PHyCLIP: ℓ_1-Product of Hyperbolic Factors Unifies Hierarchy and Compositionality in Vision–Language Representation Learning

Anonymous authors
Paper under double-blind review

Abstract

Vision–language models have achieved remarkable success in multi-modal representation learning from large-scale pairs of visual scenes and linguistic descriptions. However, they still struggle to simultaneously express two distinct types of semantic structures: the hierarchy within a concept family (e.g., *dog* $\preceq$ *mammal* $\preceq$ *animal*) and the compositionality across different concept families (e.g., "a dog in a car" $\preceq$ *dog*, *car*). Recent works have addressed this challenge by employing hyperbolic space, which efficiently captures tree-like hierarchy, yet its suitability for representing compositionality remains unclear. To resolve this dilemma, we propose *PHyCLIP*, which employs an ℓ_1-*P*roduct metric on a Cartesian product of *Hy*perbolic factors. With our design, intra-family hierarchies emerge within individual hyperbolic factors, and cross-family composition is captured by the ℓ_1-product metric, analogous to a Boolean algebra. Experiments on zero-shot classification, retrieval, hierarchical classification, and compositional understanding tasks demonstrate that PHyCLIP outperforms existing single-space approaches and offers more interpretable structures in the embedding space.

1 Introduction

Vision–language models have become a central paradigm for learning transferable representations across visual and textual modalities. As exemplified by CLIP (Radford et al., 2021), contrastive pretraining maps images and texts to embeddings and enables strong zero-shot transfer on classification, retrieval, and related tasks. However, compressing the semantics of an instance into a single point makes it challenging to faithfully encode two semantic structures at once: *hierarchy* (*is-a* relations in a concept family) and *compositionality* (conjunction across distinct concept families).

Visual and linguistic concepts linked by *is-a* relations form tree-like taxonomic *hierarchies*. For example, a dog *is a* mammal, which in turn *is an* animal, as shown in the upper part of Fig. 1. Because the number of nodes grows exponentially with depth, Euclidean geometry struggles to faithfully represent such trees, whereas hyperbolic geometry aligns well with this growth (Bridson & Haefliger, 1999; Sarkar, 2011). These observations have motivated the development of hyperbolic embeddings (Nickel & Kiela, 2017) and hyperbolic entailment cones, which encode partial orders via inclusion (Ganea et al., 2018a). Within vision–language representation learning, MERU (Desai et al., 2023) and HyCoCLIP (Pal et al., 2025) leverage these approaches to capture image–text relations; for instance, an image of a dog *is an* instance of the linguistic concept dog (see the lower part of Fig. 1).

Beyond taxonomic structure, images and texts exhibit *compositionality*. For example, the description "a dog in a car" binds concepts dog and car from distinct concept families (animals and transportation), as shown in the middle part of Fig. 1. Classical approaches express composition via logical conjunction or additive operations (e.g., Boolean algebra, bag-of-words, and vector addition in word2vec) (Hinton et al., 1986; Mikolov et al., 2013; Vendrov et al., 2016), but these struggle to encode semantic hierarchy efficiently. Conversely, while hyperbolic geometry captures hierarchy, it lacks a canonical operation for composition. Möbius addition in hyperbolic spaces (Ungar, 2008) is not aligned with standard vector addition or Boolean structures (Higgins et al., 2018). Intersections of regions (such as hyperbolic entailment cones) can approximate conjunction but offer no general guarantees of representational efficiency for arbitrary co-occurrences.

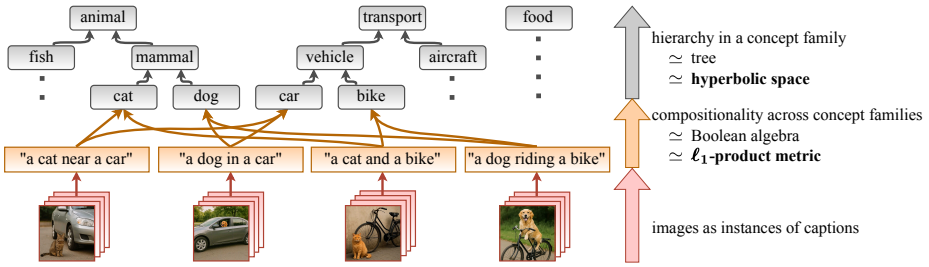


Figure 1: **Conceptual diagram of hierarchical and compositional structures.** While all arrows represent entailments ($\preceq$), they differ in nature. (upper) Linguistic concepts organize tree-like taxonomic *hierarchies* of concept families, each of which can be embedded into a hyperbolic space (Nickel & Kiela, 2017). (middle) Images and texts exhibit *compositionality* across distinct concept families, which can be captured by a Boolean algebra or an ℓ_1-product metric. (lower) Images are instances of their corresponding captions.

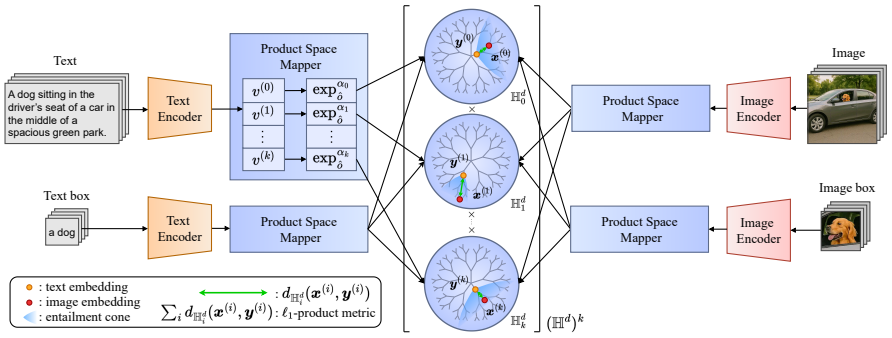


Figure 2: **Overview of PHyCLIP.** Images and texts are encoded as points $\boldsymbol{X}$ in an ℓ_1-product metric space of hyperbolic factors, $(\mathbb{H}^d)^k$, that is, as tuples of points $\boldsymbol{x}^{(i)}$ in hyperbolic spaces $\mathbb{H}^d_i$, where their distance is defined by the sum of hyperbolic distances. The entailment relations $\boldsymbol{X} \preceq \boldsymbol{Y}$ are encoded using entailment cones as $\boldsymbol{x}^{(i)} \in C(\boldsymbol{y}^{(i)})$ within hyperbolic factors $\mathbb{H}^d_i$.

To resolve this dilemma, we propose *PHyCLIP*, which leverages an ℓ_1-*P*roduct metric on a Cartesian product of *Hy*perbolic factors, as depicted in Fig. 2. Our design follows two classical correspondences: (i) metric trees admit low-distortion embeddings into hyperbolic spaces, so hyperbolic factors embed *intra-family* taxonomies (Sarkar, 2011; Nickel & Kiela, 2017; Ganea et al., 2018a); and (ii) finite Boolean algebras with the Hamming distance embed isometrically into an ℓ_1 space, so an ℓ_1-product metric naturally supports *cross-family* Boolean-like composition (Deza & Laurent, 1997). Intuitively, each bit for an atomic concept (e.g., dog, cat, horse) in the Boolean algebra is replaced with a hyperbolic factor for a concept family (e.g., animals), and the activation of multiple factors expresses composition (e.g., "dog and car"). Unlike previous mixed-curvature models (Gu et al., 2019; Wang et al., 2024; Gao et al., 2025), our space uses an ℓ_1-product metric rather than a Riemannian (ℓ_2) product metric and constrains each factor to have negative curvature. Our contributions are summarized as follows.

Balancing Hierarchy and Compositionality. We introduce PHyCLIP, a vision–language model that leverages an ℓ_1-product metric space of hyperbolic factors to jointly capture *hierarchy* (within factors) and *compositionality* (across factors).

Theoretical Support. We formally link Boolean lattices to ℓ_1-product metrics and metric trees to hyperbolic factors, explaining that an ℓ_1-product metric space of hyperbolic factors aligns better with the dual semantic structures than standard metric spaces (e.g., Euclidean or hyperbolic).

Superior Performance and Interpretability. Experiments on zero-shot classification, image–text retrieval, hierarchical classification, and compositional understanding demonstrate that PHyCLIP achieves consistent improvements over baselines that use standard metric spaces. Visualizations find that intra-family taxonomies emerge within individual factors, and composing concepts leads to the simultaneous activation of the corresponding factors, analogous to a Boolean algebra.

2 THEORETICAL BACKGROUND AND MOTIVATION

Geometry and Embedding of Hierarchies. Concepts in natural language linked by *is-a* (hypernymy/hyponymy, generalization/specialization, entailment) relations form a partially ordered set (poset) and typically exhibit deep hierarchical structure. A poset $(P, \preceq)$ is a set equipped with an order relation $\preceq$ (which is reflexive, antisymmetric, and transitive). A typical example is dog $\preceq$ mammal $\preceq$ animal, where a dog *is a* type of mammal; equivalently, if an entity is a dog, then this *entails* that the entity is a mammal. Large lexical resources such as WordNet provide such relations in the form of a directed acyclic graph with multiple inheritance (e.g., dog $\preceq$ domestic animal) (Miller, 1995). For modeling or computational convenience, many studies approximate this hierarchy with a taxonomic tree (Morin & Bengio, 2005; Mnih & Hinton, 2008). The distance between two nodes (i.e., words) in a tree is often defined as the length of their shortest path, inducing a type of *metric tree*. See technical details in Appendix A.

