THE BROKEN GEOMETRY OF HYPERBOLIC VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyperbolic vision-language models (VLMs) leverage hyperbolic geometry to model hierarchical structure in multimodal representations. Recent works contrastively align images and text in a shared hyperbolic space while employing entailment cones to impose hierarchical orderings among whole and part representations. However, we find that existing hyperbolic VLMs operate in a near-Euclidean regime, with embeddings concentrated near the origin on manifolds with small curvature. Thus, most embeddings fall outside the theoretically valid domain of entailment cones, suggesting that the intended geometry and entailment mechanism are not fully exploited in practice. Motivated by our observation, we propose a simple Euclidean-based hierarchy-aware VLM that retains the same compositional supervision while eliminating hyperbolic operations and entailment cones. Despite its simpler geometry, our model consistently outperforms existing hyperbolic VLMs across diverse downstream tasks. The empirical results suggest that the gains reported by prior studies may stem primarily from additional compositional data and hierarchical supervision rather than from hyperbolic geometry itself, highlighting the need to carefully examine the contribution of hyperbolic geometry in hierarchical representation learning.

1 INTRODUCTION

Vision-language models (VLMs) aim to align visual and linguistic information within a shared representation space (Radford et al., 2021; Desai et al., 2023), yet real-world concepts are not organized in a flat structure. Images can be viewed as compositions of multiple contextual regions, while language expresses concepts at varying levels of semantic abstraction (Zeng et al., 2022; Pal et al., 2025; Kim et al., 2026). For example, "a dog" and "a dog playing with a ball" can both describe the same image, but the latter conveys a more specific scene by combining concepts grounded in the dog and ball regions. Accordingly, modeling hierarchical structures is important for VLMs to understand not only whether images and text correspond semantically, but also how their underlying concepts are organized (Desai et al., 2023; Novack et al., 2023).

Hyperbolic geometry has emerged as a natural foundation for representing hierarchical structures (Nickel & Kiela, 2017; Ganea et al., 2018a; Sala et al., 2018). Its negative curvature gives rise to exponential volume growth with radius, closely matching the branching pattern of tree-structured data (Ganea et al., 2018b; Desai et al., 2023). Consequently, deep and highly branching hierarchies can be represented with low distortion, while the radial positions of embeddings can encode relative levels of abstraction (Sala et al., 2018; Desai et al., 2023). These advantages have motivated its adoption in representation learning across graphs, language, and vision (Chami et al., 2019; Tifrea et al., 2019; Khrulkov et al., 2020). In particular, recent VLMs (Desai et al., 2023; Pal et al., 2025; Yoshikawa & Matsubara, 2026; Kim et al., 2026) leverage hyperbolic geometry to model image-text alignment and whole-part relations.

However, do hyperbolic VLMs actually exploit hyperbolic geometry? Merely placing representations in the hyperbolic space does not guarantee that the learned embeddings make meaningful use of its distinctive geometric properties. Recent hyperbolic VLMs further combine hyperbolic representations with explicit hierarchical supervision (Pal et al., 2025; Yoshikawa & Matsubara, 2026; Kim et al., 2026), making this question particularly important. Their downstream gains cannot reveal whether the improvement comes from the geometry, the supervision, or their interaction. A

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.

proper assessment must determine whether the learned representations exhibit the intended hyperbolic structure and whether downstream gains actually depend on it.

We address the question through geometric analysis of existing hyperbolic VLMs, examining both their learned geometry and the entailment cones used to encode general-to-specific relations. Across representative VLMs, the learned curvature rapidly collapses to its lower bound early in training, while the embeddings remain concentrated near the hyperbolic origin. In this near-Euclidean regime, we further find that many embeddings used as cone apices fall below the minimum admissible radius, resulting in saturated cone apertures under clipping. Overall, these findings reveal a gap between adopting a hyperbolic parameterization and realizing its intended geometric properties.

Motivated by our geometric analysis, we introduce *EuHier*, a simple Euclidean hierarchy-aware VLM that retains whole–part compositional supervision while eliminating hyperbolic parameterization and entailment cones. To examine whether such hierarchical signals remain effective without hyperbolic geometry, EuHier encodes the corresponding relations through contrastive positive pairs across whole and part representations rather than through geometric ordering constraints. Despite its simpler formulation, EuHier consistently outperforms existing hyperbolic VLMs across diverse downstream tasks, including zero-shot image classification, cross-modal retrieval, and hierarchical classification. Together, our results suggest that the gains reported by recent hyperbolic VLMs are driven largely by hierarchy-aware compositional supervision rather than by the use of hyperbolic geometry itself, underscoring the importance of disentangling geometric effects from additional supervision when evaluating hierarchical representation learning.

2 Background

2.1 Related Works

Conventional VLMs such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and ALBEF (Li et al., 2021) learn aligned visual and textual representations in a shared representation space. However, Euclidean-based VLMs are not explicitly designed to capture the hierarchical structure underlying visual and linguistic concepts. Hyperbolic VLMs have subsequently emerged to learn hierarchy-aware multimodal representations, leveraging hyperbolic geometry whose exponential volume growth naturally accommodates tree-like structures (Desai et al., 2023). Existing methods generally combine contrastive learning for visual–textual alignment with entailment learning (Ganea et al., 2018a) intended to impose asymmetric hierarchical ordering. Such hierarchies are typically encoded radially, with more general concepts placed closer to the origin and more specific concepts farther away. As an early hyperbolic VLM, MERU (Desai et al., 2023) combines Lorentzian contrastive learning with text-to-image entailment. HyCoCLIP (Pal et al., 2025) adds cropped text/image supervision to model both inter-modal and intra-modal hierarchies. PHyCLIP (Yoshikawa & Matsubara, 2026) uses multiple hyperbolic factors to capture taxonomic and compositional structures separately. UNCHA (Kim et al., 2026) models crop representativeness as radius-based uncertainty and uses it to modulate the loss temperature. However, prior works have not clearly distinguished between merely embedding representations in hyperbolic space and meaningfully exploiting its negative curvature. It remains unclear whether the gains attributed to hyperbolic geometry can be disentangled from additional parent-child supervision generated externally to the model. Hence, we quantitatively revisit existing hyperbolic VLMs by analyzing the geometric properties of the learned representations.

2.2 Preliminaries

Hyperbolic Geometry. We consider the d-dimensional Lorentz model (Nickel & Kiela, 2018) with constant negative curvature $-\kappa$, where $\kappa > 0$. Let $\mathbb{R}^{d+1}$ be equipped with the Lorentzian inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}} = -x_0 y_0 + \langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle = -x_0 y_0 + \sum_{i=1}^{d} x_i y_i, \tag{1}$$

where x_0 is the time coordinate and $\tilde{\mathbf{x}} \in \mathbb{R}^d$ denotes the spatial coordinates. The Lorentz model is defined as the upper sheet of a two-sheeted hyperboloid

$$\mathbb{H}^d_\kappa = \left\{ \mathbf{x} \in \mathbb{R}^{d+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbb{H}} = -\frac{1}{\kappa},\ x_0 > 0 \right\}, \tag{2}$$

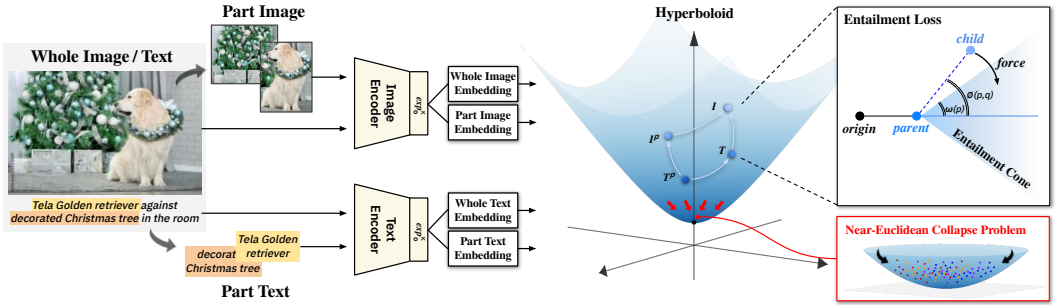


Figure 1: **Overview of recent hyperbolic VLMs and the entailment loss.** The image–text pair (I, T) and their corresponding part-level pair (I^P, T^P) are encoded by the image and text encoders, respectively. Each embedding is projected onto the hyperboloid, where cross-modal alignment is learned through a contrastive objective. The entailment loss further pulls each child embedding toward its parent's entailment cone, encouraging hierarchical transitivity in hyperbolic space. However, existing hyperbolic VLMs operate in a near-Euclidean regime, with embeddings concentrated near the origin on manifolds with small curvature.

where the origin of $\mathbb{H}^d_\kappa$ is $\mathbf{o} = (1/\sqrt{\kappa}, 0, \cdots, 0)$. For two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^d_\kappa$, the geodesic distance is

$$d_{\mathbb{H}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\kappa}} \operatorname{arcosh}(-\kappa \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{H}}) \,. \tag{3}$$

The tangent space at $\mathbf{x} \in \mathbb{H}^d_\kappa$ is defined as $T_{\mathbf{x}}\mathbb{H}^d_\kappa = \{\mathbf{v} \in \mathbb{R}^{d+1} : \langle \mathbf{x}, \mathbf{v} \rangle_{\mathbb{H}} = 0\}$. Although the Lorentzian inner product is indefinite in the ambient space, its restriction to $T_{\mathbf{x}}\mathbb{H}^d_\kappa$ is positive definite. Therefore, a tangent vector $\mathbf{v} \in T_{\mathbf{x}}\mathbb{H}^d_\kappa$ has the Riemannian norm $\|\mathbf{v}\|_{\mathbb{H}} := \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbb{H}}}$. For $\mathbf{v} \neq \mathbf{0}$, the exponential map sends a tangent vector onto the manifold as

$$\exp^\kappa_{\mathbf{x}}(\mathbf{v}) = \cosh\left(\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{H}}\right) \mathbf{x} + \frac{\sinh(\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{H}})}{\sqrt{\kappa}\|\mathbf{v}\|_{\mathbb{H}}} \mathbf{v}. \tag{4}$$

Conversely, the logarithmic map $\log^\kappa_{\mathbf{x}} : \mathbb{H}^d_\kappa \to T_{\mathbf{x}}\mathbb{H}^d_\kappa$ is the inverse of the exponential map. Further mathematical preliminaries are provided in Appendix B.

Hyperbolic Entailment Cones. As shown in Fig. 1, entailment cones (Ganea et al., 2018a; Pal et al., 2025) model asymmetric hierarchical relations by associating each general concept $\mathbf{p}$ with a conical region containing its more specific descendants $\mathbf{q}$. In the Lorentz model, the half-aperture of the cone is defined as

$$\omega(\mathbf{p}) = \arcsin\left(\frac{2K}{\sqrt{\kappa}\|\tilde{\mathbf{p}}\|}\right), \tag{5}$$

where K controls the aperture and $\tilde{\mathbf{p}}$ denotes the spatial component of the Lorentz embedding. Let $\phi(\mathbf{p}, \mathbf{q})$ denote the exterior angle of a child $\mathbf{q}$ with respect to the cone rooted at $\mathbf{p}$. The entailment objective penalizes children outside the parent cone:

$$\mathcal{L}_{\text{ent}}(\mathbf{p}, \mathbf{q}) = \max\left(0, \phi(\mathbf{p}, \mathbf{q}) - \omega(\mathbf{p})\right). \tag{6}$$

Embeddings with smaller radial coordinates therefore induce wider entailment regions, while more specific concepts are encouraged to lie within the cones of their ancestors.

3 Geometric Analysis of Existing Hyperbolic VLMs

3.1 Evaluation Protocols

Curvature Exposure Analysis. We examine whether the learned embeddings occupy a region where the effects of negative curvature are non-negligible. For a Lorentz embedding $\mathbf{x} = (x_0, \tilde{\mathbf{x}})$ and its spatial norm $\|\tilde{\mathbf{x}}\|$, we formulate the dimensionless geodesic radius as $r(\mathbf{x}) = \sqrt{\kappa}\, d_{\mathbb{H}}(\mathbf{o}, \mathbf{x}) = \operatorname{arcsinh}(\sqrt{\kappa}\, \|\tilde{\mathbf{x}}\|)$. We measure the local curvature effect using the transverse stretch factor

$$T(\mathbf{x}) = \frac{\sqrt{\kappa}\, \|\tilde{\mathbf{x}}\|}{\operatorname{arcsinh}(\sqrt{\kappa}\, \|\tilde{\mathbf{x}}\|)} = \frac{\sinh r(\mathbf{x})}{r(\mathbf{x})}. \tag{7}$$

Table 1: **Geometric analysis of existing hyperbolic VLMs.** We report the 99th percentile to sensitively capture severe distortions in the upper tail. Detailed statistics are provided in Appendix D.1.

