10	34.2±1.3	**80.6**±0.3	84.8±0.2	87.6±0.2	92.2±0.1	95.1±0.0
penumbral $(0, \delta_1)$	5	28.8±1.2 -2.1	75.1±0.5 -1.0	81.9±0.6 +2.2	85.8±0.3 +3.2	90.0±0.2 +2.7	91.2±0.3 +1.5
	10	**34.4**±0.8 +0.2	77.8±0.9 -2.8	**87.3**±0.2 +2.5	**91.9**±0.2 +4.3	**95.5**±0.1 +3.3	**96.7**±0.1 +1.6
penumbral $(0, \delta_2)$	5	24.0±0.7 -6.9	72.5±1.1 -3.6	80.1±0.4 +0.4	84.0±0.4 +1.4	89.2±0.3 +1.9	90.5±0.3 +0.8
	10	29.8±0.7 -4.4	76.9±0.4 -3.7	84.7±0.2 -0.1	89.6±0.3 +2.0	94.7±0.1 +2.5	96.6±0.1 +1.5
penumbral $(0, \delta_\infty)$	5	22.4±0.9 -8.5	74.0±0.3 -2.1	77.0±0.3 -2.7	80.2±0.2 -2.4	85.4±0.1 -1.9	88.3±0.1 -1.4
	10	26.2±0.7 -8.0	75.4±0.3 -5.2	83.0±0.2 -1.8	87.2±0.2 -0.4	92.4±0.1 +0.2	96.1±0.1 +1.0

Hearst-25k	d	non-basic edges (%) 0	10	25	50	90	recon.
Protocol of Shadow Cones (incl. half-space)							
umbral (half-space)	5	34.0±0.2	89.6±0.2	90.7±0.2	91.8±0.2	94.1±0.1	95.6±0.1
	10	**34.9**±0.3	**93.0**±0.1	**94.4**±0.1	**94.7**±0.2	**96.1**±0.1	97.5±0.0
penumbral (half-space)	5	31.7±0.4	85.0±0.4	87.9±0.2	90.4±0.2	94.2±0.3	95.7±0.1
	10	32.5±0.3	84.4±0.4	89.4±0.3	92.3±0.4	**96.1**±0.1	**97.6**±0.0
umbral (ball)	5	26.4±0.9	43.7±0.1	45.1±0.2	48.0±0.3	53.6±0.3	56.9±0.3
	10	30.0±0.9	50.2±0.2	51.7±0.1	54.2±0.1	58.3±0.2	60.6±0.2
penumbral $(\varepsilon, 0)$	5	21.6±2.0	80.6±0.2	84.0±0.2	86.5±0.2	90.9±0.1	93.0±0.1
	10	21.2±3.4	83.6±0.2	87.4±0.2	89.8±0.1	94.0±0.1	96.2±0.1
penumbral $(0, \delta_1)$	5	21.5±0.2 -0.1	80.1±0.7 -0.5	84.9±0.3 +0.9	87.4±0.2 +0.9	90.8±0.3 -0.1	92.7±0.2 -0.3
	10	21.9±0.5 +0.7	83.5±0.3 -0.1	90.1±0.2 +2.7	93.0±0.1 +3.2	95.4±0.1 +1.4	97.1±0.0 +0.9
penumbral $(0, \delta_2)$	5	24.6±0.7 +3.0	77.7±0.5 -2.9	82.5±0.4 -1.5	85.7±0.3 -0.8	90.2±0.2 -0.7	92.2±0.2 -0.8
	10	26.2±0.7 +5.0	80.6±0.4 -3.0	87.4±0.3 +0.0	91.0±0.3 +1.2	95.2±0.1 +1.2	97.0±0.1 +0.8
penumbral $(0, \delta_\infty)$	5	23.5±1.2 +1.9	78.7±0.4 -1.9	82.2±0.3 -1.8	85.2±0.3 -1.3	88.7±0.1 -2.2	91.3±0.2 -1.7
	10	26.9±1.6 +5.7	78.6±0.3 -5.0	86.7±0.3 -0.7	90.0±0.2 +0.2	94.5±0.1 +0.5	97.2±0.0 +1.0
Lorentz model (protocol of Shadow Cones)							
penumbral $(\varepsilon, 0)$	5	22.3±2.2	80.7±0.4	84.0±0.2	86.5±0.3	90.8±0.1	93.0±0.1
	10	23.0±1.4	**83.7**±0.3	87.5±0.3	89.9±0.2	94.1±0.1	96.2±0.1
penumbral $(0, \delta_1)$	5	21.2±0.3 -1.1	80.4±0.4 -0.3	84.7±0.2 +0.7	87.3±0.2 +0.8	90.8±0.2 +0.0	92.7±0.1 -0.3
	10	21.7±0.4 -1.3	**83.7**±0.4 +0.0	**90.0**±0.4 +2.5	**92.9**±0.2 +3.0	**95.5**±0.1 +1.4	97.1±0.1 +0.9
penumbral $(0, \delta_2)$	5	24.5±0.6 +2.2	77.8±0.5 -2.9	82.3±0.5 -1.7	85.7±0.3 -0.8	90.1±0.2 -0.7	92.2±0.2 -0.8
	10	26.3±0.9 +3.3	80.6±0.4 -3.1	87.1±0.2 -0.4	91.2±0.5 +1.3	95.2±0.1 +1.1	97.0±0.1 +0.8
penumbral $(0, \delta_\infty)$	5	23.2±1.0 +0.9	78.7±0.3 -2.0	82.2±0.4 -1.8	85.3±0.2 -1.2	88.8±0.2 -2.0	91.3±0.1 -1.7
	10	**27.0**±1.2 +4.0	78.4±0.4 -5.3	86.7±0.2 -0.8	90.0±0.2 +0.1	94.5±0.1 +0.4	**97.2**±0.0 +1.0

Table 8: F1 (%) under the containment criterion for the protocol of Shadow Cones on the original Hearst dataset (percentages 0–10%). The rules of the table are as in Table 7.