Theorem 1 (Hyperbolic embedding of trees (Sarkar, 2011))**.** *Let $\mathbb{H}^d$ be a d-dimensional hyperbolic space with the hyperbolic distance $d_{\mathbb{H}^d}$. For every finite metric tree T (and every infinite metric tree T with known bounds for maximum degree and minimum edge length), and for every $\varepsilon > 0$, there exist a scale $\tau > 0$ and an embedding $f : \tau T \to \mathbb{H}^2$ such that the distortion is at most $1 + \varepsilon$; that is, there exists a $(1 + \varepsilon, 0)$-quasi-isometric embedding f up to scaling.*

See Theorem 5 in Sarkar (2011) for the proof. This explains the empirical success of hyperbolic embeddings for hierarchical data (Nickel & Kiela, 2017; 2018; Ganea et al., 2018a; Sala et al., 2018; Tifrea et al., 2019). In practice, $\mathbb{H}^d$ with $d > 2$ is common for achieving better performance.

Geometry and Embedding of Compositionality. Beyond taxonomic structure, images and texts often exhibit compositionality: they mention multiple concepts to indicate the co-occurrence or conjunction of those concepts. For example, the description "a dog in a car" mentions concepts dog and car. Such data suggest another type of entailment relation, as an image of "a dog in a car" can be regarded as an image of dog as well as an image of car. The resulting structure is no longer a tree but rather a more general poset. While hyperbolic embeddings remain an option, it is natural to explore alternatives that more directly capture compositionality.

Order embeddings $(\mathbb{R}^n, \preceq)$ (Vendrov et al., 2016) assign each concept a point $\boldsymbol{x} \in \mathbb{R}^n$ and declare $\boldsymbol{x} \preceq \boldsymbol{y}$ iff $x_i \geq y_i$ for all coordinates i. This is equivalent to the inclusion relation between associated upper orthants $U(\boldsymbol{x}) := \{\boldsymbol{z} \in \mathbb{R}^n : \ z_i \geq x_i \ \forall i\}$, i.e., $\boldsymbol{x} \preceq \boldsymbol{y}$ iff $U(\boldsymbol{x}) \subseteq U(\boldsymbol{y})$. Then, the coordinate-wise $\max$ (i.e., the union of orthants) expresses conjunction (e.g., $\max($dog, car$)$ includes "a dog in a car"), and the coordinate-wise $\min$ yields shared concepts (e.g., $\min($"a dog in a car", "a dog on a sofa"$) \approx$ dog). Similarly, box embeddings use axis-aligned hyperrectangles in $\mathbb{R}^n$ (Vilnis et al., 2018; Li et al., 2019; Dasgupta et al., 2020). In hyperbolic space, hyperbolic entailment cones use geodesic conical regions (Ganea et al., 2018a), and disk embeddings use hyperballs (Suzuki et al., 2019). Compared with hyperbolic embeddings for pure hierarchies, there has been less theoretical analysis of these region-based embeddings for compositionality. Our work extends this line to capture hierarchy and compositionality simultaneously.

Boolean Lattice and Its Relation to Order Embedding. In an *is-a* taxonomy, any two nodes have at least one common generalization, whereas they need not share a common specialization. A *lattice* is a poset in which any two nodes have both a common generalization (join) and a common specialization (meet). Consider n atomic concepts $\mathcal{C} = \{c_1, \ldots, c_n\}$ (e.g., $\{$dog, car, tomato, $\ldots\}$). A subset $S \subseteq \mathcal{C}$ denotes the conjunction of the concepts specified in S, and the inclusion relation $S \supseteq T$ (e.g., $\{$dog, car$\} \supseteq \{$dog$\}, \{$car$\}$) induces the order relation $S \preceq T$ (e.g., $\{$dog, car$\} \preceq \{$dog$\}, \{$car$\}$). In this way, the *Boolean lattice* $(2^{\mathcal{C}}, \subseteq)$ over all such subsets naturally represents the compositionality of atomic concepts as a non-taxonomic poset. When focusing on operations rather than order, it is also referred to as a Boolean algebra. At the same time, using an indicator $\chi : 2^{\mathcal{C}} \to \{0, 1\}^n$, the Boolean lattice can be regarded as a metric space $(\{0, 1\}^n, d_{\mathrm{Ham}})$ with the Hamming distance. See Appendix A and Ganter & Wille (1999); Davey & Priestley (2002) for more details.

Definition 1 (ℓ_1-product metric space)**.** *Let $\{(X_i, d_i)\}_{i=1}^k$ be non-trivial metric spaces. An ℓ_1-*product metric space *of $\{(X_i, d_i)\}_{i=1}^k$ is a Cartesian product space $\prod_{i=1}^k X_i$ equipped with the ℓ_1-product metric $(\sum_{i=1}^k d_i)((\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}), (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(k)})) = \sum_{i=1}^k d_i(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$. If not ambiguous even without subscripts, this space is denoted by (X^k, d_{X^k}) for brevity.*

Proposition 1 (Embedding of Boolean Lattice). *A Boolean lattice $(2^{\mathcal{C}}, \preceq)$ for n atomic concepts can be embedded into the poset $(\mathbb{R}^n, \preceq)$ used by order embeddings while preserving the order relations. As a metric space $(\{0,1\}^n, d_{\mathrm{Ham}})$, it is isometrically embedded into an ℓ_1-product metric space $(\prod_{i=1}^k X_i, \sum_{i=1}^k d_i)$ for any $k \geq n$ after appropriate per-factor scaling. However, it admits no isometric embedding into a hyperbolic space $\mathbb{H}^d$ for any $d \geq 2$ and $n \geq 2$.*

See Appendix B for the proof. A Boolean lattice has remarkable expressivity for compositionality, but it is often too coarse as it considers combinations of all atomic concepts. Order embeddings enrich it by replacing each bit $\{0,1\}$ with $\mathbb{R}$, whereas a single hyperbolic space does not.

3 PHyCLIP and its Loss Functions

Embedding into an ℓ_1-Product Metric Space of Hyperbolic Factors. We extend the Boolean lattice by replacing each bit $\{0,1\}$ with a metric tree T_i and leverage the expressive power of both the hyperbolic embeddings for hierarchy and the Boolean lattice for compositionality. In this setting, the description "a dog in a car" is represented by a pair of nodes in metric trees T_1 and T_2 that encode *is-a* taxonomies of animals (e.g., $\mathsf{dog} \preceq \mathsf{mammal} \preceq \mathsf{animal}$) and transportation (e.g., $\mathsf{car} \preceq \mathsf{vehicle} \preceq \mathsf{transport}$), respectively. Notably, a *single* hyperbolic space cannot capture this product geometry (see Proposition 2 in Appendix B), whereas an ℓ_1-product metric space of hyperbolic factors can.

Theorem 2 (Embedding into an ℓ_1-product metric space of hyperbolic factors). *Let $T_1, \ldots, T_k$ be finite metric trees (or infinite metric trees with known bounds for maximum degree and minimum edge length) with metrics $d_{T_1}, \ldots, d_{T_k}$. For every $\varepsilon > 0$, there exists a $(1+\varepsilon, 0)$-quasi-isometric embedding from the ℓ_1-product metric space of these metric trees, $(\prod_{i=1}^k T_i,\ \sum_{i=1}^k d_{T_i})$, into an ℓ_1-product metric space of k two-dimensional hyperbolic factors, $((\mathbb{H}^2)^k, d_{(\mathbb{H}^2)^k})$, after appropriate per-factor scaling.*

Given the above, we propose embeddings into an ℓ_1-product metric space of k copies of d-dimensional hyperbolic factors $\mathbb{H}^d$, $((\mathbb{H}^d)^k,\ d_{(\mathbb{H}^d)^k})$. The total dimension of $(\mathbb{H}^d)^k$ is kd. Each hyperbolic factor $\mathbb{H}^d_i$ is intended to represent the taxonomy of a concept family as well as aspects of inter-object relations (e.g., "a dog riding on something", "a car loading something"). An instance is embedded as a tuple $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)})$ for $\boldsymbol{x}^{(i)} \in \mathbb{H}^d_i$. Within each factor $\mathbb{H}^d_i$, we use standard hyperbolic embeddings (Nickel & Kiela, 2017) together with hyperbolic entailment cones (Ganea et al., 2018a) to encode *intra-family* hierarchy and image–text entailment, while *cross-family* compositionality is captured by the additive geometry of the ℓ_1-product metric space.