(a) Point geometric analysis.

Model	Modality	r	T
MERU	text	0.200	1.0067
	img	0.434	1.0317
HyCoCLIP	part text	0.121	1.0024
	text	0.141	1.0033
	part img	0.203	1.0069
	img	0.205	1.0070
PHyCLIP	part text	0.165	1.0045
	text	0.191	1.0061
	part img	0.251	1.0105
	img	0.260	1.0113
UNCHA	part text	0.201	1.0068
	text	0.200	1.0067
	part img	0.413	1.0287
	img	0.413	1.0287

(b) Pairwise geometric and cone validity analysis.

Model	Modality	$\mathbf{G}_{\mathrm{pos}}$	$\mathbf{G}_{\mathrm{neg}}$	NNA@1	$\mathbf{S}_{\mathrm{clamp}}$	$\mathbf{S}_{\mathrm{final}}$
MERU	text → img	0.65 %	0.65 %	99.91 %	90.18 %	94.45 %
HyCoCLIP	part text → part img	0.23 %	0.22 %	99.83 %	100.0 %	100.0 %
	part text → img	0.23 %	0.22 %	99.88 %	100.0 %	100.0 %
	text → part img	0.33 %	0.33 %	99.94 %	100.0 %	100.0 %
	text → img	0.33 %	0.33 %	99.96 %	100.0 %	100.0 %
PHyCLIP	part text → part img	0.28 %	0.28 %	99.84 %	98.15 %	99.86 %
	part text → img	0.28 %	0.27 %	99.83 %	98.15 %	99.86 %
	text → part img	0.39 %	0.39 %	99.84 %	95.32 %	99.54 %
	text → img	0.39 %	0.39 %	99.95 %	95.32 %	99.54 %
UNCHA	part text → part img	0.63 %	0.60 %	99.01 %	88.90 %	92.27 %
	part text → img	0.63 %	0.60 %	99.04 %	88.90 %	92.27 %
	text → part img	0.62 %	0.61 %	99.59 %	75.03 %	93.80 %
	text → img	0.60 %	0.60 %	99.69 %	75.03 %	93.80 %

Values close to 1 indicate a locally near-Euclidean embedding, whereas larger values indicate stronger curvature-induced expansion. Further derivation is provided in Appendix C.1.

Pairwise Distance Analysis. Beyond local metric stretch, we examine whether negative curvature materially alters the pairwise distances that govern similarity and retrieval decisions. With the learned hyperbolic embeddings and curvature held fixed, we recover their corresponding representations in the tangent space at the origin as $\mathbf{z}(\mathbf{x}) = \sqrt{\kappa}\log_{\mathbf{o}}^{\kappa}(\mathbf{x})$. We compute the gap between the exact dimensionless hyperbolic distance $\delta_H(\mathbf{x},\mathbf{y}) = \sqrt{\kappa}\, d_{\mathbb{H}}(\mathbf{x},\mathbf{y})$ and the Euclidean distance $\delta_T(\mathbf{x},\mathbf{y}) = \|\mathbf{z}(\mathbf{x}) - \mathbf{z}(\mathbf{y})\|_2$ on the same frozen latent coordinates as

$$G(\mathbf{x},\mathbf{y}) = 100 \times \left| 1 - \frac{\delta_T(\mathbf{x},\mathbf{y})}{\delta_H(\mathbf{x},\mathbf{y})} \right|. \tag{8}$$

A small G indicates that the hyperbolic metric is well approximated by the Euclidean metric in the origin tangent space. We report G_{pos} for matched positive pairs and G_{neg} for the hard negative pairs.

We further evaluate decision-level sensitivity using Nearest-Neighbor Agreement at top-k (*NNA@k*). With frozen models, *NNA@1* is the fraction of queries whose top-1 neighbor remains unchanged when hyperbolic distance is replaced by Euclidean distance. Thus, while G measures distance-level distortion, *NNA@k* measures its effect on decisions relevant to downstream tasks.

Cone Validity Analysis. From Eq. (5), nested transitivity constrains the half-aperture angle to at most $\omega = \pi/2$, corresponding to a maximum full aperture angle of π. However, as the cone aperture becomes excessively wide, a broad range of child embeddings can satisfy the same entailment constraint, substantially weakening its angular selectivity for distinguishing hierarchical directions. We therefore measure how frequently the entailment cones of the parent embeddings saturate to the maximum aperture. S_{clamp} denotes the saturation rate of cones at the first training step when the 99th percentile of $r(\mathbf{x})$ enters the near-Euclidean region, and S_{final} denotes the corresponding rate at the final checkpoint.

3.2 Experimental Setup

For fair comparisons, we independently retrained all baselines three times to conduct repeated experiments, using a ViT-B/16 backbone and the same training dataset. Following the original protocol, we reconstructed the GRIT (Peng et al., 2023) dataset, consisting of whole images and cropped part images paired with their corresponding whole and part text descriptions. All methods were trained for 500,000 iterations using 4 × H200 GPUs with a batch size of 768. For geometric analysis, we computed statistics on 20K training samples and report the 99th percentile, which more sensitively captures severe distortions in the upper tail that may be obscured by the mean or median.

3.3 FINDINGS

Existing Hyperbolic VLMs Operate in a Near-Euclidean Regime. Tab. 1a shows that most of the learned embeddings remain close to the Euclidean limit. Across all models and modalities, the 99th percentile of r ranges from 0.121 to 0.434, while the corresponding transverse stretch factor T ranges from 1.0024 to 1.0317. Even the largest observed transverse expansion is only 3.17% relative to the Euclidean reference $T = 1$. Image embeddings generally lie farther from the origin than text embeddings, as intended by existing hyperbolic VLMs that explicitly model images as more specific than their paired texts. Importantly, however, both modalities remain at relatively small radii, suggesting that the expected modality ordering is present without substantial departure from the near-Euclidean regime.

The pairwise results in Tab. 1b show the same pattern. Across all evaluated relations, G_{pos} and G_{neg} range from 0.22% to 0.65%, indicating that both positive and hard negative distances are closely approximated by Euclidean distances in the tangent space at the origin. Moreover, *NNA@1* ranges from 99.01% to 99.96%, meaning that the top-1 neighbor is preserved for almost all queries after replacing the hyperbolic metric with the Euclidean metric. Despite residing on a hyperbolic manifold, the learned representations behave nearly Euclidean at both the metric and decision levels.

Entailment Cones Operate Predominantly in the Saturated Regime. We conservatively define the near-Euclidean region as one where tangential expansion relative to the Euclidean reference is less than 1% $(i.e., T < 1.01)$. Across all baselines, the learned curvature parameter reaches its imposed lower bound of $\kappa = 0.1$ at approximately 7K iterations and remains there for the rest of the 500K iteration training. By approximately 10K iterations, the 99th percentile of $r(\mathbf{x})$ for parent embeddings enters the near-Euclidean region. We therefore report the cone saturation rate S_{clamp} at the 10K checkpoint, together with the S_{final} at the final checkpoint in Tab. 1b. By this early stage, more than 75% of the cones are already saturated across all models, with HyCoCLIP cones reaching the maximum aperture. At the final checkpoint, saturation remains above 92% for every relations, indicating that the saturated cone regime persists throughout the remainder of training.

3.4 WHY IS THE HYPERBOLIC GEOMETRY BROKEN?

Aperture Clipping Does Not Enforce the Cone-Domain Constraint. When parent embeddings yield an $\arcsin$ argument outside its domain [-1,1], existing hyperbolic VLMs apply clipping to the argument. Accordingly, Eq. (5) is implemented as

$$\hat{\omega}(\mathbf{p}) = \arcsin\left(\min\left\{1, \frac{2K}{\sqrt{\kappa}\|\tilde{\mathbf{p}}\|}\right\}\right). \tag{9}$$

When $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| < 2K$, the half-aperture of the entailment cone saturates at $\pi/2$. The aperture calculation thus remains numerically defined even though the parent lies outside the original cone domain. In contrast, the original Poincaré ϵ-ball formulation (Ganea et al., 2018a) excludes an ϵ-ball around the origin and projects embeddings outside it during optimization. *Clipping is not equivalent to this projection*—it changes the aperture value but leaves the cone apex unchanged, so the exterior angle is still evaluated at the original parent location. Within the saturated region, the half-aperture is fixed at $\pi/2$ and no longer varies with the parent radius. The parent embedding still determines the cone orientation, but the wide aperture admits child embeddings from a broad range of directions.

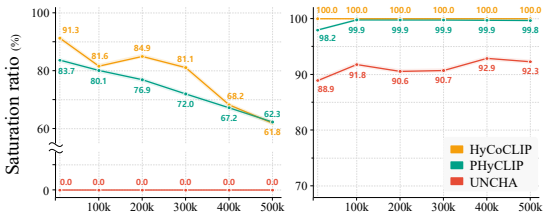
The domain constraint also depends on the learned curvature. With K fixed, as in prior works, Eq. (5) requires $\|\tilde{\mathbf{p}}\| \geq 2K/\sqrt{\kappa}$. A smaller κ therefore raises the minimum admissible spatial norm. Clipping alone does not ensure that parent embeddings satisfy the constraint during training.

Radial Contraction Can Relax the Entailment Constraint. An active entailment penalty can decrease through parent contraction alone, without adjusting the child embedding. To isolate the contribution of the parent norm, we differentiate the entailment hinge with respect to $\|\tilde{\mathbf{p}}\|$, holding κ and $\mathbf{q}$ fixed. Before saturation, where $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| > 2K$, we obtain

$$\frac{\partial \mathcal{L}_{\text{ent}}}{\partial \|\tilde{\mathbf{p}}\|} = \underbrace{\frac{\partial \phi(\mathbf{p}, \mathbf{q})}{\partial \|\tilde{\mathbf{p}}\|}}_{\text{exterior angle}} + \underbrace{\frac{2K}{\|\tilde{\mathbf{p}}\|\sqrt{\kappa\|\tilde{\mathbf{p}}\|^2 - 4K^2}}}_{\text{aperture}}. \tag{10}$$

For noncollinear radial directions, both terms are positive. Reducing the parent norm therefore decreases the exterior angle and widens the aperture, even when the child remains at the same location and neither embedding changes its direction from the origin. The penalty can consequently decrease while the parent approaches the boundary of the valid cone domain. Under the idealized clipping rule, the aperture term in Eq. (10) grows without bound as $\|\tilde{\mathbf{p}}\|$ approaches $2K/\sqrt{\kappa}$ from above for fixed K and κ. If the hinge remains active and the derivative of the exterior angle remains bounded during the approach, the aperture term dominates the parent radial derivative. Inside the saturated region, the aperture gradient vanishes, leaving only the exterior angle contribution for an active entailment hinge. Further results including empirical gradient dynamics are provided in Appendix D.2.

Cone Saturation During Training. We additionally track the saturation ratio across training checkpoints to determine whether aperture clipping is merely a transient phenomenon. Across the hierarchical relations, saturation remains substantial in most cases throughout training, as shown in Fig. 2 and Appendix D.3. For the *part image→image* relation, saturation decreases over training in HyCoCLIP and PHyCLIP but remains substantial at the final checkpoint, whereas UNCHA exhibits negligible saturation across the measured checkpoints. Under Eq. (9), a cone saturates when its parent norm falls below the minimum admissible value $2K/\sqrt{\kappa}$. Thus, a substantial fraction of the parent embeddings remains in the clipped regime even late in training, where the half-aperture no longer adapts to the parent radius.



(b) Part image → Image **(a) Part text → Text**

Figure 2: **Cone saturation ratio across 500K training iterations for intra-modal relations.**

4 METHOD

Motivated by our geometric analysis, we hypothesize that the gains of existing hyperbolic VLMs can be largely attributed to the additional supervision provided by whole–part relations rather than to negative curvature itself. To empirically assess the hypothesis, we introduce **EuHier**, a simple **Eu**clidean-based **Hier**archy-aware vision-language model that aligns visual and textual representations at multiple levels of granularity without any hyperbolic operations.