Hearst	d	non-basic edges (%) 0	1	2	5	10
Protocol of Shadow Cones (incl. half-space)						
umbral (half-space)	5	31.1±0.2	72.4±0.3	78.1±0.4	82.3±0.4	82.5±0.9
	10	**32.9**±0.1	**75.4**±0.2	**83.5**±0.1	**88.4**±0.2	**88.5**±0.5
penumbral (half-space)	5	30.9±0.1	63.6±0.3	72.6±0.3	79.7±0.3	82.2±0.5
	10	30.1±0.1	56.3±0.3	69.3±0.2	80.6±0.3	85.1±0.3
umbral (ball)	5	26.3±0.3	39.2±0.1	36.9±0.1	37.2±0.0	39.0±0.1
	10	27.5±0.5	46.3±0.1	44.3±0.1	43.4±0.1	44.6±0.1
penumbral $(\varepsilon, 0)$	5	22.4±2.5	58.7±0.5	68.2±0.3	73.0±0.2	75.2±0.1
	10	21.3±2.9	60.4±0.4	70.8±0.2	78.3±0.2	80.8±0.1
penumbral $(0, \delta_1)$	5	20.3±0.3 -2.1	59.7±0.2 +1.0	61.5±1.1 -6.7	71.8±0.6 -1.2	76.5±0.7 +1.3
	10	20.3±0.2 -1.0	61.6±0.2 +1.2	66.4±0.1 -4.4	78.5±0.1 +0.2	84.9±0.2 +4.1
penumbral $(0, \delta_2)$	5	20.2±0.2 -2.2	59.3±0.3 +0.6	60.6±0.1 -7.6	69.5±0.6 -3.5	74.0±0.7 -1.2
	10	20.5±0.5 -0.8	60.1±0.2 -0.3	65.9±0.2 -4.9	75.5±0.4 -2.8	82.3±0.3 +1.5
penumbral $(0, \delta_\infty)$	5	20.7±0.6 -1.7	54.6±0.6 -4.1	63.8±0.4 -4.4	69.9±0.1 -3.1	72.2±0.2 -3.0
	10	21.0±0.8 -0.3	50.9±0.5 -9.5	62.3±0.4 -8.5	75.9±0.2 -2.4	80.9±0.2 +0.1
Lorentz model (protocol of Shadow Cones)						
penumbral $(\varepsilon, 0)$	5	21.7±2.5	58.7±0.8	68.1±0.4	73.1±0.2	75.2±0.1
	10	**22.1**±3.8	60.3±0.6	**70.8**±0.3	78.3±0.2	80.8±0.1
penumbral $(0, \delta_1)$	5	20.4±0.3 -1.3	59.7±0.4 +1.0	62.1±1.1 -6.0	71.8±0.6 -1.3	76.5±0.9 +1.3
	10	20.5±0.4 -1.6	**61.6**±0.2 +1.3	66.5±0.2 -4.3	**78.5**±0.4 +0.2	**85.0**±0.2 +4.2
penumbral $(0, \delta_2)$	5	20.2±0.3 -1.5	59.1±0.2 +0.4	60.7±0.1 -7.4	69.3±0.9 -3.8	73.8±0.6 -1.4
	10	20.6±0.3 -1.5	60.1±0.2 -0.2	66.0±0.2 -4.8	75.5±0.4 -2.8	82.4±0.3 +1.6
penumbral $(0, \delta_\infty)$	5	20.5±0.7 -1.2	54.7±0.6 -4.0	64.1±0.6 -4.0	70.0±0.2 -3.1	72.2±0.2 -3.0
	10	21.3±0.4 -0.8	50.7±0.3 -9.6	62.7±0.3 -8.1	75.7±0.3 -2.6	80.8±0.2 +0.0

M Additional Experimental Details and Results

Construction of Hearst-25k The distributed Hearst dataset (Le et al., 2019) has a transitive closure of 6,846,245 edges, and the settings with a non-basic edge percentage of 25% or higher, which use most of the closure for training, are computationally impractical. We therefore truncate the distributed data (the extraction result cut off by the authors at the top 50,000 rows) further at the top 25,000 rows, and rebuild it by the same procedure as the original (DAG conversion, reduction, and splits); the result is Hearst-25k. Its closure has 1,060,373 edges, comparable to MCG (1,134,348 edges), preserves the deep hierarchy characteristic of Hearst, and its fraction of nodes with two or more parents is almost the same as the original (12.3% versus 13.9%).