PHyCLIP for Vision–Language Representation Learning. Here, we propose PHyCLIP for vision–language representation learning, depicted in Fig. 2. Let I and T denote instances of images and texts, respectively. From an instance, a kd-dimensional feature vector is produced, which is then sliced into k segments $\boldsymbol{v}^{(i)}$ of dimension d for $i = 1, \ldots, k$, and each segment $\boldsymbol{v}^{(i)}$ is lifted via the exponential map to its corresponding hyperbolic factor $\mathbb{H}^d_i$ as $\boldsymbol{x}^{(i)}$, yielding the embedding $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(k)}) \in (\mathbb{H}^d)^k$. We denote the embeddings of the image I and text T by $\boldsymbol{I}$ and $\boldsymbol{T}$, respectively. Let B denote the index set of instances in a mini-batch; we write the mini-batch of images as $\{\boldsymbol{I}_b\} = \{\boldsymbol{I}_b\}_{b \in B}$ for brevity.

An image I is typically more specific than its corresponding text T ($I \preceq T$) as the text T may ignore some details of the image I. Following HyCoCLIP (Pal et al., 2025), we suppose that the training data are enriched with box information: the image boxes I^{box} are object-level crops of the original images I, and the text boxes T^{box} are the corresponding nouns/phrases within the text T ($I^{\mathrm{box}} \preceq T^{\mathrm{box}}$). An image box I^{box} and a text box T^{box} are more general than the full image I and the full text T ($I \preceq I^{\mathrm{box}}$, $T \preceq T^{\mathrm{box}}$), respectively, since they omit objects and words outside the boxes.

We will introduce the contrastive loss $\mathcal{L}_{\mathrm{cont}}$ and entailment loss $\mathcal{L}_{\mathrm{ent}}$, and the final objective is their sum weighted by a hyperparameter γ:

$$\mathcal{L}_{\mathrm{overall}} = \mathcal{L}_{\mathrm{cont}} + \gamma \mathcal{L}_{\mathrm{ent}}. \tag{1}$$

Contrastive Loss. To represent each hyperbolic factor $\mathbb{H}^d_i$, we adopt the Lorentz model with a learnable curvature $-\alpha_i$ (Cannon et al., 1997; Nickel & Kiela, 2018; Lee, 2018). See Appendix C

for implementation details. Following Definition 1, we define the distance on the ℓ_1-product metric space $(\mathbb{H}^d)^k$ and its averaged version as

$$d_{(\mathbb{H}^d)^k}(\boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{k} d_{\mathbb{H}^d_i}(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}), \quad d_{\text{avg}}(\boldsymbol{X}, \boldsymbol{Y}) = \tfrac{1}{k} d_{(\mathbb{H}^d)^k}(\boldsymbol{X}, \boldsymbol{Y}). \tag{2}$$

To pull an embedding $\boldsymbol{X}_b$ close to its positive pair $\boldsymbol{Y}_b$ while pushing it away from negatives $\boldsymbol{Y}_a$ for $a \neq b$, we use the standard InfoNCE loss (Radford et al., 2021; Desai et al., 2023; Pal et al., 2025):

$$\mathcal{L}_{\text{cont}}(\{\boldsymbol{X}_b\}, \{\boldsymbol{Y}_b\}) = -\sum_{b \in B} \log \frac{\exp(-d_{\text{avg}}(\boldsymbol{X}_b, \boldsymbol{Y}_b)/\tau)}{\sum_{a \in B} \exp(-d_{\text{avg}}(\boldsymbol{X}_b, \boldsymbol{Y}_a)/\tau)} \tag{3}$$

where τ is a learnable temperature parameter. We average this loss over known pairs:

$$\mathcal{L}_{\text{cont}} = \tfrac{1}{4}(\mathcal{L}_{\text{cont}}(\{\boldsymbol{I}_b\}, \{\boldsymbol{T}_b\}) + \mathcal{L}_{\text{cont}}(\{\boldsymbol{T}_b\}, \{\boldsymbol{I}_b\}) + \mathcal{L}_{\text{cont}}(\{\boldsymbol{I}_b^{\text{box}}\}, \{\boldsymbol{T}_b^{\text{box}}\}) + \mathcal{L}_{\text{cont}}(\{\boldsymbol{T}_b^{\text{box}}\}, \{\boldsymbol{I}_b^{\text{box}}\})). \tag{4}$$

Entailment Loss. We also employ hyperbolic entailment cones to capture the entailment relations (Ganea et al., 2018a). See Appendix C for implementation details. For every point $\boldsymbol{y}^{(i)}$ in each hyperbolic factor $\mathbb{H}^d_i$, we define a geodesic conical region $C(\boldsymbol{y}^{(i)})$ with apex at $\boldsymbol{y}^{(i)}$ and half-aperture $\omega(\boldsymbol{y}^{(i)})$, where all points $\boldsymbol{x}^{(i)} \in C(\boldsymbol{y}^{(i)})$ are considered more specific than $\boldsymbol{y}^{(i)}$ (i.e., $\boldsymbol{x}^{(i)} \preceq \boldsymbol{y}^{(i)}$). Then, $\boldsymbol{x}^{(i)} \in C(\boldsymbol{y}^{(i)})$ iff $\phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) < \omega(\boldsymbol{y}^{(i)})$ for the exterior angle $\phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$. To penalize the violation of the inclusion relation $\boldsymbol{x}^{(i)} \in C(\boldsymbol{y}^{(i)})$ for a pair $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ such that $\boldsymbol{x}^{(i)} \preceq \boldsymbol{y}^{(i)}$, the entailment loss L_{ent} is calculated as

$$L_{\text{ent},i}(\boldsymbol{X}, \boldsymbol{Y}) = \max(0, \phi(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}) - \eta\omega(\boldsymbol{y}^{(i)})), \quad L_{\text{ent}}(\boldsymbol{X}, \boldsymbol{Y}) = \tfrac{1}{k}\sum_{i=1}^{k} L_{\text{ent},i}(\boldsymbol{X}, \boldsymbol{Y}), \tag{5}$$

where hyperparameter η controls the margin (Pal et al., 2025). We sum this loss over known pairs:

$$\mathcal{L}_{\text{ent}} = \sum_{b \in B} \left(L_{\text{ent}}(\boldsymbol{I}_b, \boldsymbol{T}_b) + L_{\text{ent}}(\boldsymbol{I}_b^{\text{box}}, \boldsymbol{T}_b^{\text{box}}) + L_{\text{ent}}(\boldsymbol{I}_b, \boldsymbol{I}_b^{\text{box}}) + L_{\text{ent}}(\boldsymbol{T}_b, \boldsymbol{T}_b^{\text{box}}) \right). \tag{6}$$

4 Experiments

4.1 Training Details

Datasets. We trained all models on the Grounded Image–Text Pairs (GRIT) dataset (Peng et al., 2023), which consists of automatically annotated image–text pairs with bounding boxes and corresponding nouns/phrases. Although the dataset is documented to contain 20.5 million pairs with 35.9 million box annotations, we were able to obtain 14.0 million pairs with 26.6 million box annotations due to outdated public links. This scale remains considerably larger than manually annotated resources such as Flickr30K Entities (Plummer et al., 2015).

Baselines. We compare PHyCLIP with CLIP (Radford et al., 2021), MERU (Desai et al., 2023), and HyCoCLIP (Pal et al., 2025). CLIP is a seminal vision–language model trained with contrastive learning in a Euclidean space. MERU extends CLIP by lifting embeddings to hyperbolic space and using hyperbolic entailment cones to represent hierarchy. HyCoCLIP further leverages box annotations to better capture intra-modal hierarchy. All models were trained from scratch on GRIT for fair comparison. For PHyCLIP, we set the number of factors to $k = 64$ and the dimension of each factor to $d = 8$, resulting in a total dimension of 512. We followed the training protocols and hyperparameters used in the official implementations of HyCoCLIP (Pal et al., 2025); see Appendix C for details. We report results obtained with the base Vision Transformer as an image encoder (Dosovitskiy et al., 2021). Supplementary results are provided in Appendix D.1.

4.2 Experimental Results

Zero-shot Image Classification. We evaluated the geometry of embedding space via zero-shot image classification, following the protocol standardized by CLIP (Radford et al., 2021). Images are classified using the similarity to the averaged embedding of template text queries for classes across 16 datasets, grouped into General, Fine-grained, and Specialized. General datasets cover broad, heterogeneous concept families (e.g., animals, transportation, household objects). Fine-grained datasets focus on visually similar subclasses within a single concept family (e.g., specific food, dog breeds). Specialized datasets are domain-specific (e.g., texture images, satellite imagery).

Table 1: Zero-shot image classification.

		General datasets						Fine-grained datasets						Specialized datasets			
	w/ boxes	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
CLIP		39.36	75.09	48.57	51.48	73.63	92.54	50.59	13.40	7.66	2.42	46.44	19.00	23.19	35.26	42.60	5.20
CLIP	✓	38.64	76.88	47.88	50.62	74.48	93.34	50.77	13.45	8.99	3.17	46.95	20.67	21.49	36.51	41.46	4.90
MERU		37.49	75.61	46.80	49.54	71.19	93.38	52.88	10.52	7.49	3.05	44.11	**22.94**	21.70	**39.52**	41.09	4.74
MERU	✓	37.86	77.14	48.09	50.15	72.96	93.80	53.61	9.34	7.42	3.06	43.69	17.92	21.38	35.02	40.98	5.20
HyCoCLIP	✓	42.93	88.51	57.68	54.23	75.55	94.55	51.72	12.86	9.98	**4.41**	50.66	19.93	26.33	38.02	46.15	**5.65**
PHyCLIP	✓	**44.43**	**89.30**	**59.83**	**56.18**	**75.76**	**95.06**	**56.81**	**16.00**	**10.47**	3.05	**54.64**	20.41	**26.44**	33.43	**50.13**	5.42

The best and second performances are emphasized by bold fonts and underlines, respectively.