Each training sample consists of a whole image I, its corresponding caption T, and a set of part image and part text pairs (I^p, T^p). The image encoder $f_I(\cdot)$ is shared between whole images and part images, while the text encoder $f_T(\cdot)$ is shared between whole captions and part texts. Each encoder includes its respective projection heads and ℓ_2 normalization. Given a batch $\mathcal{B}$ of paired text and image samples, we formulate the directional InfoNCE (van den Oord et al., 2018) loss as

$$\mathcal{L}(T, I) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log \frac{\exp\big(f_T(\mathbf{t}_i)^\top f_I(\mathbf{i}_i)/\tau\big)}{\sum_{j \in \mathcal{B}} \exp(f_T(\mathbf{t}_i)^\top f_I(\mathbf{i}_j)/\tau)}, \tag{11}$$

where $(\mathbf{t}_i, \mathbf{i}_i)$ forms a positive pair, while the remaining images in the batch serve as in-batch negatives for the text anchor $\mathbf{t}_i$, and τ is the temperature. The ordering of the arguments in $\mathcal{L}(T, I)$ specifies the direction of contrastive matching, with T serving as the anchor set and I as the candidate set. For the local-to-global objectives, multiple parts originating from the same sample may share the same whole representation as their positive target. The total contrastive objective comprises six relations grouped into four alignment categories:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \frac{1}{2}\underbrace{\big(\mathcal{L}(T, I) + \mathcal{L}(I, T)\big)}_{\text{global}} + \frac{1}{2}\underbrace{\big(\mathcal{L}(T^p, I^p) + \mathcal{L}(I^p, T^p)\big)}_{\text{local}} \\ + \underbrace{\mathcal{L}(T^p, I) + \mathcal{L}(I^p, T)}_{\text{inter-modal}} + \underbrace{\mathcal{L}(T^p, T) + \mathcal{L}(I^p, I)}_{\text{intra-modal}}.\end{aligned} \tag{12}$$

The global objective aligns whole images with their captions, while the local objective aligns corresponding part images and part texts. The local-to-global objectives, comprising inter-modal and intra-modal terms, organize supervision from semantically general to specific—each part-level representation is treated as a broader concept and used as the anchor, while its corresponding whole

Table 2: **Zero-shot image classification evaluation.**

Model	General datasets						Fine-grained datasets						Misc. datasets				All
	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211	Average
Hyperbolic-based																	
MERU	37.4	66.6	38.9	48.8	72.2	92.4	47.6	10.7	6.3	2.5	43.9	19.0	22.3	31.4	43.2	4.5	36.7
HyCoCLIP	43.6	87.6	55.9	52.9	78.4	94.7	54.8	14.8	10.6	3.0	53.1	24.6	24.4	29.0	47.1	4.9	42.5
PHyCLIP	44.3	88.7	59.7	54.7	77.9	94.8	57.4	**17.2**	10.6	3.7	55.4	24.8	25.8	41.2	47.9	5.1	44.3
UNCHA	49.3	90.1	**63.8**	56.9	83.0	**96.3**	56.3	15.4	**13.9**	**4.2**	57.2	**26.6**	30.6	30.5	50.1	6.0	45.6
Euclidean-based																	
CLIP	37.3	64.6	41.5	49.0	73.9	92.2	47.9	11.4	6.2	2.6	42.1	21.5	22.9	**33.0**	43.4	4.4	37.1
w/ 1-part	43.8	89.2	59.5	53.4	80.5	95.6	51.9	14.1	11.4	2.8	53.5	23.3	26.6	29.3	48.2	5.1	43.0
w/ 5-part	44.8	89.7	61.3	54.7	81.6	96.0	55.9	13.5	11.2	2.8	54.5	21.6	28.3	30.8	47.5	5.3	43.7
EuHier (Ours)	**49.5**	89.9	62.3	**57.7**	**83.8**	95.9	**59.3**	16.1	13.5	3.8	**59.1**	26.1	**32.6**	27.8	**50.5**	**6.5**	**45.9**

Table 3: **Zero-shot retrieval and hierarchical evaluation.**

Model	Text retrieval				Image retrieval				Hierarchical metrics				
	COCO		Flickr		COCO		Flickr		TIE↓	LCA↓	J↑	P_H ↑	R_H ↑
	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10					
Hyperbolic-based													
MERU	67.5	78.4	91.0	95.0	55.5	66.9	81.1	88.2	3.801	2.272	0.774	0.847	0.846
HyCoCLIP	68.4	78.7	91.5	95.4	56.3	67.6	82.0	88.9	3.305	2.094	0.806	0.867	0.868
PHyCLIP	69.8	79.9	90.9	95.0	57.0	68.1	82.7	89.2	3.667	2.203	0.782	0.853	0.848
UNCHA	73.8	82.5	93.8	96.8	60.9	71.8	86.5	92.2	2.899	**1.936**	0.831	**0.885**	0.885
Euclidean-based													
CLIP	69.2	79.1	91.8	96.1	55.6	67.3	82.5	89.3	3.823	2.298	0.774	0.845	0.846
w/ 1-part	70.6	80.4	91.8	95.8	57.1	68.3	83.4	89.8	3.317	2.096	0.804	0.867	0.866
w/ 5-part	71.1	80.7	92.8	96.9	57.9	69.4	84.9	90.9	3.227	2.071	0.811	0.871	0.872
EuHier (Ours)	**75.0**	**83.8**	**94.7**	**97.6**	**62.4**	**72.9**	**88.1**	**93.3**	**2.884**	1.950	**0.833**	**0.885**	**0.887**

representation serves as the context-enriched, more specific target. EuHier thus retains the multi-granularity supervision used by recent hyperbolic VLMs without relying on hyperbolic geometry or entailment constraints.

5 Experiments

5.1 Baselines

We compare against representative methods (Desai et al., 2023; Pal et al., 2025; Yoshikawa & Matsubara, 2026; Kim et al., 2026) that model the representation space using hyperbolic geometry and employ entailment cones. To account for the effect of data augmentation, we report CLIP (Radford et al., 2021) not only when trained on the original (I, T) pairs, but also when (I^p, T^p) pairs are treated equivalently to the whole pairs during training. Following the HyCoCLIP setting, CLIP w/ 1-part additionally uses one randomly sampled part image from each whole image, whereas CLIP w/ 5-part follows the UNCHA setting and uses up to five part images per image.

5.2 Downstream Tasks

Zero-shot Image Classification. We evaluate zero-shot top-1 classification accuracy on 16 benchmark datasets, as reported in Tab. 2. Following prior works (Pal et al., 2025; Kim et al., 2026), we use multiple text prompt templates for each class label and encode the prompts with the text encoder. The average of the resulting embeddings is used as the class representation for computing image–text similarity. Our model achieves the highest average accuracy across the benchmarks. Notably,

Table 4: **Scene understanding evaluation.** O-Large (O-L), O-Medium (O-M), and O-Small (O-S) are subsets of VL-CO grouped by object bounding-box area in pixels. All results are reported as accuracy.

Model	VL-CO	VG-A	O-L	O-M	O-S
MERU	76.6	58.5	84.1	74.3	67.2
HyCoCLIP	78.9	61.5	86.5	77.3	67.6
PHyCLIP	80.0	64.3	87.1	78.4	70.8
UNCHA	81.0	65.0	88.0	79.8	71.1
CLIP	77.6	59.3	84.4	75.3	69.6
w/ 1-part	77.6	59.7	85.1	75.9	68.2
w/ 5-part	78.1	60.0	85.0	76.3	69.7
EuHier (Ours)	**82.5**	**69.3**	**88.9**	**82.0**	**71.9**

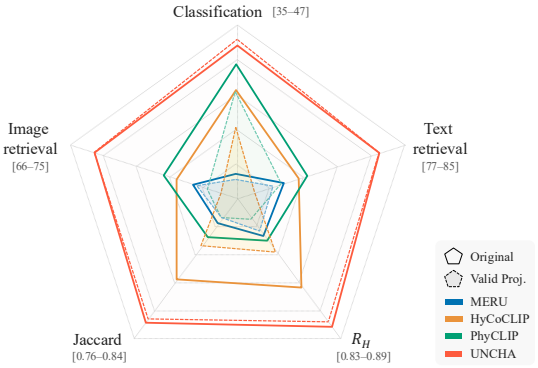


Figure 3: **Effect of valid projection on hyperbolic VLMs.** We compare the original models (—) with variants (····) that project to the admissible radius.

augmenting CLIP with part-level supervision alone substantially improves performance, resulting in a Euclidean baseline that outperforms prior hyperbolic VLMs such as MERU and HyCoCLIP.

Zero-shot Cross-modal Retrieval. As reported in Tab. 3, we assess cross-modal retrieval performance using Recall@K on the COCO (Lin et al., 2014) and Flickr-30K (Young et al., 2014) benchmarks. For text retrieval, candidate captions are ranked by their similarity to a query image, while image retrieval ranks candidate images given a query text. Consistent with the image classification results, CLIP augmented with additional part data achieves performance comparable to existing hyperbolic VLMs, and our model yields the best overall results.

Hierarchical Classification. Following the evaluation metrics introduced in HyCoCLIP, we measure the distance between the predicted and ground-truth labels in the WordNet taxonomy tree (Miller, 1995; Deng et al., 2009), as well as ancestor-based distance metrics and set-based measures. Specifically, TIE measures the tree distance between the predicted and ground-truth labels, whereas LCA measures the distance from the predicted label to their lowest common ancestor. J, P_H, and R_H measure the Jaccard similarity, precision, and recall of their ancestor sets, respectively. As reported in Tab. 3, our method achieves the best overall performance despite not relying on a hyperbolic space that explicitly models hierarchical structure.

Scene Understanding. We evaluate scene understanding on VL-CO (Zhao et al., 2022), which assesses object recognition across different sizes and locations, and VG-A (Yüksekgönül et al., 2023), which tests attribute binding by distinguishing correct captions from hard negatives. As reported in Tab. 4, EuHier consistently outperforms the baselines, with particularly strong gains on VG-A. Its performance on O-Large, O-Medium, and O-Small further indicates effective recognition across varying spatial granularities within a scene. Notably, adding part-level supervision also improves CLIP while EuHier outperforms the hyperbolic baselines, suggesting that these gains are more closely associated with hierarchy-aware supervision than with hyperbolic geometry itself.

5.3 Ablation Study

Valid Projection on the Baselines. As discussed in Sec. 3.4, we radially project any invalid apex with $r < \operatorname{asinh}(2K/\sin\omega_{\max})$ onto the admissible boundary before computing the aperture rather than clipping it. We retain the original cone parameter $K = 0.1$ and conservatively set $\omega_{\max} = 80°$ to prevent the cone from saturating toward a half-space. However, Fig. 3 shows that applying the valid projection tends to reduce baseline performance, suggesting that geometric validity alone does not guarantee more informative hierarchical representations. There are two possible explanations: (i) Even within the valid region, smaller parent radii still produce wider cones, meaning that projection alone cannot prevent solutions from remaining near the admissible boundary or ensure a diverse radial structure that reflects differences in conceptual generality. (ii) Outward projection with a capped half-aperture tightens parent–child containment relative to the clipped formulation,

Table 5: **Ablation study.** *EuHier-H* replaces the Euclidean metric with hyperbolic metric, keeping all other settings unchanged.

Model	Classification			Retrieval		Hierarchical metrics				
	General	Fine	MISC.	Text	Image	TIE ↓	LCA ↓	J ↑	P_H ↑	R_H ↑
EuHier	**73.2**	**29.7**	**29.4**	84.9	**75.3**	**2.884**	**1.950**	**0.833**	**0.885**	**0.887**
w/o inter-modal	71.5	25.8	28.4	82.6	72.6	3.133	2.043	0.817	0.875	0.877
w/o intra-modal	72.7	28.2	27.9	84.2	74.5	2.930	**1.950**	0.830	0.884	0.884
w/o local	72.4	29.5	29.1	**85.2**	75.0	2.923	1.957	0.830	0.884	0.885
EuHier-H	70.9	27.3	25.5	84.3	75.0	3.230	2.030	0.810	0.874	0.867

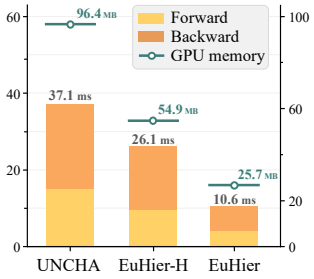


Figure 4: **Post-encoder efficiency comparisons.**

potentially shifting the balance between entailment and contrastive alignment. The results indicate that the entailment objective itself should impose hierarchical structure while remaining compatible with the contrastive objective. Detailed implementation is described in Appendix G.