Splits and negative examples For nouns we use the splits distributed by Ganea et al. as they are. For MCG and Hearst, we remove the edges that form cycles from the raw hypernymy relations to obtain a DAG, and take its transitive reduction as the basic edges and its transitive closure as the full set of edges. The splits follow the procedure of Ganea et al.: 5% of the non-basic edges are drawn at random as the test set and another 5% as the validation set, and the training set is the union of all basic edges and a random sample of the remaining non-basic edges of the specified percentage. The negative examples for evaluation are generated once and fixed: for each positive pair (x, y) (x the hypernym), 5 pairs (x, y') with y' not a descendant of x in the closure and 5 pairs (x', y) with x' not

an ancestor of y in the closure, 10 in total. Negative examples for training are sampled by the same rule at every update. The splits and the evaluation negatives are fixed, and the seed changes only the initial embeddings and the sampling of training negatives.

Settings of the protocol of Ganea et al. The initialization trains Poincaré embeddings (Nickel & Kiela, 2017) (Euclidean distance embeddings for the Euclidean cone) for 100 epochs. The learning rate is 0.03, the first 20 epochs are a burn-in with the learning rate divided by 10, and 10 negatives per positive pair are drawn with probability proportional to the word frequency to the power 0.75. The resulting embeddings are scaled by 0.7 and used as the initial values of the cones. The cones are trained for 200 epochs with the max-margin loss of Ganea et al. (margin 0.01), batch size 10, 10 negatives per positive pair drawn uniformly, and Riemannian SGD. K is fixed to 0.1. The learning rate is selected by validation F1 for each pair of method and dataset from the candidates $\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}, 3\times10^{-2}\}$. The Euclidean cone uses SGD with learning rate 10^{-4}, following Ganea et al. The cones on the Lorentz model take the Poincaré embeddings of the initialization mapped isometrically to the hyperboloid as initial values, and are trained with the same settings as the Poincaré model except for the selection of the learning rate (Appendix K). Among the reference methods, the Poincaré embedding is the same 100-epoch embedding as the initialization, evaluated with the distance-based is-a score of Nickel and Kiela with a threshold selected on validation. The order embedding has no initialization and is trained for 500 epochs with margin 1.0 and learning rate 0.1. Evaluation is performed every 20 epochs, and for each of the containment and threshold criteria the model at the epoch with the best validation F1 is used for the test evaluation. Under the containment criterion, K at evaluation time is selected by validation F1 from the candidates $\{0.01, 0.05, 0.1, 0.2, 0.3\}$ together with the trained value. For the $(\varepsilon, 0)$ cones, however, candidates whose radius ε of the infeasible region exceeds the minimum norm of the embeddings are excluded. The dimension is $d \in \{5, 10\}$ and each condition uses 10 seeds.

Conditions under which the proposed method wins We discuss how the relative merits of the $(\varepsilon, 0)$ and $(0, \delta_2)$ cones in Table 2 split by the structure of the data. The $(\varepsilon, 0)$ cones beat the $(0, \delta_2)$ cones only on the high-percentage settings of nouns (90% and reconstruction); on MCG and Hearst-25k the $(0, \delta_2)$ cones win in all settings, with large margins. This distribution is consistent with two properties of the data. The first is the abundance of nodes with two or more parents. Their fraction is 2.6% for nouns, 12.3% for Hearst-25k, and 26.8% for MCG. A node with several parents must lie in the intersection of the cones of all its ancestors, and realizing this with nested cones satisfying exact transitivity is difficult. Conversely, an order in which almost all nodes have at most one parent, as in nouns, can be realized directly by nested cones. Indeed, the two cells where the $(\varepsilon, 0)$ cones beat the $(0, \delta_2)$ cones are both high-percentage settings of nouns; when a realizable nesting is given together with most of the transitive closure, the freedom to tolerate violations offers no advantage. For the $(\varepsilon, 0)$ cones, the aperture thins exponentially away from the origin as $\sin\psi_K = K/h_{\max}$, and all points are confined outside ε, so realizing these intersections becomes harder as the number of entailments to be satisfied simultaneously grows. Indeed, on the Hearst datasets the containment-criterion F1 of the $(\varepsilon, 0)$ cones is not monotone in the percentage (Table 6), whereas the threshold criterion of the same models improves monotonically (Table 5). The ordering of the values of the loss f is being learned; only the configurations where f is exactly 0 fail to be realized. The second property is the depth of the hierarchy. The longest path of the basic edges is 18 for nouns, 31 for MCG, and 46 for Hearst-25k, and with the thin $(\varepsilon, 0)$ cones, longer chains fall out of transitivity under smaller displacements. These properties also appear in the sensitivity to the post-hoc selection of K. On nouns, the K selected on the validation set coincides with the training value for all seeds and F1 changes steeply around it, whereas on the Hearst datasets and MCG, F1 is flat over the wide range $K \geq 0.1$. In the 0% setting, where only the basic edges are given, this difficulty is most pronounced: the $(\varepsilon, 0)$ cones almost stop working, with 5.5 on MCG and 1.2 on Hearst-25k, while the $(0, \delta_2)$ cones maintain 22.2 and 16.4 (Table 4). These results suggest that the relative advantage of the relaxation to (ε, δ)-transitivity is larger when the hierarchy contains more nodes with two or more parents, is deeper, or is observed more sparsely, whereas the original formula remains competitive when the order is close to a tree and most of its transitive closure is observed. The difference between the $(0, \delta_1)$ and $(0, \delta_2)$ cones can be interpreted in the same framework. Their relative merits split by dataset and training protocol: under the protocol of Ganea et al. the $(0, \delta_2)$ cones win on nouns in every setting and on Hearst-25k at 25% or higher for $d = 10$, whereas the $(0, \delta_1)$ cones win on MCG and on the original Hearst at 10% or higher (Tables 4 and 6); under the protocol of Shadow Cones the $(0, \delta_1)$ cones win on all three datasets except at 0% (Appendix L). The