Table 2: Zero-shot retrieval and hierarchical classification.

		Text → Image				Image → Text				Hierarchical Classification				
		COCO		Flickr		COCO		Flickr		WordNet				
	w/ boxes	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	TIE(↓)	LCA(↓)	J(↑)	P_H(↑)	R_H(↑)
CLIP		56.39	67.59	83.30	89.70	70.44	80.42	**93.10**	95.70	3.705	2.254	0.7805	0.8498	0.8503
CLIP	✓	56.12	67.58	82.54	89.32	**70.72**	80.32	91.90	**96.10**	3.720	2.265	0.7797	0.8487	0.8509
MERU		55.50	66.71	82.26	88.84	69.32	78.96	89.70	95.70	3.832	2.292	0.7720	0.8451	0.8439
MERU	✓	55.93	67.29	81.68	88.36	69.72	80.02	91.00	95.70	3.793	2.277	0.7740	0.8462	0.8454
HyCoCLIP	✓	56.24	67.69	82.90	88.94	69.00	79.16	91.90	95.30	3.378	2.113	0.8008	0.8653	0.8636
PHyCLIP	✓	**58.00**	**68.74**	**83.40**	**89.92**	70.20	**80.44**	91.10	95.60	**3.285**	**2.088**	**0.8065**	**0.8684**	**0.8682**

Table 1 summarizes top-1 accuracies. PHyCLIP obtained consistent performance gains, particularly on General datasets, which we attribute to assigning concept families to hyperbolic factors that naturally support coarse-grained classifications. Within Fine-grained datasets, PHyCLIP achieved remarkable improvements on Food-101 (Bossard et al., 2014) and Oxford-IIIT Pets (Parkhi et al., 2012), implying that it also learned intra-family taxonomies without confusion with other families. Although not best on every dataset, the performance gap on Flowers-102 (Nilsback & Zisserman, 2008) is small, FGVC-Aircraft (Maji et al., 2013) and Country211 (Radford et al., 2021) remain challenging for all models due to extreme intra-class similarity, and EuroSAT (Helber et al., 2019) (comprising satellite imagery) is out-of-distribution relative to GRIT. Consistent with prior findings (Pal et al., 2025), CLIP and MERU do not yield clear improvements with box annotations. Overall, PHyCLIP is the strongest zero-shot classifier among the comparison models.

Zero-shot Image and Text Retrieval. We evaluate cross-modal alignment via zero-shot retrieval in the shared embedding space: given a text query, retrieve the nearest images, and vice versa. This is also a standard benchmark for vision–language models (Radford et al., 2021). We used the COCO validation set (Lin et al., 2014) and the Flickr30K test set (Young et al., 2014; Karpathy & Fei-Fei, 2015). We report Recall at k (R@k), the fraction of queries for which the paired instance appears in the top-k retrieved results.

Results are summarized in the left half of Table 2. PHyCLIP achieves the best performance across all metrics and datasets on image retrieval, which supports our choice of the ℓ_1-product metric in Eq. (2). This metric sums distances over hyperbolic factors; when an object specified in the text is absent from a candidate image, or an unspecified object is present, the corresponding factor incurs a large penalty. By contrast, a single hyperbolic space implicitly encodes the presence or absence of objects as hierarchical relations, which may weaken penalties for such mismatches and hinder separability of hard negatives. The vanilla CLIP works well for text retrieval, and PHyCLIP attains a competitive performance. Texts are more diverse and ambiguous than images, and some texts may accidentally match with non-paired images, which may limit the benefits of our design.

Hierarchical Classification. We evaluate the expressivity for the *is-a* taxonomy via hierarchical classification (Kosmopoulos et al., 2015; Pal et al., 2025) on ImageNet (Russakovsky et al., 2015), where class labels are enriched by WordNet (Miller, 1995) and errors between predicted and ground-truth classes are measured on the WordNet graph with unit-length edges: Tree Induced Error (TIE) is their graph distance; Lowest Common Ancestor (LCA) error is the maximum of the distances to their LCA; Jaccard similarity J, hierarchical precision P_H, and hierarchical recall R_H are similarities between the sets of ancestors.

Table 3: Compositional understanding through hard-negative classification.

	w/ boxes	VL-CheckList–Object						SugarCrepe							
		Location			Size			Replace			Swap		Add		
		Center	Mid	Margin	Large	Medium	Small	Obj	Att	Rel	Obj	Att	Obj	Att	Overall
CLIP		67.1	65.5	64.3	69.9	63.8	64.2	88.74	80.84	**69.42**	**63.67**	64.86	80.60	72.83	77.66
CLIP	✓	66.1	61.6	64.7	67.0	64.6	63.3	89.29	81.73	69.84	62.45	64.11	80.12	71.68	77.61
MERU		63.3	60.0	60.5	66.6	57.3	58.6	88.68	80.71	69.27	57.55	64.11	80.16	**74.42**	77.37
MERU	✓	62.6	58.3	59.8	62.6	60.3	59.8	89.53	79.06	69.91	56.33	66.37	79.97	75.72	77.73
HyCoCLIP	✓	65.9	65.6	63.1	67.6	63.1	63.9	91.28	80.46	67.07	54.69	63.96	81.09	72.40	77.46
PHyCLIP	✓	**73.0**	**72.0**	**71.4**	**76.4**	**69.2**	**69.0**	**91.34**	**82.11**	66.64	59.18	**66.07**	**83.56**	74.28	**78.75**

Results are summarized in the right half of Table 2. PHyCLIP achieves superior scores across all metrics, indicating not only higher classification accuracy but also that misclassifications tend to be close to the ground-truth class in the taxonomy. By handling cross-family compositionality via the ℓ_1-product metric, each hyperbolic factor can devote capacity to a cleaner intra-family *is-a* taxonomy, thereby yielding disentangled, hierarchy-aligned representations.

Compositional Understanding. We assess the expressivity of compositionality via hard negative classification using VL-CheckList (Zhao et al., 2022) and SugarCrepe (Hsieh et al., 2023). Both benchmarks require models to distinguish ground-truth captions from hard negatives created by altering objects, attributes, or relations in the ground-truth captions. Following Pal et al. (2025), we evaluate the Object subset of VL-CheckList, in which a noun for a single object in each caption is randomly replaced. The results are summarized by the replaced object's location (center/mid/margin) and size (small/medium/large) in the image. We also evaluate all seven subsets in SugarCrepe, in which objects, attributes, and relations are replaced, swapped, or added in each caption.

As shown in Table 3, PHyCLIP yields a substantial improvement on VL-CheckList–Object. It successfully represents object presence robustly with respect to location and size. On SugarCrepe, PHyCLIP obtains the best scores on four out of seven subsets and the second-best on two subsets; its average score exceeds that of the second-best model by more than 1%, whereas the other models cluster within 0.3%. Performance on attribute subsets is robust across all three operations, suggesting that our design decouples intra-family taxonomy from cross-family composition and thereby emphasizes attribute–object binding. By contrast, we observe modest drops in relation replacement and object swapping, which implies that our design is less sensitive to inter-object relations, although it potentially captures these relations within each hyperbolic factor.

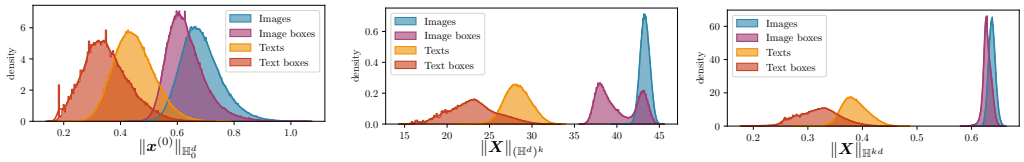
Ablation Study. We investigate the contributions of embedding space factorization and the ℓ_1-product metric through ablation studies, summarized in Table 4. We fix the total embedding dimension kd and vary the number of factors, k. When $k = 1$ (equivalent to HyCoCLIP), performance is the lowest on most metrics; increasing k generally improves results, except for text retrieval, thereby demonstrating the benefit of factorization. Performance peaks at $k = 64$ or $k = 128$, although zero-shot classification accuracy for Food-101 (Bossard et al., 2014) drops substantially at $k = 128$, indicating that overly fine factorization may impair the representation of intra-family taxonomy. Replacing the ℓ_1-product metric with the Riemannian (ℓ_2) product metric consistently degrades performance, except for text retrieval. This result supports that the ℓ_1-product metric provides a more effective way to aggregate cross-family composition.

Table 4: Ablation study.