Effects of Loss Terms. Tab. 5 shows the contribution of each supervision term. Removing the inter-modal objective causes the most consistent degradation, indicating that cross-modal hierarchical supervision contributes most strongly to learning hierarchical relations in EuHier. The intra-modal objective provides a smaller but clear benefit, suggesting that within-modality supervision further improves the learned representations. Removing the local objective has only a limited effect, leaving text retrieval unchanged and slightly reducing the remaining metrics. Overall, the full model achieves consistently strong performance across both downstream tasks and hierarchical metrics.

Effect of the Hyperbolic Metric. We further examine whether hyperbolic geometry provides benefits beyond hierarchy-aware supervision. EuHier-H retains the same architecture and hierarchical contrastive supervision as EuHier, while mapping the projected features onto the Lorentz hyperboloid via the exponential map at the origin. Cosine similarity in InfoNCE is then replaced with negative geodesic distance. As shown in Tab. 5, EuHier-H consistently yields lower performance across the evaluated metrics. Under the same hierarchy-aware supervision, the hyperbolic formulation provides no additional benefit over EuHier. Together with the loss ablations, the results show that whole–part objectives remain effective without hyperbolic parameterization.

Computational Efficiency. We compare UNCHA, EuHier-H, and EuHier using an isolated single-GPU post-encoder benchmark under matched input workloads. We measure forward latency, backward latency, and additional peak GPU memory. As shown in Fig. 4, EuHier achieves the lowest cost across all metrics. Together with its strong downstream performance, the result highlights the practical advantage of EuHier—effective hierarchy-aware representation learning can be achieved with a simple Euclidean formulation, without the additional computational overhead introduced by hyperbolic operations. Detailed protocols are provided in Appendix F.1.

6 Conclusion

In this work, we show that the combination of hierarchical contrastive learning and entailment learning, widely adopted in Hyperbolic VLMs, fails to induce the intended hyperbolic geometry. The learned representations instead concentrate in a restricted region near the origin of a low-curvature manifold. The resulting regime is mathematically near-Euclidean and does not substantially exploit the exponential volume growth of hyperbolic space. To empirically validate the observation, we propose EuHier, which learns the same hierarchical supervision in Euclidean space. EuHier outperforms existing Hyperbolic VLMs, suggesting that a substantial portion of the performance gains reported by existing methods may arise from additional part–whole supervision rather than from hyperbolic geometry itself. The results highlight the need to go beyond merely verifying that representations lie on the hyperbolic manifold to carefully examine the causal contribution of hyperbolic geometry to learned representations and downstream performance. Notably, our study focuses on identifying the geometric failure in existing Hyperbolic VLMs. We leave redesigning the learning objectives for reliable entailment-based hierarchy formation as future work.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammatical correctness and readability of the manuscript, assist with translation, and assist in the writing of proofs. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate the reproducibility of our theoretical and empirical results. All assumptions and complete proofs of the theoretical results are provided in Appendix B-C. The experimental setup, including dataset, model configurations, and hyperparameter settings, is described in Section 3.2 and Appendix G. Detailed evaluation protocols and metrics are described in Appendix E. Detailed statistics from repeated experimental runs, including measures of variability, are reported throughout the Appendix D-F to support the robustness and reproducibility of the empirical results. All source code, configuration files, and supporting materials required to reproduce our results have been prepared and will be publicly released upon publication.

REFERENCES

Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. Fully hyperbolic convolutional neural networks for computer vision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

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)*, 2014.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017.

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

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

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2004.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

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)*, 2018a.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018b.

Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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*, 2019.

Chuong Huynh, Hossein Souri, Abhinav Kumar, Vitali Petsiuk, Deen Dayal Mohan, and Suren Kumar. Argent: Adaptive hierarchical image-text representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

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)*, 2020.

Hayeon Kim, Ji Ha Jang, Junghun James Kim, and Se Young Chun. Uncertainty-guided compositional alignment with part-to-whole semantic representativeness in hyperbolic vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Sanghwan Kim, Rui Xiao, Mariana-Iuliana Georgescu, Stephan Alaniz, and Zeynep Akata. COSMOS: Cross-modality self-distillation for vision language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

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 *4th International IEEE Workshop on 3D Representation and Recognition (3dRR)*, 2013.

Dmitri V. Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 82:036106, 2010.

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

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

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, 2018.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Tsung-Yi Lin, Michael Maire, Serge J. 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)*, 2014.

Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Teng Long, Pascal Mettes, Heng Tao Shen, and Cees G. M. Snoek. Searching for actions on the hyperbole. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38:39–41, 1995.

Gal Mishne, Zhengchao Wan, Yusu Wang, and Sheng Yang. The numerical stability of hyperbolic representation learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 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)*, 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)*, 2008.

Zachary Novack, Julian McAuley, Zachary C. Lipton, and Saurabh Garg. CHiLS: Zero-shot image classification with hierarchical label sets. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

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)*, 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)*, 2024.

Scott E. Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium Graph Drawing and Network Visualization*, 2011.

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

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Max van Spengler, Erwin Berkhout, and Pascal Mettes. Poincaré resnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Menglin Yang, Min Zhou, Rex Ying, Yankai Chen, and Irwin King. Hyperbolic representation learning: Revisiting and advancing. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Daiki Yoshikawa and Takashi Matsubara. PHyCLIP: ℓ1-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.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel C. F. Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

APPENDIX

TABLE OF CONTENTS

A Further Related Work **15**

A.1 Vision-Language Representation Learning 15

A.2 Hyperbolic Representation Learning 16

A.3 Hyperbolic Vision-Language Models 16

B Additional Preliminaries **16**

B.1 Mathematical Preliminaries 16

B.2 Additional Details on Hyperbolic Entailment Cones 17

C Derivations for Geometric Analysis **17**

C.1 Curvature Exposure Metric 17

C.2 Pairwise Distance Gap 18

C.3 Entailment Loss Decomposition 18

D Details on Geometric Analysis **19**

D.1 Detailed Statistics for Geometric Analysis 19

D.2 Entailment Gradients During Early Training 20

D.3 Saturation Dynamics Throughout Training 20

E Details on Downstream Experiments **21**

E.1 Benchmarks 21

E.2 Detailed Statistics for Downstream Tasks 22

F Further Experiments **23**

F.1 Efficiency Experiments 23

G Implementation Details **24**

H Limitation **25**

A FURTHER RELATED WORK

A.1 VISION-LANGUAGE REPRESENTATION LEARNING

Vision-language models (VLMs) have achieved strong performance on downstream tasks such as image-text retrieval and zero-shot image classification by aligning visual and textual representations in a shared embedding space (Radford et al., 2021; Jia et al., 2021). Early methods used ranking objectives to associate images with category labels (Frome et al., 2013), sentences (Kiros et al., 2014), or sentence fragments (Karpathy et al., 2014). Fine-grained approaches leveraged natural-language descriptions of discriminative visual properties for recognition (Reed et al., 2016; He & Peng, 2017), while image regions were aligned with words (Karpathy & Fei-Fei, 2015; Lee et al., 2018) or structured semantic components (Wu et al., 2019). Subsequent work increasingly adopted large-scale vision-language pretraining, where objectives such as masked prediction and image-text matching facilitated transfer to diverse vision-language tasks (Lu et al., 2019; Chen et al., 2020). CLIP (Radford et al., 2021) demonstrated that contrastive pretraining on large-scale image-text pairs yields transferable representations and effective zero-shot recognition. ALIGN (Jia et al., 2021) scaled the contrastive paradigm to massive noisy image-text collections, while ALBEF (Li et al., 2021) combined contrastive alignment with cross-attention fusion and momentum distillation. More recent

work has moved beyond global image-text alignment by incorporating fine-grained cross-modal correspondences (Yao et al., 2022; Zhong et al., 2022) and global-local consistency objectives (Kim et al., 2025).

A.2 Hyperbolic Representation Learning

Hyperbolic geometry is a non-Euclidean geometry characterized by constant negative curvature. Its exponential volume growth makes hyperbolic space particularly well suited to representing tree-structured data (Krioukov et al., 2010; Nickel & Kiela, 2017). Moreover, finite trees admit low-distortion embeddings even in low-dimensional hyperbolic space (Sarkar, 2011; Sala et al., 2018). Early approaches demonstrated that learned embeddings can jointly capture similarity and latent hierarchical structure (Nickel & Kiela, 2017; 2018). In particular, Hyperbolic entailment cones explicitly model the directionality of hierarchical relations by encoding partial orders through nested cones (Ganea et al., 2018a; Le et al., 2019). Hyperbolic neural networks extend representation learning beyond directly optimizing entity embeddings by introducing learnable feature transformations in hyperbolic space (Ganea et al., 2018b; Shimizu et al., 2021; Chen et al., 2022). Hyperbolic representations have been applied to graph learning (Chami et al., 2019; Liu et al., 2019) and hierarchical word embeddings (Tifrea et al., 2019; Le et al., 2019). In computer vision, hyperbolic learning has been explored for visual representation learning (Khrulkov et al., 2020; Ge et al., 2023), semantic segmentation (Atigh et al., 2022), and hierarchical action search (Long et al., 2020). Further work has developed fully hyperbolic visual backbones using convolutional and residual architectures (van Spengler et al., 2023; Bdeir et al., 2024). Alongside these developments, prior work has examined the numerical limitations of hyperbolic representations (Mishne et al., 2023) and whether optimization recovers the intended hierarchical structure (Yang et al., 2023).

A.3 Hyperbolic Vision-Language Models

Building on advances in vision-language representation learning, hyperbolic VLMs seek to capture semantic structure beyond image–text alignment. MERU (Desai et al., 2023) motivates hyperbolic geometry as a compact and interpretable space for visual-semantic hierarchies, where asymmetric entailment distinguishes general textual concepts from their more specific visual instances. HyCoCLIP (Pal et al., 2025) extends hierarchical supervision to grounded object crops and text fragments, treating isolated parts as more general than the wholes that provide their context. Its compositional objectives model relationships across and within modalities, with reported gains in zero-shot and hierarchical classification. UNCHA (Kim et al., 2026) further refines whole–part alignment by using radius-based uncertainty to account for differences in part representativeness, reporting more accurate whole–part ordering and improved multi-object understanding. PHyCLIP (Yoshikawa & Matsubara, 2026) extends the embedding geometry to an ℓ_1-product of hyperbolic factors, jointly modeling taxonomic hierarchies within concept families and composition across families.

Despite these developments, it remains unclear whether trained hyperbolic VLMs actually realize and rely on the geometric mechanisms they are designed to exploit. A related line of work has identified conflicts in hyperbolic objectives and instabilities in cone-based supervision (Ramasinghe et al., 2024; Huynh et al., 2026), but primarily seeks improved hyperbolic formulations rather than examining whether hyperbolic geometry is meaningfully exploited. Notably, we audit geometric utilization at the representation, metric, and decision levels by jointly measuring learned curvature and embedding radii. We further analyze cone saturation and whether valid projection restores the intended geometry. Finally, we disentangle geometric effects from compositional supervision through Euclidean comparisons.