differences are mostly within 3 points and exceed 5 points only at 0% on MCG under the protocol of Ganea et al. and at 10% on nouns under the protocol of Shadow Cones. By (8), $q = 1$ has a looser budget than $q = 2$ in the region far from the origin, at the price of a more conservative aperture at intermediate geodesic distances, and which is advantageous depends on the form of the loss.

Transitivity violations of the learned cones We measure how often the relation predicted by the learned cones violates transitivity and where the violations occur. For the models of Table 4 at 50% non-basic edges and $d = 5$, we form the predicted relation $A = \{(x, y) : f(x, y) = 0\}$ over all ordered pairs of nodes, where $f(x, y) = \max(0, \Xi(x, y) - \psi(\|x\|))$ is the containment criterion, without reference to the ground truth. A 2-path $(x, y), (y, z) \in A$ is a violation if $(x, z) \notin A$, and $f(x, z)$ is its violation angle. Table 9 reports the number of predicted edges and 2-paths, the share of 2-paths whose start x lies inside the infeasible region of the $(\varepsilon, 0)$ cones ($\|x\| < \varepsilon$, $\varepsilon = 0.0990$ for $K = 0.1$), the violation rate among the 2-paths, overall and split by $\|x\| \geq \varepsilon$ and $\|x\| < \varepsilon$, the median and maximum of f, and, for $\|x\| \geq \varepsilon$, the maximum ratio of f to the bound $\psi_K - \psi$ of Proposition C.1. Figure 11 shows the same quantities as functions of $\|x\|$. Embeddings on the Lorentz model are mapped to the ball by $p = x_s/(1 + x_0)$ and evaluated with the same formulas. Since K is selected on the balanced evaluation with sampled negatives, the predicted relation over all pairs is denser than the transitive closure (for the $(0, \delta_2)$ cones, 4 to 7 times its size), and the 2-paths are dominated by the top-level concepts, which the proposed cones place inside ε (Figure 12).

The $(\varepsilon, 0)$ cones violate transitivity in none of the checkpoints, as Theorem 3.1 guarantees for their embeddings, which all lie outside ε. The $(0, \delta_2)$ cones violate in 0.57% (nouns), 5.1% (MCG), and 8.5% (Hearst-25k) of the 2-paths, and the $(0, \delta_1)$ cones in 0.96%, 3.9%, and 6.5%. Almost all violations start inside ε: outside ε the rate is at most 0.33% for the $(0, \delta_2)$ cones and 1.1% for the $(0, \delta_1)$ cones, and the ratio $f/(\psi_K - \psi)$ stays below 0.38, so the bound of Proposition C.1 holds with margin. The median violation angle is 2.5° to 5° and the maximum 50° for the $(0, \delta_2)$ cones. Inside ε only the trivial bound $\pi - \psi$ applies. Restricted to the 2-paths formed by two edges of the transitive closure, the rates of the $(0, \delta_2)$ cones drop to 0.26%, 0.96%, and 3.2%. The $(0, \delta_\infty)$ cones, whose aperture is the constant $\pi/2$ inside ε, violate in 28% (MCG) and 29% (Hearst-25k) of their 2-paths, whereas the smooth saturation of $q = 1, 2$ reduces this to at most 8.5%. Smooth saturation thus matters for the consistency of the predictions as well as for F1. The Lorentz model behaves the same way.

Table 9: Transitivity violations of the relation predicted by the learned cones over all ordered pairs (50% non-basic edges, $d = 5$; mean and standard deviation ($\pm$) over seeds, except the maxima, which are over all seeds). Edges and 2-paths: their numbers in the predicted relation (M: millions); inside/outside: 2-paths whose start x lies inside ($\|x\| < \varepsilon$) or outside the infeasible region of the $(\varepsilon, 0)$ cones; the inside column is the share of such 2-paths.