# of factors, k	# of dims., d	product metric	classification		retrieval COCO, R@5		hierarchical	
			ImageNet	Food-101	Image	Text	TIE	J
1	512	–	42.93	51.71	56.24	69.00	3.378	0.8008
8	64	ℓ_1	44.26	52.16	57.28	69.38	3.288	0.8061
16	32	ℓ_1	44.03	54.89	56.78	67.62	3.292	0.8063
32	16	ℓ_1	43.90	54.48	56.70	66.92	3.324	0.8035
64	8	ℓ_1	**44.43**	**56.81**	**58.00**	70.20	3.285	0.8065
128	4	ℓ_1	44.08	52.61	57.82	**71.44**	**3.278**	**0.8073**
64	8	ℓ_2	43.46	51.44	57.72	71.40	3.377	0.7998

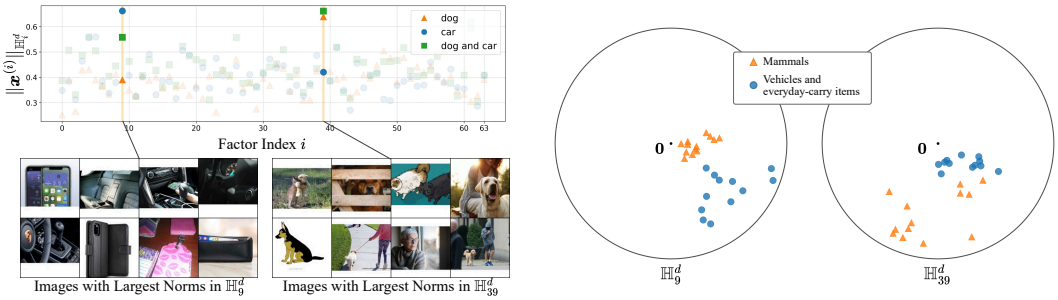
4.3 Visualizations of Hyperbolic Factors

Norm Distributions. Figure 3 plots the empirical distributions of embedding norms. As shown in (b) and (c), in both PHyCLIP and HyCoCLIP, image norms are consistently larger than text norms and are tightly concentrated. These models consider images to be more specific than their paired texts, $I_b \preceq T_b$, which encourages the image embedding $\boldsymbol{I}_b$ to lie within the text's hyperbolic entailment cone $C(\boldsymbol{T}_b)$ (i.e., $\boldsymbol{I}_b \in C(\boldsymbol{T}_b)$), yielding larger image norms. However, within individual hyperbolic factors



(a) PHyCLIP (factor-wise norms) (b) PHyCLIP (overall norms) (c) HyCoCLIP

Figure 3: **Norm distributions.** In (b) and (c), image norms are consistently larger than text norms, because images are more specific than their paired texts ($I_b \preceq T_b$). However, in a single hyperbolic factor shown in (a), image and text norms largely overlap, as PHyCLIP may keep some factors unused for instances that do not contain the corresponding concept families.



(a) dog and car (b) Embeddings projected onto 2D disks by HoroPCA.

Figure 4: **Visualization of factor-wise embeddings.** (a) Each concept (e.g., dog or car) activates a distinct factor (i.e., $i = 39$ or $i = 9$), and their composition (e.g., "a dog and a car") activates the corresponding factors simultaneously. (b) A set of relevant concepts (e.g., hyponyms of mammals) forms a hierarchical structure in the corresponding factor (e.g., $i = 39$), while they cluster near the origin in another factor (e.g., $i = 9$).

of PHyCLIP in (a), the image and text distributions largely overlap and are broadly dispersed. This is because instances without a particular concept family lie near the origin in the corresponding factor, in other words, factors are used selectively on a per-instance basis. Consequently, PHyCLIP leverages a broader portion of the embedding space and facilitates meaningful distances and taxonomic structures under contrastive learning.

Composition via ℓ_1-Product Metric. To examine the behavior of the ℓ_1-product metric, we obtain factor-wise embeddings $\boldsymbol{x}^{(i)}$ of single-concept prompts (e.g., "a photo of a dog" and "a photo of a car") and their compositions (e.g., "a photo of a dog and a car"). Figure 4 (a) shows factor-wise embedding norms $\|\boldsymbol{x}^{(i)}\|_{\mathbb{H}^d_i}$ across $k = 64$ factors. The "dog" embedding exhibits its largest norm in factor $i = 39$ while remaining near the origin in factor $i = 9$. Conversely, the "car" embedding peaks in factor $i = 9$ and is suppressed in factor $i = 39$. Then, their composition produces large norms in *both* factors $i = 39$ and $i = 9$, meaning that composing concepts simultaneously activates the corresponding factors. We observe the same pattern for "boy and bicycle" and "sunset and ocean" (see Appendix D.2). This pattern aligns with the behavior of a Boolean algebra, where multiple concepts are specified by the union of concept subsets (or, the element-wise $\max$ of binary indicators).

Figure 4 (a) also provides GRIT images randomly sampled from the top 0.1% by the embedding norm for each factor. Factor $i = 39$ yields various mammals, suggesting a family of mammals, whereas factor $i = 9$ shows vehicles and everyday-carry items. Embeddings visualized using HoroPCA (Chami et al., 2021) in Fig. 4(b) support this interpretation. Terms related to mammals form a hierarchical structure (i.e., captured) in factor $i = 39$ and concentrate near the origin (i.e., not captured) in factor $i = 9$. Embeddings related to vehicles and everyday-carry items exhibit the opposite pattern. See Appendix D.2 for more visualizations.

We emphasize that, while we give a hierarchy between samples, we do not provide any explicit supervision for factor assignments; this specialization of factors emerges automatically through training. Consistent with Theorem 2, these observations empirically support that PHyCLIP organizes intra-family taxonomies within individual hyperbolic factors and expresses inter-family compositionality via the simultaneous activation of multiple factors, analogous to a Boolean algebra.

5 Related Work

Vision–Language Models and Representation Learning. Vision–language representation learning contributes to retrieval (Mori et al., 1999), semantic segmentation (Barnard et al., 2003), and image generation (Ramesh et al., 2021; Labs, 2025). Early works learned alignments through object-level, word-based classification and detection (Karpathy & Fei-Fei, 2015; He & Peng, 2017; Engilberge et al., 2018) or through text–image generation (Peng et al., 2017; Gu et al., 2018), but they often required complex annotation and network designs (Zhao et al., 2022). A more generic approach maps an entire image or text to a single vector and learns a shared embedding space with a contrastive objective. Representative systems include DeViSE (Frome et al., 2013), VSE++ (Faghri et al., 2018), CLIP (Radford et al., 2021), and ALIGN (Jia et al., 2021). Our model, PHyCLIP, follows this line, while implicitly extracting individual concepts through the geometry of an ℓ_1-product metric space.

Hyperbolic Representations in Deep Learning. Data often exhibit hierarchical, tree-like structures. Many approaches have attempted to encode such structure (Nguyen et al., 2017; Vulic & Mrksic, 2018), and hyperbolic spaces have become influential due to their empirical performance and theoretical support (Sala et al., 2018; Sonthalia & Gilbert, 2020). As discussed in Section 2, tree metrics admit quasi-isometric embeddings into the two-dimensional hyperbolic plane, which enhances generalization and interpretability (Bridson & Haefliger, 1999; Sarkar, 2011). Hyperbolic embeddings have been applied to words (Nickel & Kiela, 2017; 2018; Tifrea et al., 2019), sentences (Dhingra et al., 2018), graphs (Liu et al., 2019), and images (Khrulkov et al., 2020; Atigh et al., 2022; van Spengler et al., 2023; Qiu et al., 2024). There is also extensive work on building neural networks on hyperbolic spaces (Ganea et al., 2018b; Shimizu et al., 2021; Takeuchi et al., 2022; Peng et al., 2022) and on optimization over Riemannian manifolds (Bonnabel, 2013; Bécigneul & Ganea, 2019). Within vision–language learning, MERU adapts CLIP to hyperbolic geometry (Desai et al., 2023). Our method also leverages hyperbolic geometry and embeds tree-like structures efficiently.

For non-hierarchical data, Euclidean, hyperspherical, or toroidal geometries can be effective (Ebisu & Ichise, 2018), and several studies explore representations in a Riemannian (ℓ_2) product of such spaces as mixed-curvature representations (Gu et al., 2019; Wang et al., 2024; Gao et al., 2025). PHyCLIP also employs a product space, but all factors are hyperbolic and the product metric is ℓ_1; we justified both choices theoretically in Section 2.

Region-based Embeddings for Structured Representations. Hierarchical relations can be viewed as a form of inclusion relations. Order embeddings (Vendrov et al., 2016) represent an instance as an upper orthant of Euclidean space, and box embeddings (Vilnis et al., 2018) represent it as an axis-aligned hyperrectangle, where the inverse of set inclusion encodes the hierarchical relation. Euclidean variants include Gaussian embeddings (Vilnis & Mccallum, 2015), and hyperbolic variants include disk embeddings (Nickel & Kiela, 2018) and hyperbolic entailment cones (Ganea et al., 2018a). These approaches have also been employed in the vision–language setting (Ren et al., 2016; Desai et al., 2023; Pal et al., 2025). We summarize their theoretical connections in Appendix A.2. These region-based approaches permit composition via intersection of regions, which allows multiple parents and richer semantic composition. However, their compositional expressivity has not yet been fully characterized. We showed in Section 2 that order embeddings and PHyCLIP support compositionality at the level of a Boolean algebra, while a single hyperbolic space may not.