B Additional Preliminaries

B.1 Mathematical Preliminaries

For an ambient vector $\mathbf{a} \in \mathbb{R}^{d+1}$, its Lorentz-orthogonal projection onto $T_{\mathbf{x}}\mathbb{H}^d_\kappa$ is

$$\mathrm{proj}^\kappa_{\mathbf{x}}(\mathbf{a}) = \mathbf{a} + \kappa\langle \mathbf{x}, \mathbf{a}\rangle_{\mathbb{H}}\mathbf{x}. \tag{13}$$

For $\mathbf{x} \neq \mathbf{y}$, the logarithmic map has the closed form

$$\log_{\mathbf{x}}^{\kappa}(\mathbf{y}) = \frac{\operatorname{arcosh}(-\kappa\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}})}{\sqrt{(\kappa\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{H}})^2 - 1}} \operatorname{proj}_{\mathbf{x}}^{\kappa}(\mathbf{y}), \tag{14}$$

with $\log_{\mathbf{x}}^{\kappa}(\mathbf{x}) = \mathbf{0}$. Since the tangent space at the origin is isometric to Euclidean $\mathbb{R}^d$, an encoder output $\tilde{\mathbf{v}} \in \mathbb{R}^d$ is interpreted as the tangent vector $\mathbf{v} = (0, \tilde{\mathbf{v}}) \in T_{\mathbf{o}}\mathbb{H}_{\kappa}^d$. The exponential map then maps $\mathbf{v}$ to a point on the Lorentz hyperboloid as

$$\exp_{\mathbf{o}}^{\kappa}(\mathbf{v}) = \left(\frac{\cosh(\sqrt{\kappa}\|\tilde{\mathbf{v}}\|)}{\sqrt{\kappa}}, \frac{\sinh(\sqrt{\kappa}\|\tilde{\mathbf{v}}\|)}{\sqrt{\kappa}\|\tilde{\mathbf{v}}\|}\tilde{\mathbf{v}} \right), \tag{15}$$

with $\exp_{\mathbf{o}}^{\kappa}(\mathbf{0}) = \mathbf{o}$. Conversely, for $\mathbf{x} \neq \mathbf{o}$, the logarithmic map at the origin reduces to

$$\log_{\mathbf{o}}^{\kappa}(\mathbf{x}) = \left(0, \frac{\operatorname{arcsinh}(\sqrt{\kappa}\|\tilde{\mathbf{x}}\|)}{\sqrt{\kappa}\|\tilde{\mathbf{x}}\|}\tilde{\mathbf{x}} \right), \tag{16}$$

where $\log_{\mathbf{o}}^{\kappa}(\mathbf{o}) = \mathbf{0}$.

B.2 Additional Details on Hyperbolic Entailment Cones

For parent and child embeddings $\mathbf{p}, \mathbf{q} \in \mathbb{H}_{\kappa}^d$ with $\mathbf{p} \neq \mathbf{o}$ and $\mathbf{q} \neq \mathbf{p}$, the exterior angle $\phi(\mathbf{p}, \mathbf{q})$ is formulated at $\mathbf{p}$ between the outward radial direction and the initial direction of the geodesic toward $\mathbf{q}$. It has the closed form

$$\phi(\mathbf{p}, \mathbf{q}) = \arccos\left(\frac{q_0 + \kappa p_0 \langle \mathbf{p}, \mathbf{q}\rangle_{\mathbb{H}}}{\|\tilde{\mathbf{p}}\|\sqrt{(\kappa\langle \mathbf{p}, \mathbf{q}\rangle_{\mathbb{H}})^2 - 1}} \right). \tag{17}$$

Let $r(\mathbf{p}) = \sqrt{\kappa}\, d_{\mathbb{H}}(\mathbf{o}, \mathbf{p})$ denote the dimensionless geodesic radius. Since $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| = \sinh r(\mathbf{p})$, the half-aperture in Eq. (5) can equivalently be written as

$$\omega(\mathbf{p}) = \arcsin\left(\frac{2K}{\sinh r(\mathbf{p})} \right). \tag{18}$$

For $K > 0$, the canonical half-aperture is real-valued if and only if $r(\mathbf{p}) \geq \operatorname{arcsinh}(2K)$, equivalently $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| \geq 2K$.

C Derivations for Geometric Analysis

C.1 Curvature Exposure Metric

For $\mathbf{x} = (x_0, \tilde{\mathbf{x}}) \in \mathbb{H}_{\kappa}^d$, the hyperboloid constraint implies $-\kappa\langle \mathbf{o}, \mathbf{x}\rangle_{\mathbb{H}} = \sqrt{\kappa}\, x_0 = \sqrt{1 + \kappa\|\tilde{\mathbf{x}}\|^2}$. Substituting this identity into the geodesic distance gives the dimensionless geodesic radius

$$r(\mathbf{x}) = \sqrt{\kappa}\, d_{\mathbb{H}}(\mathbf{o}, \mathbf{x}) = \operatorname{arcosh}\left(\sqrt{1 + \kappa\|\tilde{\mathbf{x}}\|^2} \right) = \operatorname{arcsinh}\left(\sqrt{\kappa}\, \|\tilde{\mathbf{x}}\| \right). \tag{19}$$

For $\mathbf{x} \neq \mathbf{o}$, let $\mathbf{s}_{\mathbf{x}} = \tilde{\mathbf{x}}/\|\tilde{\mathbf{x}}\|$ denote its spatial direction. Writing $r_{\mathbf{x}} = r(\mathbf{x})$ and using $\mathbf{s}_{\mathbf{x}}$ for the spatial direction, the dimensionless Lorentz coordinates are

$$\sqrt{\kappa}\, \mathbf{x} = (\cosh r_{\mathbf{x}}, \sinh r_{\mathbf{x}}\, \mathbf{s}_{\mathbf{x}}). \tag{20}$$

Identifying $T_{\mathbf{o}}\mathbb{H}_{\kappa}^d$ with $\mathbb{R}^d$, we represent the corresponding dimensionless tangent coordinates as $\mathbf{z}(\mathbf{x}) = \sqrt{\kappa}\log_{\mathbf{o}}^{\kappa}(\mathbf{x}) = r_{\mathbf{x}}\mathbf{s}_{\mathbf{x}}$, with $\mathbf{z}(\mathbf{o}) = \mathbf{0}$. Since $\|\mathbf{s}_{\mathbf{x}}\| = 1$, we have $\mathbf{s}_{\mathbf{x}}^{\top} d\mathbf{s}_{\mathbf{x}} = 0$. Differentiating Eq. (20) gives

$$\sqrt{\kappa}\, d\mathbf{x} = (\sinh r_{\mathbf{x}}\, dr_{\mathbf{x}}, \cosh r_{\mathbf{x}}\, \mathbf{s}_{\mathbf{x}}\, dr_{\mathbf{x}} + \sinh r_{\mathbf{x}}\, d\mathbf{s}_{\mathbf{x}}). \tag{21}$$

Applying the Lorentzian metric yields

$$\kappa\langle d\mathbf{x}, d\mathbf{x}\rangle_{\mathbb{H}} = -(\sinh r_{\mathbf{x}}\, dr_{\mathbf{x}})^2 + \|\cosh r_{\mathbf{x}}\, \mathbf{s}_{\mathbf{x}}\, dr_{\mathbf{x}} + \sinh r_{\mathbf{x}}\, d\mathbf{s}_{\mathbf{x}}\|^2 = dr_{\mathbf{x}}^2 + \sinh^2 r_{\mathbf{x}}\, d\Omega^2, \tag{22}$$

where $d\Omega^2 = \|d\mathbf{s}_{\mathbf{x}}\|^2$ denotes the standard metric on the unit sphere $\mathbb{S}^{d-1}$. Since $\|\mathbf{z}(\mathbf{x})\| = r_{\mathbf{x}}$, the pair $(r_{\mathbf{x}}, \mathbf{s}_{\mathbf{x}})$ gives the ordinary polar coordinates of $\mathbf{z}(\mathbf{x})$ in $T_{\mathbf{o}}\mathbb{H}_{\kappa}^d \simeq \mathbb{R}^d$. For the same infinitesimal

variation, $d\mathbf{z} = \mathbf{s}_\mathbf{x}\, dr_\mathbf{x} + r_\mathbf{x}\, d\mathbf{s}_\mathbf{x}$. Using $\mathbf{s}_\mathbf{x}^\top d\mathbf{s}_\mathbf{x} = 0$ again, the Euclidean line element between neighboring dimensionless logarithmic coordinates is

$$\|d\mathbf{z}\|^2 = dr_\mathbf{x}^2 + r_\mathbf{x}^2 d\Omega^2. \tag{23}$$

Comparing Eq. (22) and Eq. (23) shows that their radial components coincide—both dimensionless line elements reduce to $|dr_\mathbf{x}|$ for $d\Omega = 0$. For a nonzero transverse displacement, characterized by $dr_\mathbf{x} = 0$, the ratio of the corresponding dimensionless lengths is

$$T(\mathbf{x}) = \frac{\sqrt{\kappa\langle d\mathbf{x}, d\mathbf{x}\rangle_\mathbb{H}}}{\|d\mathbf{z}\|} = \frac{\sinh r_\mathbf{x}}{r_\mathbf{x}} = \frac{\sqrt{\kappa}\|\tilde{\mathbf{x}}\|}{\operatorname{arcsinh}(\sqrt{\kappa}\|\tilde{\mathbf{x}}\|)}. \tag{24}$$

At the origin, the factor is defined by continuity as $T(\mathbf{o}) = 1$. Its expansion around $r_\mathbf{x} = 0$ is

$$T(\mathbf{x}) - 1 = \frac{r_\mathbf{x}^2}{6} + \mathcal{O}\big(r_\mathbf{x}^4\big)\,. \tag{25}$$

Consequently, the relative transverse length distortion $T(\mathbf{x}) - 1$, measured against the origin-centered tangent representation, vanishes quadratically as $r_\mathbf{x} \to 0$.

C.2 Pairwise Distance Gap

For two non-origin points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^d_\kappa$, let $r_\mathbf{x} = r(\mathbf{x})$, $r_\mathbf{y} = r(\mathbf{y})$, and $\cos\theta_{\mathbf{x},\mathbf{y}} = \mathbf{s}_\mathbf{x}^\top \mathbf{s}_\mathbf{y}$. On the tangent plane at the origin, we define the dimensionless representation as $\mathbf{z}(\mathbf{x}) = \sqrt{\kappa}\log_\mathbf{o}^\kappa(\mathbf{x})$. Since its time coordinate is zero, it can be identified with its spatial component as $\mathbf{z}(\mathbf{x}) = (0, r_\mathbf{x}\mathbf{s}_\mathbf{x})$ and $\|\mathbf{z}(\mathbf{x})\| = r_\mathbf{x}$. Using Eq. (20), the exact dimensionless hyperbolic distance is

$$\delta_\mathbb{H}(\mathbf{x}, \mathbf{y}) = \sqrt{\kappa}\, d_\mathbb{H}(\mathbf{x}, \mathbf{y}) = \operatorname{arcosh}(\cosh r_\mathbf{x} \cosh r_\mathbf{y} - \sinh r_\mathbf{x} \sinh r_\mathbf{y} \cos\theta_{\mathbf{x},\mathbf{y}})\,. \tag{26}$$

The corresponding distance in the dimensionless origin tangent space is

$$\delta_\mathrm{T}(\mathbf{x}, \mathbf{y}) = \|\mathbf{z}(\mathbf{x}) - \mathbf{z}(\mathbf{y})\| = \sqrt{r_\mathbf{x}^2 + r_\mathbf{y}^2 - 2r_\mathbf{x} r_\mathbf{y} \cos\theta_{\mathbf{x},\mathbf{y}}}. \tag{27}$$

As $r_\mathbf{x}, r_\mathbf{y} \to 0$, their squared-distance difference has the expansion

$$\delta^2_\mathbb{H}(\mathbf{x}, \mathbf{y}) = \delta^2_\mathrm{T}(\mathbf{x}, \mathbf{y}) + \frac{1}{3} r_\mathbf{x}^2 r_\mathbf{y}^2 \sin^2\theta_{\mathbf{x},\mathbf{y}} + \mathcal{O}\big((r_\mathbf{x} + r_\mathbf{y})^6\big)\,. \tag{28}$$

Thus, the leading curvature-dependent correction scales as $r_x^2 r_y^2 \sin^2\theta_{\mathbf{x},\mathbf{y}}$. Since this correction is fourth order in the radii whereas the squared tangent-space distance is second order, the relative gap G vanishes for nondegenerate pairs as $r_x, r_y \to 0$, providing a quantitative measure of pairwise metric agreement in the near-origin regime.