		edges	2-paths	inside (%)	violation rate (%)			$f(x, z)$ (deg)		$f/(\psi_K - \psi)$
					all	outside	inside	median	max	max, outside
Poincaré model										
WordNet nouns	$(\varepsilon, 0)$	1.89M	12.1M	0.0±0.0	0.00±0.00	0.000±0.000	0.00±0.00	0.00±0.00	0.0	0.00
	$(0, \delta_1)$	4.46M	81.6M	75.0±2.7	0.96±0.21	0.238±0.034	1.20±0.29	1.82±0.13	23.8	0.27
	$(0, \delta_2)$	4.90M	89.6M	61.3±2.1	0.57±0.13	0.061±0.041	0.89±0.17	2.54±0.16	32.4	0.38
	$(0, \delta_\infty)$	3.02M	29.2M	27.0±3.7	0.17±0.11	0.000±0.000	0.61±0.33	4.75±2.11	55.4	0.00
MCG	$(\varepsilon, 0)$	1.51M	8.1M	0.0±0.0	0.00±0.00	0.000±0.000	0.00±0.00	0.00±0.00	0.0	0.00
	$(0, \delta_1)$	7.25M	254.7M	96.5±1.0	3.90±0.25	0.630±0.057	4.02±0.24	2.31±0.13	38.3	0.26
	$(0, \delta_2)$	8.20M	294.7M	98.4±0.3	5.13±0.43	0.332±0.033	5.21±0.42	3.54±0.14	38.5	0.37
	$(0, \delta_\infty)$	2.57M	148.6M	99.3±0.2	27.59±2.02	0.000±0.000	27.78±2.03	10.25±0.95	66.4	0.00
Hearst-25k	$(\varepsilon, 0)$	1.37M	6.5M	0.0±0.0	0.00±0.00	0.000±0.000	0.00±0.00	0.00±0.00	0.0	0.00
	$(0, \delta_1)$	3.54M	133.5M	99.3±0.4	6.50±0.56	1.057±0.236	6.54±0.55	3.33±0.32	42.0	0.24
	$(0, \delta_2)$	3.96M	182.0M	99.6±0.2	8.54±0.82	0.281±0.054	8.57±0.81	4.99±0.50	50.0	0.30
	$(0, \delta_\infty)$	2.18M	119.6M	99.4±0.1	28.74±1.32	0.000±0.000	28.92±1.32	9.10±0.87	68.3	0.00
Lorentz model										
MCG	$(\varepsilon, 0)$	1.46M	7.7M	0.0±0.0	0.00±0.00	0.000±0.000	0.00±0.00	0.00±0.00	0.0	0.00
	$(0, \delta_1)$	7.43M	266.4M	95.9±0.7	3.69±0.20	0.672±0.061	3.82±0.20	2.21±0.08	21.5	0.29
	$(0, \delta_2)$	8.46M	310.8M	98.3±0.3	5.07±0.40	0.339±0.052	5.15±0.40	3.52±0.15	46.3	0.37
	$(0, \delta_\infty)$	2.64M	154.0M	99.3±0.2	27.72±2.28	0.000±0.000	27.91±2.30	10.37±1.12	68.9	0.00
Hearst-25k	$(\varepsilon, 0)$	1.35M	6.3M	0.0±0.0	0.00±0.00	0.000±0.000	0.00±0.00	0.00±0.00	0.0	0.00
	$(0, \delta_1)$	3.53M	139.3M	99.3±0.3	6.25±0.73	1.118±0.252	6.28±0.73	3.30±0.37	39.6	0.28
	$(0, \delta_2)$	3.96M	186.6M	99.6±0.1	8.30±0.89	0.310±0.076	8.33±0.89	5.08±0.55	61.3	0.32
	$(0, \delta_\infty)$	2.18M	120.2M	99.3±0.1	28.69±1.22	0.000±0.000	28.88±1.21	9.13±0.79	69.3	0.00

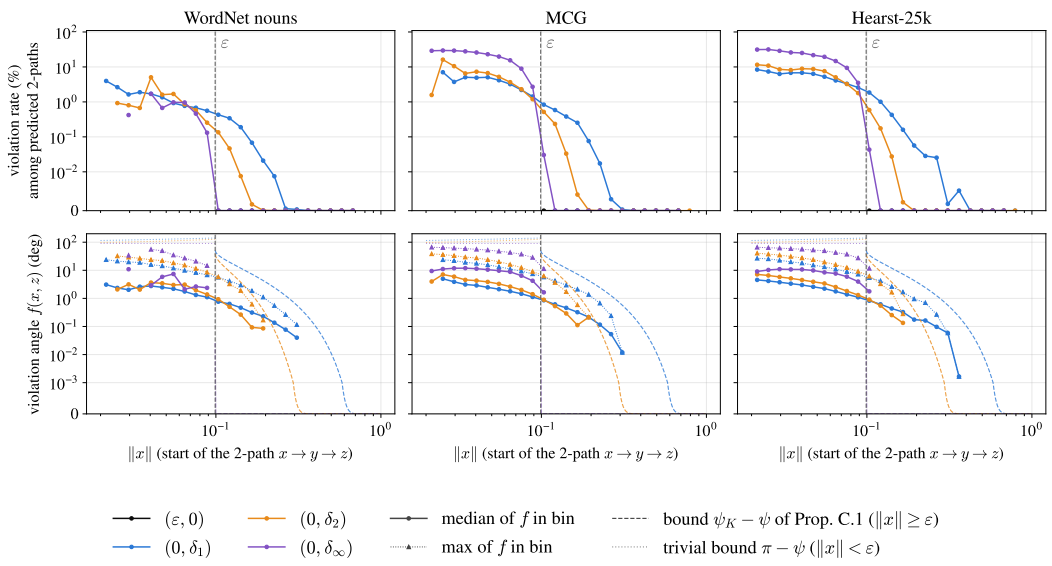

Figure 11: Violation rate (top) and violation angle $f(x, z)$ (bottom) of the predicted relation as functions of $\|x\|$ (Poincaré model, seeds pooled, logarithmic bins). The dashed vertical line is ε. The dashed curves are the bound $\psi_K - \psi$ of Proposition C.1 for $\|x\| \geq \varepsilon$, and the dotted curves the trivial bound $\pi - \psi$ for $\|x\| < \varepsilon$.