6 Conclusion

We introduced PHyCLIP, a vision–language model that learns representations using an ℓ_1-product metric space of hyperbolic factors. We theoretically and empirically demonstrated that it simultaneously captures compositionality across concept families through the ℓ_1-product metric, as well as *is-a* taxonomies within hyperbolic spaces via hyperbolic embeddings. This design yields state-of-the-art performance across various downstream tasks and provides an interpretable embedding structure. While the main focus is on object composition, it also performs well for attribute binding because it decouples intra-family taxonomy from cross-family composition. By contrast, the relational structure remains unexplored; incorporating its algebraic structure is a promising direction for future work.

Ethics Statement

This study is purely focused on vision–language representation learning, and it is not expected to have any direct negative impact on society or individuals.

Reproducibility Statement

The environment, datasets, methods, evaluation metrics, and other experimental settings are provided in Section 4 and Appendix C. For full reproducibility, the source code is attached as supplementary material.

References

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

Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching Words and Pictures. *Journal of Machine Learning Research*, 2003.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian Adaptive Optimization Methods. In *International Conference on Learning Representations (ICLR)*, 2019.

Silvère Bonnabel. Stochastic Gradient Descent on Riemannian Manifolds. *IEEE Transactions on Automatic Control*, 2013.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining Discriminative Components with Random Forests. In *European Conference on Computer Vision (ECCV)*, 2014.

Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature*. Springer, 1999.

James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic Geometry. *Flavors of Geometry*, 1997.

Ines Chami, Albert Gu, Dat Nguyen, and Christopher Ré. HoroPCA: Hyperbolic Dimensionality Reduction via Horospherical Projections. In *International Conference on Machine Learning (ICML)*, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 2017.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing Textures in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *Artificial Intelligence and Statistics (AISTATS)*, 2011.

Shib Sankar Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Lorraine Li, and Andrew McCallum. Improving Local Identifiability in Probabilistic Box Embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

B. A. Davey and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge University Press, 2002.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic Image-text Representations. In *International Conference on Machine Learning (ICML)*, 2023.

Michel Marie Deza and Monique Laurent. *Geometry of Cuts and Metrics*. Springer, 1997.

Bhuwan Dhingra, Christopher J. Shallue, Mohammad Norouzi, Andrew M. Dai, and George E. Dahl. Embedding Text in Hyperbolic Spaces. In *ACL Workshop on Graph-Based Methods for Natural Language Processing*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Takuma Ebisu and Ryutaro Ichise. TorusE: Knowledge Graph Embedding on a Lie Group. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Martin Engilberge, Louis Chevallier, Patrick Perez, and Matthieu Cord. Finding Beans in Burgers: Deep Semantic-Visual Embedding with Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference (BMVC)*, 2018.

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. *CVPR Workshop*, 2004.

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. In *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.

Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, 1999.

Yuxiao Gao, Fuwei Zhang, Zhao Zhang, Xiaoshuang Min, and Fuzhen Zhuang. Mixed-Curvature Multi-Modal Knowledge Graph Completion. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

Albert Gu, Frederic Sala, Beliz Gunel, and Christopher Ré. Learning Mixed-Curvature Representations in Product Spaces. In *International Conference on Learning Representations (ICLR)*, 2019.

Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiangteng He and Yuxin Peng. Fine-Grained Image Classification via Combining Vision and Language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Patrick Helber, Benjamin Bischke, Andreas R. 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.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv*, 2018.

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Distributed Representations. In *Parallel Distributed Processing, Volume 1: Explorations in the Microstructure of Cognition: Foundations*. The MIT Press, 1986.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *NeurIPS Datasets and Benchmarks*, 2023.

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 *International Conference on Machine Learning (ICML)*, 2021.

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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

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*, 2015.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *ICCV Workshops*, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009.

Black Forest Labs. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv*, 2025.

John M. Lee. *Introduction to Riemannian Manifolds, Second Edition*. Springer International Publishing AG, 2018.

Xiang Li, Luke Vilnis, Dongxu Zhang, Michael Boratko, and Andrew McCallum. Smoothing the Geometry of Probabilistic Box Embeddings. In *International Conference on Learning Representations (ICLR)*, 2019.

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 *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.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-Grained Visual Classification of Aircraft. *arXiv*, 2013.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 1995.

Andriy Mnih and Geoffrey Hinton. A Scalable Hierarchical Distributed Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2008.

Yasuhide Mori, Hironobu Takahashi, and Ryu ichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *IEEE International Conference on Image Processing*, 1999.

Frederic Morin and Yoshua Bengio. Hierarchical Probabilistic Neural Network Language Model. In *Artificial Intelligence and Statistics (AISTATS)*, 2005.

Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. Hierarchical Embeddings for Hypernymy Detection and Directionality. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

Maximilian Nickel and Douwe Kiela. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Maximilian Nickel and Douwe Kiela. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In *International Conference on Machine Learning (ICML)*, 2018.

M-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional Entailment Learning for Hyperbolic Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and Dogs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Yuxin Peng, Jinwei Qi, and Yuxin Yuan. CM-GANs: Cross-modal Generative Adversarial Networks for Common Representation Learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2017.

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*, 2023.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015.

Zexuan Qiu, Jiahong Liu, Yankai Chen, and Irwin King. HiHPQ: Hierarchical Hyperbolic Product Quantization for Unsupervised Image Retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. *arXiv*, 2021.

Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint Image-Text Representation by Gaussian Visual-Semantic Embedding. In *ACM International Conference on Multimedia (ACMMM)*, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015.

Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation Tradeoffs for Hyperbolic Embeddings. In *International Conference on Machine Learning (ICML)*, 2018.

Rik Sarkar. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In *International Symposium on Graph Drawing (GD)*, 2011.

Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic Neural Networks++. In *International Conference on Learning Representations (ICLR)*, 2021.

Rishi Sonthalia and Anna C. Gilbert. Tree! I Am No Tree! I Am a Low Dimensional Hyperbolic Embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ryota Suzuki, Ryusuke Takahama, and Shun Onoda. Hyperbolic Disk Embeddings for Directed Acyclic Graphs. In *International Conference on Machine Learning (ICML)*, 2019.

Jun Takeuchi, Qian Chen, Hiroshi Noji, and Isao Goto. Neural Networks in a Product of Hyperbolic Spaces. In *North American Chapter of the Association for Computational Linguistics (NAACL) Student Research Workshop*, 2022.

Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincare Glove: Hyperbolic Word Embeddings. In *International Conference on Learning Representations (ICLR)*, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021.

Abraham Albert Ungar. *A Gyrovector Space Approach to Hyperbolic Geometry*. Morgan and Claypool, 2008.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-Embeddings of Images and Language. In *International Conference on Learning Representations (ICLR)*, 2016.

Luke Vilnis and Andrew Mccallum. Word Representations via Gaussian Embedding. In *International Conference on Learning Representations (ICLR)*, 2015.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Ivan Vulic and Nikola Mrksic. Specialising Word Vectors for Lexical Entailment. In *North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2018.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, 2011.

Jiapu Wang, Zheng Cui, Boyue Wang, Shirui Pan, Junbin Gao, Baocai Yin, and Wen Gao. IME: Integrating Multi-curvature Shared and Specific Embedding for Temporal Knowledge Graph Completion. In *International Conference on World Wide Web (WWW)*, 2024.

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

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*, 2014.

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*, 2022.

A Background Theory

A.1 Quasi-isometric Embeddings

We adopt standard notions from Bridson & Haefliger (1999). Let (X, d_X) be a metric space. A geodesic segment $[x, y] \subset X$ is an isometric image of an interval whose endpoints are mapped to x and y in X. The space X is *geodesic* if every pair of points can be joined by a geodesic segment. For $\delta \geq 0$, a geodesic triangle is δ-*slim* if each side is contained in the δ-neighborhood of the union of the other two sides. A geodesic metric space X is δ-*hyperbolic* (in the sense of Gromov) if every geodesic triangle in X is δ-slim. A metric tree is a geodesic metric space in which any two nodes are joined by a unique geodesic, and every geodesic triangle is a tripod; hence, it is 0-hyperbolic. Any tree with positive edge lengths, equipped with path length as distance, is a metric tree. Euclidean spaces $\mathbb{R}^n$ are not Gromov-hyperbolic for $n \geq 2$, whereas hyperbolic spaces $\mathbb{H}^n$ are δ-hyperbolic, where δ depends only on the curvature.

Definition 2 (Quasi-isometric embedding (Bridson & Haefliger, 1999)). *Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : (X, d_X) \to (Y, d_Y)$ is a (λ, c)-quasi-isometric embedding if it is injective and there exist a distortion $\lambda \geq 1$ and an error $c \geq 0$ such that*

$$\tfrac{1}{\lambda}\, d_X(\boldsymbol{x}, \boldsymbol{x}') - c \;\leq\; d_Y\big(f(\boldsymbol{x}), f(\boldsymbol{x}')\big) \;\leq\; \lambda\, d_X(\boldsymbol{x}, \boldsymbol{x}') + c \quad \textit{for all } \boldsymbol{x}, \boldsymbol{x}' \in X. \tag{7}$$

If $\lambda = 1$ and $c = 0$, the embedding is isometric.