C.3 Entailment Loss Decomposition

We derive the parent-radial derivative used in the main text. Consider a strictly active entailment hinge, $\phi(\mathbf{p}, \mathbf{q}) > \omega(\mathbf{p})$, in the valid cone domain $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| > 2K$. We vary $\|\tilde{\mathbf{p}}\|$ while holding K, κ, $\mathbf{q}$, and the spatial direction $\tilde{\mathbf{p}}/\|\tilde{\mathbf{p}}\|$ fixed. Accordingly, $p_0 = \sqrt{1/\kappa + \|\tilde{\mathbf{p}}\|^2}$ varies so that $\mathbf{p}$ remains on $\mathbb{H}^d_\kappa$. Differentiating the half-aperture gives

$$-\frac{\partial\omega(\mathbf{p})}{\partial\|\tilde{\mathbf{p}}\|} = \frac{2K}{\|\tilde{\mathbf{p}}\|\sqrt{\kappa\|\tilde{\mathbf{p}}\|^2 - 4K^2}} > 0. \tag{29}$$

For nonzero, noncollinear spatial components, let $\theta \in (0, \pi)$ denote the Euclidean angle between $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{q}}$. Direct differentiation of the exterior-angle expression gives

$$\frac{\partial\phi(\mathbf{p}, \mathbf{q})}{\partial\|\tilde{\mathbf{p}}\|} = \frac{-\kappa^2\|\tilde{\mathbf{q}}\|\langle\mathbf{p}, \mathbf{q}\rangle_\mathbb{H} \sin\theta}{\sqrt{1 + \kappa\|\tilde{\mathbf{p}}\|^2}\left[\big(\kappa\langle\mathbf{p}, \mathbf{q}\rangle_\mathbb{H}\big)^2 - 1\right]} > 0. \tag{30}$$

Indeed, $-\kappa\langle\mathbf{p}, \mathbf{q}\rangle_\mathbb{H} = \cosh(\sqrt{\kappa}\, d_\mathbb{H}(\mathbf{p}, \mathbf{q})) > 1$ for $\mathbf{p} \neq \mathbf{q}$, while $\sin\theta > 0$. Therefore, while the hinge is active,

$$\frac{\partial\mathcal{L}_\mathrm{ent}}{\partial\|\tilde{\mathbf{p}}\|} = \frac{\partial\phi(\mathbf{p}, \mathbf{q})}{\partial\|\tilde{\mathbf{p}}\|} + \frac{2K}{\|\tilde{\mathbf{p}}\|\sqrt{\kappa\|\tilde{\mathbf{p}}\|^2 - 4K^2}} > 0. \tag{31}$$

Table 6: **Detailed point geometric analysis of existing hyperbolic VLMs.**

Model	Modality	r	T	Model	Modality	r	T
MERU	part text	–	–	PHyCLIP	part text	0.1650 ± 0.0000	1.0045 ± 0.0001
	text	0.2000 ± 0.0000	1.0067 ± 0.0000		text	0.1910 ± 0.0000	1.0061 ± 0.0000
	part img	–	–		part img	0.2510 ± 0.0000	1.0105 ± 0.0001
	img	0.4340 ± 0.0026	1.0317 ± 0.0003		img	0.2603 ± 0.0006	1.0113 ± 0.0001
HyCoCLIP	part text	0.1207 ± 0.0006	1.0024 ± 0.0000	UNCHA	part text	0.2010 ± 0.0000	1.0068 ± 0.0001
	text	0.1410 ± 0.0010	1.0033 ± 0.0001		text	0.2000 ± 0.0000	1.0067 ± 0.0000
	part img	0.2030 ± 0.0000	1.0069 ± 0.0000		part img	0.4133 ± 0.0006	1.0287 ± 0.0001
	img	0.2050 ± 0.0000	1.0070 ± 0.0000		img	0.4130 ± 0.0010	1.0287 ± 0.0001

Table 7: **Detailed pairwise geometric and cone validity analysis.**

Model	Modality	$\mathbf{G}_{\text{pos}}$	$\mathbf{G}_{\text{neg}}$	NNA@1	$\mathbf{S}_{\text{clamp}}$	$\mathbf{S}_{\text{final}}$
MERU	text → img	0.65 ± 0.00	0.65 ± 0.01	99.91 ± 0.02	90.18 ± 2.03	94.45 ± 0.15
HyCoCLIP	part text → part img	0.23 ± 0.00	0.22 ± 0.01	99.83 ± 0.06	100.00 ± 0.00	100.00 ± 0.00
	part text → img	0.23 ± 0.01	0.22 ± 0.01	99.88 ± 0.05	100.00 ± 0.00	100.00 ± 0.00
	text → part img	0.33 ± 0.01	0.33 ± 0.01	99.94 ± 0.01	100.00 ± 0.00	100.00 ± 0.00
	text → img	0.33 ± 0.01	0.33 ± 0.01	99.96 ± 0.02	100.00 ± 0.00	100.00 ± 0.00
PHyCLIP	part text → part img	0.28 ± 0.00	0.28 ± 0.01	99.84 ± 0.06	98.15 ± 0.16	99.86 ± 0.01
	part text → img	0.28 ± 0.01	0.27 ± 0.00	99.83 ± 0.00	98.15 ± 0.16	99.86 ± 0.01
	text → part img	0.39 ± 0.00	0.39 ± 0.00	99.84 ± 0.00	95.32 ± 0.17	99.54 ± 0.02
	text → img	0.39 ± 0.00	0.39 ± 0.00	99.95 ± 0.03	95.32 ± 0.17	99.54 ± 0.02
UNCHA	part text → part img	0.63 ± 0.00	0.60 ± 0.00	99.01 ± 0.31	88.90 ± 2.24	92.27 ± 0.22
	part text → img	0.63 ± 0.00	0.60 ± 0.00	99.04 ± 0.26	88.90 ± 2.24	92.27 ± 0.22
	text → part img	0.62 ± 0.00	0.61 ± 0.00	99.59 ± 0.02	75.03 ± 1.46	93.80 ± 0.61
	text → img	0.60 ± 0.00	0.60 ± 0.00	99.69 ± 0.03	75.03 ± 1.46	93.80 ± 0.61

Hence, decreasing $\|\tilde{\mathbf{p}}\|$ decreases $\mathcal{L}_{\text{ent}}$ along this radial path, provided that the hinge remains active. Approaching the validity boundary from the valid cone domain as $\|\tilde{\mathbf{p}}\| \to (2K/\sqrt{\kappa})^+$, we have

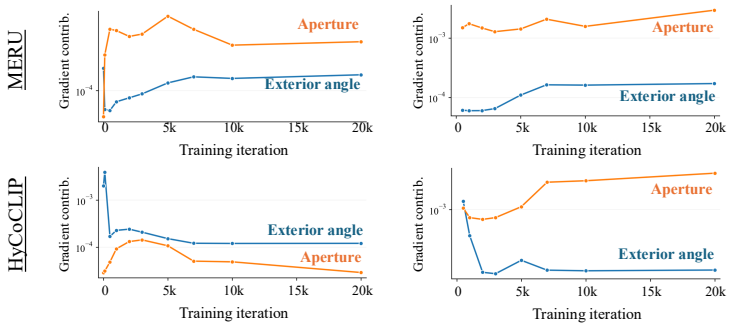
$$\frac{2K}{\|\tilde{\mathbf{p}}\|\sqrt{\kappa\|\tilde{\mathbf{p}}\|^2 - 4K^2}} \to +\infty, \qquad \omega(\mathbf{p}) \to \frac{\pi}{2}. \tag{32}$$

The divergence therefore concerns the radial slope rather than the aperture or loss value. For fixed $\mathbf{q}$ and fixed noncollinear directions, the exterior-angle derivative remains finite at the boundary. The hinge may nevertheless become inactive before reaching the boundary. Under the idealized clipped extension, $\omega(\mathbf{p}) = \pi/2$ for $\sqrt{\kappa}\|\tilde{\mathbf{p}}\| < 2K$, so the aperture derivative vanishes in the saturated region. The boundary divergence above pertains to the unclamped expression used for theoretical analysis—the argument of $\arcsin$ is clamped to $1 - \epsilon$ in the implementation. The analysis identifies a possible radial route to saturation rather than an inevitable outcome of joint optimization.

D Details on Geometric Analysis

D.1 Detailed Statistics for Geometric Analysis

We provide detailed results from repeated runs of the geometric analysis. All baselines are trained using three random seeds (0, 42, 2026) while keeping the training configuration of each baseline fixed across seeds. We compute the 99th percentile and additionally report the standard deviation across seeds. Tab. 6 presents detailed statistics for the point geometric analysis, while Tab. 7 reports those for the pairwise geometric and cone validity analysis. Notably, for PHyCLIP, we analyze 64 Lorentz factors using their respective curvatures. For factor-level diagnostics, rate-based statistics are averaged across factors, which is equivalent to computing the corresponding rates over all factor–sample pairs because each factor contains the same number of observations. Percentiles are computed from measurements pooled across factors within each seed, rather than by averaging factor-level percentiles. The pooled final-checkpoint analysis is performed after verifying that all factor curvatures coincide. For pairwise diagnostics, we preserve PHyCLIP's ℓ_1-product aggregation for both the



(a) All active parents (b) Valid-near active parents

Figure 5: **Entailment gradients during early training.** Mean exterior-angle and aperture contributions to the parent log-radius gradient for (a) all pairs with an active entailment hinge and (b) those whose parents satisfy $0 < m(\mathbf{p}) \leq 0.05$. The vertical axes are logarithmic.

hyperbolic distances and their tangent-space counterparts. We compute the pairwise metrics from the aggregated distances, rather than by pooling or averaging factor-level gaps or nearest-neighbor agreements.

D.2 Entailment Gradients During Early Training

Fig. 5 compares the exterior-angle and aperture contributions from the first $20{,}000$ training iterations. We focus on this interval to capture cone saturation and how the two contributions evolve. We report the mean contributions to $\partial \mathcal{L}_{\text{ent}} / \partial \log r(\mathbf{p})$ over parent–child pairs with an active entailment hinge, where $r(\mathbf{p}) = \sqrt{\kappa}\, d_{\mathbb{H}}(o, \mathbf{p})$. For each pair, the child embedding, curvature, and parent radial direction are held fixed. A positive contribution indicates that reducing the parent radius locally decreases the entailment loss through that term. The mean is taken over all such pairs across the entailment relations of each model, retaining the loss weights and reduction factors used in training. Fig. 5(a) includes all active pairs, regardless of cone validity. We define the normalized cone-domain margin as $m(\mathbf{p}) = \sqrt{\kappa}\|\tilde{\mathbf{p}}\|/(2K) - 1$. For Fig. 5(b), we retain only active pairs whose parents satisfy $0 < m(\mathbf{p}) \leq 0.05$, corresponding to spatial norms within 5% above the minimum admissible norm.

In Fig. 5(a), the mean aperture contribution is larger in MERU after initialization, whereas the mean exterior-angle contribution remains larger in HyCoCLIP at the evaluated checkpoints. In Fig. 5(b), the mean aperture contribution exceeds the exterior-angle contribution in both models at the evaluated checkpoints from $1{,}000$ iterations onward. Thus, in HyCoCLIP, the aperture contribution is larger near the cone-domain boundary even though it is smaller when averaged over all active pairs. These results are consistent with Eq. 10—near the admissible boundary, the aperture term can contribute more strongly to the parent-radial gradient than the exterior-angle term. In the near-boundary subset, a small radial contraction therefore reduces the active penalty, on average, more through cone widening than through the accompanying change in exterior angle. This relaxes the entailment constraint without moving the child embedding or changing either embedding's radial direction. The curves report loss gradients evaluated at the saved checkpoints, not the magnitude of actual optimizer updates.