Qualitative evaluation (norm distribution of top-level concepts) Figure 12 shows the empirical cumulative distribution of the norms of the nodes of depth at most 1 (the roots and their children), for embeddings trained at 50% non-basic edges and $d = 5$ as in Table 2 (a single seed, the best checkpoint under the containment criterion). With the $(\varepsilon, 0)$ cones, no embedding norm can fall below ε, so top-level concepts either stick to the boundary of the infeasible region or are pushed further outward. In contrast, the $(0, \delta_1)$ and $(0, \delta_2)$ cones place a portion of the top-level concepts inside ε, that is, inside the infeasible region of the $(\varepsilon, 0)$ cones, confirming that the full-domain feasibility shown in §5 is actually used during training.

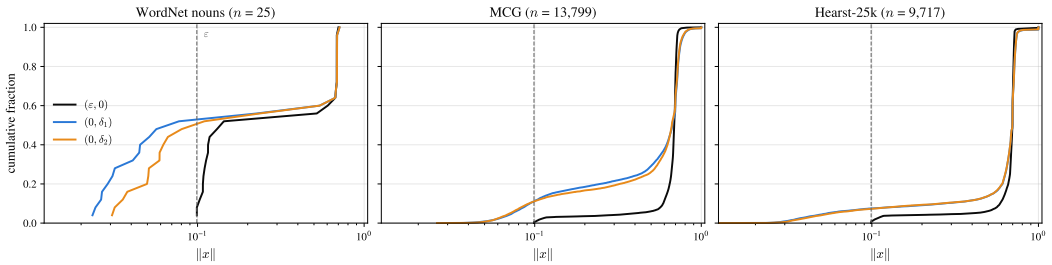

Figure 12: Empirical cumulative distribution of the norms of top-level concepts (nodes of depth at most 1). Embeddings are from a single seed trained at 50% non-basic edges ($d = 5$, $K = 0.1$) as in Table 2. The dashed line marks the boundary ε of the infeasible region.

Qualitative evaluation (visualization of two-dimensional embeddings) Figure 13 shows embeddings of WordNet mammals for all four cones (the main text, Figure 4, shows the $(\varepsilon, 0)$ and $(0, \delta_2)$ cones), and Figure 14 those of a balanced tree of depth 6 and branching factor 4, each trained directly in two dimensions. This experiment follows the visualization setting of Ganea et al. ($d = 2$, 150 epochs of Poincaré initialization, 150 epochs of cone training, $K = 0.1$, a single seed), whose training conditions, starting with the dimension, are entirely different from the accuracy comparison of Table 2 ($d \in \{5, 10\}$, the setting of §6.1). These figures are therefore a qualitative evaluation of the layout of the embeddings, and should not be used to compare the accuracy of the methods. In

both figures the $(0, \delta_1)$ cones behave like the $(0, \delta_2)$ cones and move the shallowest nodes inside the boundary ε, whereas the $(0, \delta_\infty)$ cones lie between the $(\varepsilon, 0)$ cones and the two: they move only a few nodes inside and keep the others on the boundary (on the balanced tree their innermost nodes have norms 0.095 and 0.097 against $\varepsilon = 0.099$).

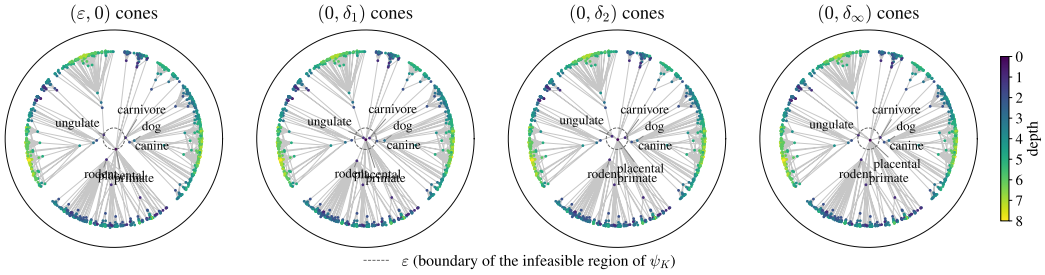

Figure 13: Full version of Figure 4: embeddings of WordNet mammals trained directly in two dimensions with all four formulas. Node colors indicate the depth from the root, gray lines the basic edges, and the dashed circle the boundary ε of the infeasible region of the $(\varepsilon, 0)$ cones.

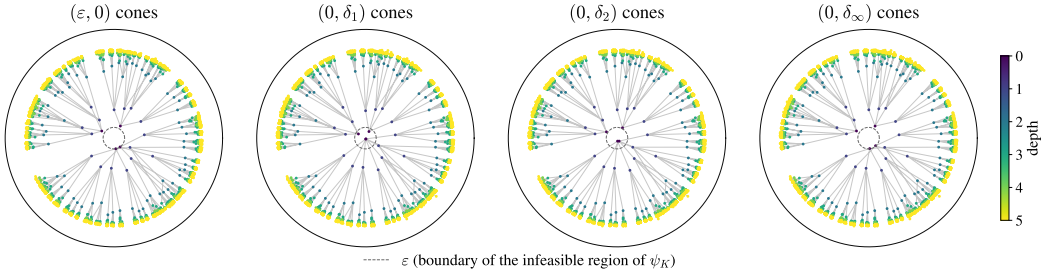

Figure 14: The same visualization as Figure 13 for a balanced tree of depth 6 and branching factor 4.