A.2 Representations of Poset and Lattice

A *poset* $(P, \preceq)$ is a set P with a reflexive, antisymmetric, and transitive relation $\preceq$. Its Hasse diagram places an edge from $\boldsymbol{x}$ to $\boldsymbol{y}$ when $\boldsymbol{x} \prec \boldsymbol{y}$ and there is no $\boldsymbol{z}$ such that $\boldsymbol{x} \prec \boldsymbol{z} \prec \boldsymbol{y}$; hence the existence of an upward path from $\boldsymbol{x}$ to $\boldsymbol{y}$ implies $\boldsymbol{x} \preceq \boldsymbol{y}$ (Ganter & Wille, 1999; Davey & Priestley, 2002). Given $\boldsymbol{x}, \boldsymbol{y} \in P$, a *meet* $\boldsymbol{x} \sqcap \boldsymbol{y}$ (greatest lower bound) and a *join* $\boldsymbol{x} \sqcup \boldsymbol{y}$ (least upper bound) may or may not exist. A *meet-semilattice* (*join-semilattice*) is a poset in which the meet $\boldsymbol{x} \sqcap \boldsymbol{y}$ (the join $\boldsymbol{x} \sqcup \boldsymbol{y}$) exists for all pairs $\boldsymbol{x}, \boldsymbol{y}$, and a *lattice* has both for all pairs.

If a rooted tree is ordered by the ancestor relation with the root $\boldsymbol{o}$ at the bottom (so that $\boldsymbol{o} \preceq \boldsymbol{x}$ for all $\boldsymbol{x}$), then the meet $\boldsymbol{x} \sqcap \boldsymbol{y}$ of a pair $\boldsymbol{x}, \boldsymbol{y}$ always exists, while joins need not exist. Hence, a rooted tree is naturally a meet-semilattice in this orientation. Conversely, an *is-a* taxonomy often uses the *entailment* order $\boldsymbol{x} \preceq \boldsymbol{y}$ interpreted as "$\boldsymbol{x}$ entails $\boldsymbol{y}$" or more roughly "$\boldsymbol{x}$ is more specific than $\boldsymbol{y}$," with the root $\boldsymbol{o}$ at the top. The join $\boldsymbol{x} \sqcup \boldsymbol{y}$ always exists, while meets need not exist; the poset is then a join-semilattice.

Let $\mathcal{C} = \{c_1, \ldots, c_n\}$ be n atomic concepts (e.g., dog, car, tomato,…). A subset $S \subseteq \mathcal{C}$ expresses the conjunction or co-occurrence of the concepts specified in S. We define the *Boolean lattice* $(2^{\mathcal{C}}, \subseteq)$ over the power set $2^{\mathcal{C}}$ of $\mathcal{C}$, in which the order relation $\preceq$ is the inclusion relation $\subseteq$. Meet/join are given by intersection/union, respectively. $T \subseteq S$ means that S specifies all concepts in T, so S entails T. Let $\chi : 2^{\mathcal{C}} \to \{0, 1\}^n$ be the indicator map with $\chi(S)_i = 1$ iff $c_i \in S$. Then, $T \preceq S$ iff $\chi(T)_i \leq \chi(S)_i$ for all i, and meet/join become bit-wise AND/OR, respectively. We summarize the correspondence between different representations in Table 5. In this lattice, each node is defined *intensionally* as a set of concepts.

From the dual perspective, each node can be defined *extensionally* as a set of instances that contain specified concepts, in the context of formal concept analysis (Ganter & Wille, 1999). Let $\mathcal{Z}$ be a universe of instances and let $I \subseteq \mathcal{Z} \times \mathcal{C}$ be an incidence relation (i.e., $\boldsymbol{z}\, I\, c$ means that $\boldsymbol{z}$ has concept c). For $S \subseteq \mathcal{C}$, define an operation $S' = \{\boldsymbol{z} \in \mathcal{Z} \mid \boldsymbol{z}\, I\, c \text{ for all } c \in S \subseteq \mathcal{C}\}$, which forms a Galois connection: $S \subseteq T$ implies $T' \subseteq S'$. Also, subsets $S' \subseteq \mathcal{Z}$ form the dual lattice of $(2^{\mathcal{C}}, \subseteq)$, where $S \subseteq T \Leftrightarrow S \preceq T \Rightarrow T' \subseteq S' \Leftrightarrow T' \preceq S'$. If $S = S''$ for any subset $S \subseteq \mathcal{C}$, $S \subseteq T \Leftrightarrow T' \subseteq S'$.

An *is-a* taxonomy is typically realized as a join-subsemilattice of this dual lattice. Order embeddings (Vendrov et al., 2016) can be regarded as an extension of the Boolean lattice, where each bit $\{0, 1\}$ is replaced with a real number $\mathbb{R}$. They declare "$\boldsymbol{x}$ entails $\boldsymbol{y}$" iff $x_i \geq y_i$ for all i, similarly to the indicators of a Boolean lattice. Indeed, the ambient poset $(\mathbb{R}^n, \preceq)$ of order embeddings is a lattice with meet/join given by coordinate-wise $\max$/$\min$, respectively. When regarding an embedding $\boldsymbol{x}$ as an orthant $U(\boldsymbol{x}) \subseteq \mathbb{R}^n$, the entailment is represented as $U(\boldsymbol{x}) \subseteq U(\boldsymbol{y})$, similarly to the dual lattice. When treating the orthant $U(\boldsymbol{y})$ as the set of all instances that contain the specified concepts

Table 5: Correspondence of generalization, specialization, and entailment in different representations.

	Generalization (hypernymy)	Specialization (hyponymy)	Space	Entailment ($\boldsymbol{x}$ or S entails $\boldsymbol{y}$ or T)
Tree of *is-a* Relations (*is-a* Taxonomy)	join $\sqcup$	(meet $\sqcap$)	T	$\boldsymbol{x} \preceq \boldsymbol{y}$
Order Embedding (as points)	min	max	$\mathbb{R}^n$	$x_i \geq y_i$ for all i
Order Embedding (as orthants)			orthants in $\mathbb{R}^n$	$U(\boldsymbol{x}) \subseteq U(\boldsymbol{y})$
Order Embedding (for entailment)			orthants in $\mathbb{R}^n$	$\boldsymbol{x} \in U(\boldsymbol{y})$
Hyperbolic Entailment Cone	(union $\cup$)	intersection $\cap$	cones in $\mathbb{H}^n$	$\boldsymbol{x} \in C(\boldsymbol{y})$
Boolean Lattice (as a power set)	intersection $\cap$	union $\cup$	$2^{\mathcal{C}}$	$S \supseteq T$
Boolean Lattice (as a lattice)	meet $\sqcap$	join $\sqcup$		$S \succeq T$
Boolean Lattice (with indicator)	AND	OR	$\{0,1\}^{\lvert\mathcal{C}\rvert}$	$\chi(S)_i \geq \chi(T)_i$ for all i
Dual Lattice (as a set)	union $\cup$	intersection $\cap$		$S' \subseteq T'$
Dual Lattice (as a lattice)	join $\sqcup$	meet $\sqcap$		$S' \preceq T'$
Product of Trees	join $\sqcup$	(meet $\sqcap$)	$\prod_{i=1}^k T_i$	$\boldsymbol{x}^{(i)} \preceq \boldsymbol{y}^{(i)}$ for all i
PHyCLIP	(union $\cup$)	intersection $\cap$	cones in $(\mathbb{H}_i^d)^k$	$\boldsymbol{x}^{(i)} \in C_i(\boldsymbol{y}^{(i)})$ for all i

$\boldsymbol{y}$, the entailment is represented as $\boldsymbol{x} \in U(\boldsymbol{y})$, which aligns with the definition of the dual lattice. Hyperbolic entailment cones (Ganea et al., 2018a) are a hyperbolic extension of the last interpretation of order embeddings, where an orthant $U(\boldsymbol{y})$ is replaced with a geodesic conical region $C(\boldsymbol{y})$.

Also, our proposed PHyCLIP can be regarded as an extension of a Boolean lattice, where each bit $\{0,1\}$ is replaced with a metric tree T_i, which is embedded into a hyperbolic factor $\mathbb{H}_i^d$.