D.3 Saturation Dynamics Throughout Training

Figure 6 shows that cone saturation is not a transient phenomenon confined to S_{clamp}, but remains prevalent throughout the subsequent training process. For all relations in Figs. 6(a)–(e), the saturation ratio is already high at S_{clamp} or increases further toward the final checkpoint. HyCoCLIP remains fully saturated at 100%, while PHyCLIP approaches nearly complete saturation for most relations. MERU also maintains a saturation ratio above 90% where applicable. Although UNCHA begins with a comparatively lower ratio of 75.03% for the text-to-image relations, its saturation ratio steadily increases to 93.80%; for the relations involving part text, it remains consistently high at approximately 89%–93%. These results indicate that the cone aperture stays at its clipped upper bound during most of the remaining optimization for the majority of relation instances. Consequently, the

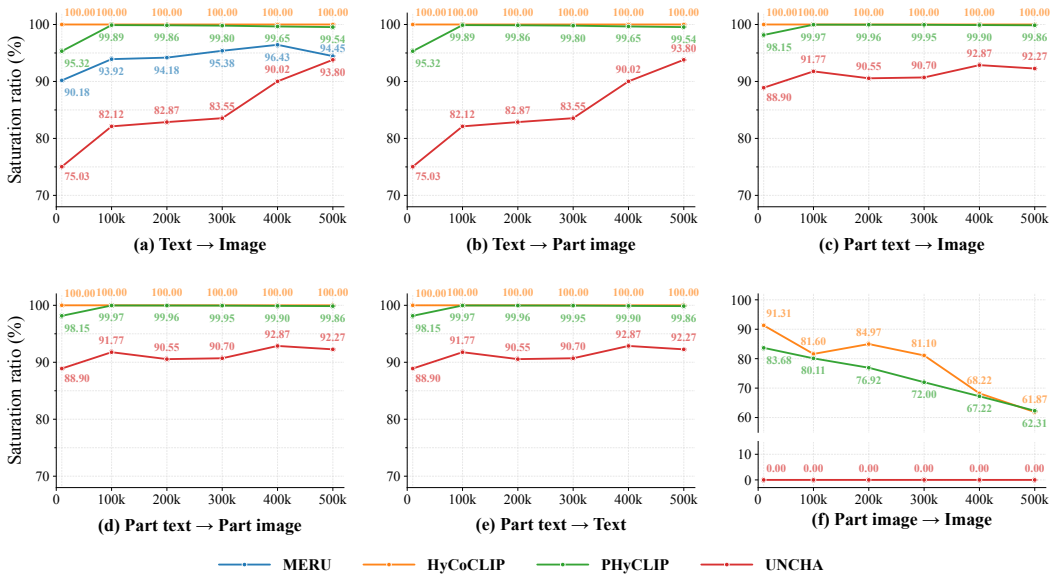


Figure 6: **Saturation dynamics throughout training.** We track how the entailment cone saturation ratio evolves for each hierarchical relation from S_{clamp} to the final training step at 500K.

aperture becomes insensitive to changes in the embedding norm and no longer supplies the norm-dependent gradient intended to organize embeddings according to their hierarchical levels. Although the full entailment objective still produces gradients through the exterior-angle term when its hinge remains active, its aperture-based geometric mechanism is largely inactive for saturated pairs.

The only clear exception is the part-image-to-image relation in Fig. 6(f), for which UNCHA exhibits a saturation ratio of 0% throughout training. In this case, the embeddings associated with the cone aperture maintain sufficiently large radial norms to remain outside the clipping region, allowing the aperture to retain its dependence on the embedding radius. HyCoCLIP and PHyCLIP also show decreasing saturation for this relation, from 91.31% and 83.68% at S_{clamp} to 61.87% and 62.31% at 500K, respectively. Nevertheless, more than 60% of the corresponding relations remain saturated even at the final checkpoint. Overall, the observed dynamics suggest that cone saturation governs most hierarchical relations during training, substantially limiting the extent to which the entailment objective can exploit its intended norm-dependent cone geometry.

E Details on Downstream Experiments

E.1 Benchmarks

Zero-Shot Image Classification. We evaluate zero-shot recognition on 16 benchmarks: ImageNet (Deng et al., 2009), CIFAR-10 (Krizhevsky, 2009), 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 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), RESISC45 (Cheng et al., 2017), and Country211 (Radford et al., 2021; Desai et al., 2023). For each class label, we construct multiple text prompts using dataset-specific templates, encode them with the text encoder, and average the resulting embeddings to obtain the class representation. Each image is assigned to the label with the highest image–text similarity. We report class-averaged top-1 accuracy for each dataset and use the unweighted mean across the 16 benchmarks as the overall classification score.

Zero-Shot Cross-Modal Retrieval. We evaluate image–text retrieval on the COCO (Lin et al., 2014) and Flickr-30K (Young et al., 2014; Karpathy & Fei-Fei, 2015). For text retrieval, each image is used to rank all candidate captions, whereas image retrieval ranks all candidate images for a text

query. A query is counted as correct at rank K when at least one ground-truth paired item appears among the top-K results. Recall@1, Recall@5, and Recall@10 are computed in both directions and averaged over all queries.

Hierarchical Classification. Following prior work (Pal et al., 2025; Kim et al., 2026), we evaluate hierarchical classification using the WordNet (Miller, 1995) hierarchy of ImageNet (Deng et al., 2009) labels. Given a predicted class $\hat{y}$ and its ground-truth class y, let $E(i, j)$ denote the edges on the shortest path connecting nodes i and j in the WordNet graph. We denote by $\mathrm{lca}(i, j)$ their lowest common ancestor, by $d_G(i, j)$ the shortest-path distance in the graph, and by w_e the weight of edge e, with $w_e = 1$ for all WordNet edges. Tree-Induced Error (TIE) measures the complete graph distance between the two labels, whereas Lowest Common Ancestor (LCA) error measures the distance from the predicted label to the deepest ancestor shared with the ground-truth label as

$$\mathrm{TIE}(\hat{y}, y) = \sum_{e \in E(\hat{y}, y)} w_e, \qquad \mathrm{LCA}(\hat{y}, y) = d_G(\hat{y}, \mathrm{lca}(\hat{y}, y)) \,. \tag{33}$$

Let $\hat{Y}_{\mathrm{anc}}$ and Y_{anc} denote the ancestor sets of $\hat{y}$ and y, respectively. We compute Jaccard similarity J, hierarchical precision P_H, and hierarchical recall R_H (Kosmopoulos et al., 2015) as

$$J = \frac{|\hat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|\hat{Y}_{\mathrm{anc}} \cup Y_{\mathrm{anc}}|}, \qquad P_H = \frac{|\hat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|\hat{Y}_{\mathrm{anc}}|}, \qquad R_H = \frac{|\hat{Y}_{\mathrm{anc}} \cap Y_{\mathrm{anc}}|}{|Y_{\mathrm{anc}}|}. \tag{34}$$

Lower TIE and LCA indicate smaller hierarchical errors, whereas higher J, P_H, and R_H indicate greater agreement between the predicted and ground-truth ancestor sets.

Scene Understanding. We evaluate compositional scene understanding using VL-Checklist-Object (VL-CO) (Zhao et al., 2022) and VG-Attribution (VG-A) (Yüksekgönül et al., 2023). Both benchmarks present an image with a correct caption and a negative caption, and count a prediction as correct when the model assigns greater image–text similarity to the correct caption.

- **VL-CO** evaluates whether the model correctly identifies objects under variations in their size and image location. Negative captions are generated by replacing an object expression in the original caption with an unrelated noun phrase. Using the object bounding-box area A, O-Small (O-S), O-Medium (O-M), and O-Large (O-L) correspond to $A < 32 \times 32$, $32 \times 32 \leq A \leq 96 \times 96$, and $A > 96 \times 96$, respectively. For location-based subsets, let $r = d/h$, where d is the distance between the object and image centers and h is half of the image diagonal. O-Center, O-Mid, and O-Margin correspond to $r \leq 1/3$, $1/3 < r \leq 2/3$, and $r > 2/3$, respectively.
- **VG-A** evaluates whether the model correctly associates attributes with their corresponding objects in multi-object scenes. Each hard negative exchanges the attributes assigned to two objects while retaining the same objects and attribute words. The task therefore requires correct binding between attribute and object rather than recognition of the individual concepts alone. We report the average pairwise accuracy, defined as the proportion of examples for which the model assigns higher image–text similarity to the correct caption than to the corresponding hard negative with exchanged attributes.

E.2 Detailed Statistics for Downstream Tasks

We provide detailed results from repeated runs of the downstream experiments. All baselines are trained under identical settings using three random seeds (0, 42, 2026). For each evaluation metric, we report the mean and standard deviation across the seeds. Tab. 8 presents detailed statistics for the zero-shot image classification, while Tab. 9 reports those for the zero-shot retrieval. Tab. 10 presents detailed statistics for the hierarchical evaluation. Further, Tab. 11 reports detailed results for the scene understanding task.

Table 8: **Detailed zero-shot image classification evaluation.**

Model	General datasets						Fine-grained datasets						Misc. datasets			
	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
MERU	37.4 ± 0.4	66.6 ± 2.1	38.9 ± 2.6	48.8 ± 0.6	72.2 ± 0.4	92.4 ± 0.7	47.6 ± 1.9	10.7 ± 0.6	6.3 ± 0.2	2.5 ± 0.2	43.9 ± 0.7	19.0 ± 0.9	22.3 ± 1.3	31.4 ± 5.6	43.2 ± 1.1	4.5 ± 0.2
HyCoCLIP	43.6 ± 0.7	87.6 ± 0.5	55.9 ± 0.7	52.9 ± 0.4	78.4 ± 1.5	94.7 ± 0.2	54.8 ± 0.9	14.8 ± 0.7	10.6 ± 0.2	3.0 ± 0.4	53.1 ± 1.9	24.6 ± 1.3	24.4 ± 1.3	29.0 ± 4.3	47.1 ± 0.2	4.9 ± 0.2
PHyCLIP	44.3 ± 0.7	88.7 ± 0.9	59.7 ± 0.8	54.7 ± 0.5	77.9 ± 0.3	94.8 ± 0.3	57.4 ± 1.3	**17.2** ± 0.4	10.6 ± 0.6	3.7 ± 0.5	55.4 ± 4.0	24.8 ± 2.2	25.8 ± 1.7	**41.2** ± 3.0	47.9 ± 1.2	5.1 ± 0.3
UNCHA	49.3 ± 0.4	**90.1** ± 0.5	**63.8** ± 0.4	56.9 ± 0.3	83.0 ± 0.8	**96.3** ± 0.2	56.3 ± 2.3	15.4 ± 2.0	**13.9** ± 0.5	**4.2** ± 0.8	57.2 ± 3.4	**26.6** ± 1.3	30.6 ± 1.3	30.5 ± 2.3	50.1 ± 0.3	6.0 ± 0.2
CLIP	37.3 ± 1.0	64.6 ± 2.2	41.5 ± 0.7	49.0 ± 1.1	73.9 ± 1.0	92.2 ± 0.8	47.9 ± 1.9	11.4 ± 1.5	6.2 ± 0.4	2.6 ± 0.7	42.1 ± 0.9	21.5 ± 1.4	22.9 ± 0.8	33.0 ± 4.8	43.4 ± 1.1	4.4 ± 0.4
w/ 1-part	43.8 ± 0.7	89.2 ± 0.2	59.5 ± 0.7	53.4 ± 0.3	80.5 ± 0.6	95.6 ± 0.3	51.9 ± 1.4	14.1 ± 1.7	11.4 ± 1.4	2.8 ± 0.5	53.5 ± 1.7	23.3 ± 2.5	26.6 ± 1.9	29.3 ± 3.4	48.2 ± 1.9	5.1 ± 0.1
w/ 5-part	44.8 ± 0.4	89.7 ± 1.4	61.3 ± 0.5	54.7 ± 0.8	81.6 ± 0.4	96.0 ± 0.2	55.9 ± 1.1	13.5 ± 1.1	11.2 ± 0.9	2.8 ± 0.5	54.5 ± 2.7	21.6 ± 5.3	28.3 ± 1.4	30.8 ± 2.5	47.5 ± 0.8	5.3 ± 0.1
EuHier (Ours)	**49.5** ± 0.4	89.9 ± 0.7	62.3 ± 0.1	**57.7** ± 0.3	**83.8** ± 0.5	95.9 ± 0.4	**59.3** ± 1.0	16.1 ± 1.4	13.5 ± 1.3	3.8 ± 0.2	**59.1** ± 1.8	26.1 ± 2.3	**32.6** ± 0.4	27.8 ± 1.2	**50.5** ± 1.0	**6.5** ± 0.2