Settings of MERU The training data is GRIT (Peng et al., 2024) preprocessed with the procedure provided in the public implementation of HyCoCLIP (Pal et al., 2025), used without further changes, except that 3 tar files (19,215 pairs) are held out for validation and excluded from training. HyCoCLIP recommends GRIT over RedCaps (Desai et al., 2021) for pretraining that exploits hierarchical pairs, because grounded regions in RedCaps are poorly localized within the image and few pairs remain after noise filtering. The training of MERU does not use the grounded regions; only pairs of the whole image and its caption are used. The model uses a ViT-B/16 image encoder and the same text encoder as the original work, and we keep the original settings unchanged, namely batch size 2048, 120,000 iterations, AdamW (learning rate 5×10^{-4}, weight decay 0.2), a cosine schedule with 4000 warmup iterations, fp16 mixed precision, entailment-loss weight 0.2, and learnable curvature and scale; only the aperture formula of the entailment loss is replaced by that of the $(\varepsilon, 0)$, $(0, \delta_1)$, or $(0, \delta_2)$ cones. We use 3 seeds (seeds 1, 2, and 3 in the main text), which change the initial values and the order of the data. A checkpoint is saved every 5000 iterations, and the curves in Figures 5 and 15 are the entailment loss of each checkpoint recomputed with the original aperture on all validation pairs. Accuracies are evaluated on the final checkpoint following the protocol of HyCoCLIP (including its set of prompts): zero-shot classification on 16 datasets (ImageNet (Deng et al., 2009), CIFAR-10 and CIFAR-100 (Krizhevsky, 2009), SUN397 (Xiao et al., 2010), Caltech-101 (Fei-Fei et al., 2004), STL-10 (Coates et al., 2011), Food-101 (Bossard et al., 2014), CUB (Wah et al., 2011), Cars (Krause et al., 2013), Aircraft (Maji et al., 2013), Pets (Parkhi et al., 2012), Flowers-102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), Country211 (Radford et al., 2021)), image-text retrieval Recall@5 and Recall@10 on COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), and the hierarchy metrics on ImageNet.

Definition of the hierarchy metrics The hierarchy metrics on ImageNet follow the evaluation protocol of HyCoCLIP (Pal et al., 2025). We use the directed graph (1,372 nodes) obtained by restricting the WordNet hypernymy relation to the 1,000 ImageNet classes and their ancestors, and write $A(c)$ for the set of ancestor nodes of class c (excluding c itself). For each validation image, the predicted class $\hat{y}$ is obtained with the same classifier as in zero-shot classification, the following five quantities are computed for the pair $(\hat{y}, y)$ with the true class y, and they are averaged over all images. TIE (tree induced error) (Dekel et al., 2004) is the length of the shortest path between $\hat{y}$ and y when the direction of the graph is ignored (edge weight 1). The LCA (lowest common ancestor) error is the distance from $\hat{y}$ to the deepest common ancestor, computed as $|A(\hat{y}) \setminus A(y)| + 1$. The Jaccard similarity, hierarchical precision, and hierarchical recall (Kosmopoulos et al., 2015) are set-based quantities on the ancestor sets,

$$J = \frac{|A(\hat{y}) \cap A(y)|}{|A(\hat{y}) \cup A(y)|}, \qquad P_H = \frac{|A(\hat{y}) \cap A(y)|}{|A(\hat{y})|}, \qquad R_H = \frac{|A(\hat{y}) \cap A(y)|}{|A(y)|}.$$

Lower TIE and LCA and higher values of the other three indicate that misclassifications are closer to the true class in the hierarchy.

Complete results for hyperbolic vision-language models We present the results complementing Table 3 and Figure 5 of the main text. Figure 15 shows the validation entailment loss of the models pretrained as the three cones ($(\varepsilon, 0)$, $(0, \delta_1)$, and $(0, \delta_2)$ cones), and Tables 10 and 11 list, for each seed, the zero-shot classification accuracy, the image-text retrieval Recall, the ImageNet hierarchy metrics, and the mean normalized radius $\bar{s}$ of the text and image embeddings of the same three formulas. Bold marks the best value in each column within each seed. The classification and retrieval accuracies are comparable for all formulas, with no consistent trend across seeds.

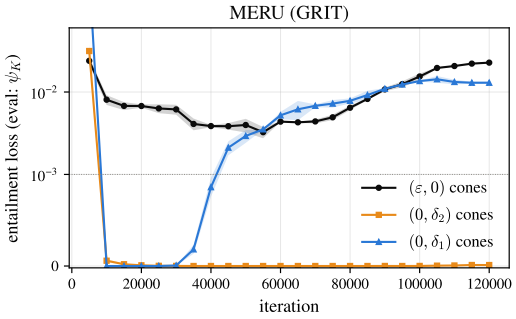

Figure 15: Validation entailment loss of MERU (ViT-B/16) pretrained on GRIT (full version): means over three seeds with standard-deviation bands. Each checkpoint of the three models pretrained as the $(\varepsilon, 0)$, $(0, \delta_1)$, and $(0, \delta_2)$ cones is evaluated with the loss of the original formula ψ_K. The vertical axis is a logarithmic scale, linear below 10^{-3}. The $(0, \delta_1)$ cones start to increase partway and merge with the curve of the $(\varepsilon, 0)$ cones.