Table 9: **Detailed zero-shot retrieval evaluation.**

Model	Text retrieval						Image retrieval					
	COCO			Flickr			COCO			Flickr		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MERU	40.1 ± 0.2	67.5 ± 0.3	78.4 ± 0.2	68.0 ± 1.4	91.0 ± 0.5	95.0 ± 0.7	29.2 ± 0.5	55.5 ± 0.5	66.9 ± 0.2	55.6 ± 0.3	81.1 ± 0.3	88.2 ± 0.3
HyCoCLIP	41.7 ± 0.6	68.4 ± 0.3	78.7 ± 0.3	69.9 ± 0.9	91.5 ± 1.0	95.4 ± 0.6	30.2 ± 0.2	56.3 ± 0.2	67.6 ± 0.2	57.3 ± 0.8	82.0 ± 0.4	88.9 ± 0.5
PHyCLIP	42.8 ± 0.3	69.8 ± 0.4	79.9 ± 0.6	71.3 ± 1.3	90.9 ± 0.1	95.0 ± 0.6	31.2 ± 0.3	57.0 ± 0.4	68.1 ± 0.3	58.4 ± 0.9	82.7 ± 0.1	89.2 ± 0.1
UNCHA	46.8 ± 0.2	73.8 ± 0.7	82.5 ± 0.1	75.1 ± 1.3	93.8 ± 0.8	96.8 ± 0.3	34.4 ± 0.0	60.9 ± 0.1	71.8 ± 0.1	63.4 ± 0.5	86.5 ± 0.4	92.2 ± 0.4
CLIP	41.7 ± 0.6	69.2 ± 1.0	79.1 ± 0.4	70.6 ± 2.4	91.8 ± 0.8	96.1 ± 0.5	29.3 ± 0.2	55.6 ± 0.5	67.3 ± 0.5	56.7 ± 0.8	82.5 ± 0.3	89.3 ± 0.5
w/ 1-part	43.0 ± 0.4	70.6 ± 0.8	80.4 ± 0.9	71.9 ± 1.5	91.8 ± 0.9	95.8 ± 0.4	30.6 ± 0.4	57.1 ± 0.7	68.3 ± 0.5	58.7 ± 0.5	83.4 ± 1.0	89.8 ± 0.3
w/ 5-part	44.3 ± 0.6	71.1 ± 0.4	80.7 ± 0.5	74.1 ± 0.7	92.8 ± 0.6	96.9 ± 0.4	31.4 ± 0.3	57.9 ± 0.2	69.4 ± 0.3	60.1 ± 0.5	84.9 ± 0.3	90.9 ± 0.2
EuHier (Ours)	**48.2** ± 0.7	**75.0** ± 0.5	**83.8** ± 0.3	**76.6** ± 0.6	**94.7** ± 0.5	**97.6** ± 0.8	**35.7** ± 0.3	**62.4** ± 0.2	**72.9** ± 0.2	**65.8** ± 0.4	**88.1** ± 0.4	**93.3** ± 0.4

Table 10: **Detailed hierarchical evaluation.**

Model	TIE↓	LCA↓	J↑	$\mathbf{P_H}$ ↑	$\mathbf{R_H}$ ↑
MERU	3.801 ± 0.038	2.272 ± 0.020	0.774 ± 0.002	0.847 ± 0.002	0.846 ± 0.002
HyCoCLIP	3.305 ± 0.051	2.094 ± 0.027	0.806 ± 0.004	0.867 ± 0.003	0.868 ± 0.002
PHyCLIP	3.667 ± 0.047	2.203 ± 0.025	0.782 ± 0.002	0.853 ± 0.002	0.848 ± 0.001
UNCHA	2.899 ± 0.019	**1.936** ± 0.006	0.831 ± 0.001	**0.885** ± 0.001	0.885 ± 0.001
CLIP	3.823 ± 0.068	2.298 ± 0.021	0.774 ± 0.005	0.845 ± 0.003	0.846 ± 0.004
w/ 1-part	3.317 ± 0.035	2.096 ± 0.013	0.804 ± 0.003	0.867 ± 0.002	0.866 ± 0.002
w/ 5-part	3.227 ± 0.040	2.071 ± 0.019	0.811 ± 0.003	0.871 ± 0.002	0.872 ± 0.002
EuHier (Ours)	**2.884** ± 0.013	1.950 ± 0.010	**0.833** ± 0.001	**0.885** ± 0.001	**0.887** ± 0.001

Table 11: **Detailed scene understanding evaluation.**

Model	VL-CO	O-Large	O-Medium	O-Small	O-Center	O-Mid	O-Margin	VG-A
MERU	76.6 ± 0.6	84.1 ± 0.5	74.3 ± 0.3	67.2 ± 1.5	83.6 ± 0.3	79.7 ± 1.1	70.7 ± 1.6	58.5 ± 0.9
HyCoCLIP	78.9 ± 0.6	86.5 ± 0.3	77.3 ± 1.1	67.6 ± 1.8	85.7 ± 0.5	81.8 ± 0.4	74.5 ± 0.8	61.5 ± 0.5
PHyCLIP	80.0 ± 0.3	87.1 ± 0.1	78.4 ± 0.2	70.8 ± 1.4	86.1 ± 0.2	83.0 ± 0.3	74.9 ± 0.2	64.3 ± 1.3
UNCHA	81.0 ± 0.4	88.0 ± 0.3	79.8 ± 0.1	71.1 ± 0.8	87.0 ± 0.1	83.7 ± 0.5	76.3 ± 0.8	65.0 ± 0.1
CLIP	77.6 ± 0.9	84.4 ± 0.3	75.3 ± 0.9	69.6 ± 2.9	83.9 ± 0.4	79.7 ± 0.8	72.4 ± 0.9	59.3 ± 1.1
w/ 1-part	77.6 ± 0.2	85.1 ± 0.3	75.9 ± 0.4	68.2 ± 1.5	84.9 ± 0.3	80.2 ± 0.6	71.5 ± 0.3	59.7 ± 0.4
w/ 5-part	78.1 ± 1.1	85.0 ± 1.0	76.3 ± 0.4	69.7 ± 1.9	84.3 ± 1.0	80.5 ± 1.4	72.7 ± 1.6	60.0 ± 1.4
EuHier (Ours)	**82.5** ± 0.1	**88.9** ± 0.1	**82.0** ± 0.5	**71.9** ± 1.0	**87.8** ± 0.5	**85.1** ± 0.1	**79.6** ± 0.6	**69.3** ± 0.9

F Further Experiments

F.1 Efficiency Experiments

Settings. We collected all efficiency measurements on a server equipped with four NVIDIA H200 GPUs and dual Intel Xeon Platinum 8462Y+ processors, using an allocation of 60 CPU core-equivalents and 900 GiB of host memory. To summarize training step time, we excluded the warm-up iterations and partitioned a single run of each model into 19 consecutive blocks of 500 iterations.

Table 12: **Computational efficiency comparison.** Training step time is measured during multi-GPU training, while post-encoder time and memory are measured separately on a single GPU excluding encoders and collective communication.

Method	Full Training		Isolated Post-Encoder		
	Training step time ↓ (ms/step)	Peak GPU memory ↓ (GB)	Forward time ↓ (ms)	Backward time ↓ (ms)	GPU memory ↓ (MB)
UNCHA	402.5 ± 3.5	52.63	15.12 ± 0.04	22.03 ± 0.54	96.4
EuHier-H	389.8 ± 2.1	**51.62**	9.49 ± 0.04	16.58 ± 0.53	54.9
EuHier	$\mathbf{387.4 \pm 3.6}$	**51.62**	$\mathbf{4.23 \pm 0.02}$	$\mathbf{6.38 \pm 0.02}$	**25.7**

We computed the mean iteration time within each block and reported the standard deviation across the 19 block-level means. To summarize post-encoder time, we first computed the median across repeated measurements for each of 32 saved iteration–rank cases and then reported the mean and standard deviation across the resulting case-level medians. The reported standard deviations therefore characterize within-run measurement variation rather than variability across independent runs.

We used the following metrics to compare computational efficiency.

- **Training step time** measures the wall-clock time required for one training iteration, including data loading, host-to-device transfer, encoder computation, loss computation, backpropagation, optimizer updates, and distributed communication.
- **Peak training GPU memory** is the maximum of the per-GPU peak allocated memory recorded during the training measurement interval across all GPUs.
- **Post-encoder forward time** measures the time required to transform the encoder outputs and compute similarity scores and losses, while excluding encoder computation and collective communication.
- **Post-encoder backward time** measures the time required to compute gradients with respect to the encoder outputs and post-encoder parameters, while excluding backpropagation through the encoders and collective communication.
- **Post-encoder GPU memory** estimates the additional peak allocated GPU memory attributable to the post-encoder computation in an isolated benchmark that excludes the encoders and NCCL workspaces.

Computational Efficiency Comparison. Table 12 compares the computational efficiency of UNCHA, EuHier-H, and EuHier under matched input workloads. EuHier-H, which retains the architecture and hierarchical contrastive supervision of EuHier while operating in Lorentz space, is already substantially more efficient than UNCHA in the isolated post-encoder benchmark. The comparison shows that substantial post-encoder savings are possible with a simpler hierarchy-aware contrastive objective, even within Lorentz space. Replacing the exponential map and geodesic distance in EuHier-H with Euclidean operations yields a further pronounced reduction—EuHier requires 55.4% less forward time, 61.5% less backward time, and 53.2% less additional post-encoder memory than EuHier-H. The corresponding difference in full training step time is much smaller, consistent with the substantial cost of the remaining training pipeline, including encoder computation, optimizer updates, data transfer, and distributed communication. Overall, the results demonstrate that EuHier reduces post-encoder computational costs by avoiding the hyperbolic computation path while retaining the whole-part contrastive supervision used by EuHier-H.

G Implementation Details

Architecture. EuHier uses a ViT-B/16 image encoder with 224×224 inputs and a 12-layer Transformer text encoder with a hidden width of 512 and a maximum context length of 77 tokens. Modality-specific linear projections map the encoder outputs into a shared 512-dimensional embedding space, followed by ℓ_2 normalization. Within each modality, the encoder and projection layer are shared between whole and part inputs.

Optimization. The six relation losses are summed with unit weights and share a learnable temperature initialized to $\tau = 0.07$, with the inverse temperature capped at 100. We train EuHier from scratch for $500{,}000$ iterations on four NVIDIA H200 GPUs with a global batch size of 768 whole image–text pairs, each accompanied by up to five part pairs. We use AdamW (Loshchilov & Hutter, 2019) with a peak learning rate of 5×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.98)$, and a weight decay of 0.2, excluding biases, LayerNorm parameters, and the learnable logit scale from weight decay. The learning rate increases linearly during the first $4{,}000$ iterations and then decays to zero following a cosine schedule (Loshchilov & Hutter, 2017).

Implementation of EuHier-H. EuHier-H is a Lorentz-space counterpart of EuHier that retains the same image and text encoder architecture, linear projection heads, and six contrastive relations with unit loss weights. Instead of applying ℓ_2 normalization, we map each unnormalized projected feature $\mathbf{z}_m$ onto the Lorentz hyperboloid through the origin exponential map, $\mathbf{x}_m = \exp_o^\kappa(\alpha_m \mathbf{z}_m)$, where $m \in \{I, T\}$ denotes the modality. The learnable modality scales α_m are shared between whole and part inputs, initialized to $1/\sqrt{d}$ for embedding dimension d, and capped at 1. All representations share a learnable curvature magnitude κ, initialized to 1 and constrained to $[0.1, 10]$. Both the curvature magnitude and modality scales are parameterized in log space. We replace cosine similarity with negative Lorentz geodesic distance, yielding contrastive logits $-d_{\mathbb{H}_\kappa}(\mathbf{x}, \mathbf{y})/\tau$. All six losses share a learnable temperature initialized to $\tau = 0.07$, with the inverse temperature capped at 100. As in EuHier, the whole image–text and part image–text losses are bidirectional, whereas the four part-to-whole losses use each part as a query and its corresponding whole representation as the positive target. No entailment loss, entailment cones, or minimum-radius projection is used.

H LIMITATION

Our work focuses on representative hyperbolic VLMs evaluated under a matched training protocol and data setting. The observed near-Euclidean representations and entailment cone saturation characterize the evaluated models and training conditions, and do not rule out the possibility that hyperbolic geometry may provide practical benefits under alternative architectures, objectives, data scales, or hierarchical structures. Hyperbolic space retains established theoretical advantages for representing tree-like structures with low distortion, and identifying the learning conditions that translate these geometric properties into practical gains for VLMs remains an important direction for future work. In addition, EuHier demonstrates that whole–part hierarchical supervision can remain effective without hyperbolic parameterization or entailment cones, but its contrastive formulation does not explicitly enforce directional parent–child orderings or transitive partial-order structure in the embedding space. Future work should evaluate these capabilities more directly and clarify when hierarchy-aware supervision alone is sufficient and when explicit geometric ordering provides additional value.