Table 10: Zero-shot classification accuracy (%) of MERU (ViT-B/16) pretrained on GRIT, per seed and aperture formula. Bold marks the best within each seed.

		ImageNet	Food-101	CIFAR-10	CIFAR-100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	Country211	Average
seed 1	$(\varepsilon, 0)$	38.0	53.9	**76.5**	43.6	13.2	49.2	9.1	1.8	18.8	**51.3**	75.5	**25.8**	**92.8**	33.1	**40.4**	**5.1**	39.3
	$(0, \delta_1)$	38.9	**54.4**	74.2	44.2	**16.4**	**50.3**	8.2	2.8	20.9	48.5	75.7	23.2	92.0	**41.2**	38.9	4.9	**39.7**
	$(0, \delta_2)$	**39.1**	51.4	71.9	**46.1**	16.1	49.9	**9.6**	**3.2**	**22.8**	50.8	**76.9**	20.7	**92.8**	34.1	39.9	4.8	39.4
seed 2	$(\varepsilon, 0)$	38.1	51.5	71.3	44.2	13.5	48.7	8.5	3.1	21.3	48.2	**77.8**	23.3	92.4	**45.2**	42.0	4.8	39.6
	$(0, \delta_1)$	**39.1**	**55.5**	71.0	44.6	**16.8**	49.9	**9.4**	**3.6**	**22.5**	**48.5**	75.7	**23.6**	**93.2**	43.3	39.7	5.0	**40.1**
	$(0, \delta_2)$	38.8	52.2	**73.2**	**46.5**	16.3	**50.4**	9.1	3.4	20.7	48.2	77.5	23.4	92.4	37.3	**42.1**	**5.3**	39.8
seed 3	$(\varepsilon, 0)$	38.4	50.9	**77.8**	44.5	14.0	50.0	9.0	**3.6**	21.9	46.5	74.8	**22.2**	92.0	33.2	**42.8**	4.7	39.1
	$(0, \delta_1)$	**39.2**	**53.8**	76.7	**46.7**	14.8	**50.2**	**9.4**	2.8	19.6	44.6	**76.8**	21.7	**92.1**	**40.9**	41.4	**5.1**	**39.7**
	$(0, \delta_2)$	38.6	51.7	72.1	43.1	**15.4**	49.4	**9.4**	3.0	**22.7**	**47.2**	75.6	21.7	91.8	31.4	42.4	4.9	38.8

Table 11: Image-text retrieval Recall (%), ImageNet hierarchy metrics, and mean normalized radius $\bar{s}$ of the validation text and image embeddings of MERU (ViT-B/16) pretrained on GRIT, per seed and aperture formula. Bold marks the best within each seed (not applied to $\bar{s}$).

		COCO				Flickr30k				ImageNet hierarchy					$\bar{s}$	
		T→I@5	I→T@5	T→I@10	I→T@10	T→I@5	I→T@5	T→I@10	I→T@10	TIE↓	LCA↓	Jaccard↑	Hier. prec.↑	Hier. rec.↑	text	image
seed 1	$(\varepsilon, 0)$	55.3	**69.9**	66.8	79.7	**82.3**	90.7	89.0	95.3	3.877	2.339	0.768	0.840	0.841	0.97	1.99
	$(0, \delta_1)$	55.6	68.6	66.7	79.4	81.9	**91.5**	88.5	**95.7**	**3.760**	2.283	**0.775**	**0.846**	**0.847**	1.14	4.99
	$(0, \delta_2)$	**55.8**	**69.9**	**67.1**	**79.8**	82.1	91.3	**89.5**	95.3	3.762	**2.282**	**0.775**	0.845	0.846	0.85	3.27
seed 2	$(\varepsilon, 0)$	**55.5**	**70.1**	**67.2**	**80.1**	**82.2**	90.7	88.3	**95.5**	3.870	2.327	0.768	0.840	0.840	0.97	1.99
	$(0, \delta_1)$	55.3	68.2	66.3	78.3	81.5	90.7	**88.6**	95.3	**3.740**	**2.284**	**0.777**	**0.846**	**0.849**	1.15	4.98
	$(0, \delta_2)$	**55.5**	68.9	67.0	79.5	81.6	**91.3**	**88.6**	95.3	3.775	2.299	0.774	0.845	0.846	0.85	3.26
seed 3	$(\varepsilon, 0)$	**55.9**	69.2	67.1	79.2	82.2	91.2	88.6	95.4	3.829	2.300	0.771	0.843	0.842	0.97	1.99
	$(0, \delta_1)$	55.8	**70.2**	67.2	**79.9**	82.3	**91.9**	**88.9**	**95.9**	**3.752**	**2.281**	**0.776**	**0.847**	**0.847**	1.14	4.96
	$(0, \delta_2)$	55.7	69.0	**67.4**	79.3	**82.4**	90.7	**88.9**	95.1	3.809	2.329	0.774	0.843	**0.847**	0.85	3.25

Computing environment All experiments on hierarchical word embeddings were run on CPUs, distributed over several machines with one core per job. The training and evaluation of MERU were run on a machine with four NVIDIA RTX 6000 Ada GPUs, using one GPU per training run.