B Propositions, Theorems, and Proofs

B.1 Proof of Proposition 1

Let $(2^{\mathcal{C}}, \subseteq)$ be a Boolean lattice over all subsets of atomic concepts $\mathcal{C} = \{c_1, \ldots, c_n\}$. The indicator χ maps subsets $S, T \subseteq \mathcal{C}$ to binary sequences $\chi(S), \chi(T) \in \{0,1\}^n$, where $\chi(S)_i = 1$ if $c_i \in S$ and $\chi(S)_i = 0$ otherwise. Then, $S \subseteq T$ iff $\chi(S)_i \leq \chi(T)_i$ for all i. The Hamming distance d_{Ham} is defined as $d_{\mathrm{Ham}}(\chi(S), \chi(T)) = \sum_{i=1}^n |\chi(S)_i - \chi(T)_i|$. Consider a map $f : \{0,1\}^n \to \mathbb{R}^n, \chi(S) \mapsto \boldsymbol{x} = (x_1, \ldots, x_n) = (1 - \chi(S)_1, \ldots, 1 - \chi(S)_n)$ and the product order $\boldsymbol{x} \preceq \boldsymbol{y}$ iff $x_i \geq y_i$ for all i on $\mathbb{R}^n$. Then, the map $f \circ \chi$ embeds the Boolean lattice $(2^{\mathcal{C}}, \subseteq)$ into the poset $(\mathbb{R}^n, \preceq)$ used by order embeddings while preserving the order relations.

By definition, the Hamming distance is $d_{\mathrm{Ham}}(\chi(S), \chi(T)) = \|\chi(S) - \chi(T)\|_1 = \sum_{i=1}^n |\chi(S)_i - \chi(T)_i|$. Hence, the metric space $(\{0,1\}^n, d_{\mathrm{Ham}})$ is equivalent to an ℓ_1-product metric space $(\{0,1\}^n, \sum_{i=1}^n |\cdot|)$. Consider a map f_i that maps 0 to the base point of the metric space X_i and 1 to a point with a finite non-zero distance $1/\tau_i > 0$ from the base point. The map f_i is an isometric embedding from $\{0,1\}$ to X_i after scaled by τ_i. The map $f = (f_1, \ldots, f_n)$ is an isometric embedding from $(\{0,1\}^n, d_{\mathrm{Ham}})$ to $(\prod_{i=1}^k \tau_i X_i, \sum_{i=1}^k d_{X_i})$ for any $k \geq n$.

Assume by contradiction that an isometric embedding $f : (\{0,1\}^n, d_{\mathrm{Ham}}) \to (\mathbb{H}^d, d_{\mathbb{H}^d})$ exists for some $n \geq 2$ and $d \geq 2$. Take four points

$$A = (0,0,0,\ldots), \quad B = (1,0,0,\ldots), \quad C = (1,1,0,\ldots), \quad D = (0,1,0,\ldots).$$

in $\{0,1\}^n$. Let $\boldsymbol{a} = f(A)$, $\boldsymbol{b} = f(B)$, $\boldsymbol{d} = f(D)$, and $\boldsymbol{c} = f(C)$. Since f is an isometric embedding,

$$d_{\mathbb{H}^d}(\boldsymbol{a}, \boldsymbol{b}) = d_{\mathbb{H}^d}(\boldsymbol{b}, \boldsymbol{c}) = d_{\mathbb{H}^d}(\boldsymbol{a}, \boldsymbol{d}) = d_{\mathbb{H}^d}(\boldsymbol{d}, \boldsymbol{c}) = 1$$

and

$$d_{\mathbb{H}^d}(\boldsymbol{a}, \boldsymbol{b}) + d_{\mathbb{H}^d}(\boldsymbol{b}, \boldsymbol{c}) = d_{\mathbb{H}^d}(\boldsymbol{a}, \boldsymbol{d}) + d_{\mathbb{H}^d}(\boldsymbol{d}, \boldsymbol{c}) = d_{\mathbb{H}^d}(\boldsymbol{a}, \boldsymbol{c}) = 2.$$

In a hyperbolic space, a geodesic segment is unique, and its midpoint is unique, so both $\boldsymbol{b}$ and $\boldsymbol{d}$ are placed at the midpoint in the geodesic segment $[\boldsymbol{a}, \boldsymbol{c}]$; hence $\boldsymbol{b} = \boldsymbol{d}$. See Proposition I.4 in Bridson & Haefliger (1999). This contradicts the assumption that f is an isometric embedding (which is injective).

B.2 Proposition 2 and its Proof

Proposition 2 (ℓ_1-product of trees is not hyperbolic)**.** *Let T_1, T_2 be infinite metric trees with known bounds for maximum degree and minimum edge length. Their ℓ_1-product metric space $(T_1 \times T_2, d_{T_1} +$*

d_{T_2}) *is not δ-hyperbolic for any finite δ. Consequently, there is no (λ, c)-quasi-isometric embedding* $(T_1 \times T_2, d_{T_1} + d_{T_2}) \to \mathbb{H}^n$.

A quasi-geodesic q in X is a (λ, c)-quasi-isometric embedding $q : I \to X$, where I is an interval in $\mathbb{R}$ or the intersection of $\mathbb{Z}$ with such an interval; see Definition I.8.22 in Bridson & Haefliger (1999). In a δ-hyperbolic space Y, the stability of quasi-geodesics asserts that the Hausdorff distance between a geodesic γ and a (λ, c)-quasi-geodesic q with common endpoints is bounded by a constant $D = D(\lambda, c, \delta)$; see Theorem III.1.7 in Bridson & Haefliger (1999).

Lemma 1 (Stability of geodesic triangles under quasi-isometric embeddings)**.** *Let X be a geodesic metric space and $f : X \to Y$ be a (λ, c)-quasi-isometric embedding into a δ-hyperbolic space Y. Then every geodesic triangle in X is $\tilde{\delta}$-slim for some constant $\tilde{\delta} \leq \lambda(\delta + 2D + c)$, where $D = D(\lambda, c, \delta)$ is the quasi-geodesic stability constant in Y.*

Proof. Let Δ be a geodesic triangle in X. Each side maps to a (λ, c)-quasi-geodesic in Y. By the stability of quasi-geodesics, each image side is contained in D-neighborhood of the corresponding geodesic. Geodesic triangles in Y are δ-slim; hence, each point on one image side is contained in $\delta + 2D$-neighborhood of the union of the other two image sides. Pulling this back via the quasi-isometry inequalities yields the stated bound. □

Let T_1, T_2 be infinite trees with bounds for maximum degree and minimum edge length, which admit a geodesic ray of infinite length. For simplicity, we restrict the edge length to be 1, but the following discussion holds for arbitrary non-zero edge lengths, by replacing $\mathbb{N}$ with the ordered set of the geodesic distances from the root to the nodes in the geodesic ray.

Consider $(\mathbb{N}^2, \|\cdot\|_1)$. Let m be an even integer and take three points $A = (0, 0)$, $B = (m, 0)$, $C = (0, m)$. The midpoint $(\frac{m}{2}, \frac{m}{2})$ of a monotone geodesic from B to C is at $\frac{m}{2}$ from $[A, B] \cup [A, C]$, requiring $\delta \geq m/2$. $\delta \to \infty$ as $m \to \infty$. Hence, $(\mathbb{N}^2, \|\cdot\|_1)$ is not δ-hyperbolic for any finite δ.

Choose two geodesic rays $\gamma_i : \mathbb{N} \to T_i$ for $i = 1, 2$. The map $\Phi : \mathbb{N}^2 \to T_1 \times T_2$, $\Phi(m, n) = (\gamma_1(m), \gamma_2(n))$ is an isometric embedding from $(\mathbb{N}^2, \|\cdot\|_1)$ into $(T_1 \times T_2, d_{T_1} + d_{T_2})$. Given a $\tilde{\delta}$-slim geodesic triangle Δ in $\mathbb{N}^2$, its image $\Phi(\Delta)$ is also a $\tilde{\delta}$-slim geodesic triangle in $T_1 \times T_2$. Since $(\mathbb{N}^2, \|\cdot\|_1)$ is not δ-hyperbolic, neither is $(T_1 \times T_2, d_{T_1} + d_{T_2})$.

Assume by contradiction that $f : (T_1 \times T_2, d_{T_1} + d_{T_2}) \to \mathbb{H}^n$ is a (λ, c)-quasi-isometric embedding, where $\mathbb{H}^n$ is δ-hyperbolic for a finite δ. By Lemma 1, every geodesic triangle in $T_1 \times T_2$ is $\tilde{\delta}$-slim, where $\tilde{\delta} \leq \lambda(\delta + 2D + c)$ and $D = D(\lambda, c, \delta)$ are constants. However, $(T_1 \times T_2, d_{T_1} + d_{T_2})$ is not $\tilde{\delta}$-hyperbolic for any finite $\tilde{\delta}$, which contradicts the assumption. Therefore, there is no (λ, c)-quasi-isometric embedding $f : (T_1 \times T_2, d_{T_1} + d_{T_2}) \to \mathbb{H}^n$.

B.3 Proof of Theorem 2

Lemma 2 (Product of quasi-isometric embeddings)**.** *If $f_i : (X_i, d_{X_i}) \to (Y_i, d_{Y_i})$ are (λ_i, c_i)-quasi-isometric embeddings, then*

$$f = \prod_{i=1}^k f_i : \left(\prod_{i=1}^k X_i,\ \sum_{i=1}^k d_{X_i}\right) \longrightarrow \left(\prod_{i=1}^k Y_i,\ \sum_{i=1}^k d_{Y_i}\right) \tag{8}$$

is (λ, c)-quasi-isometric with $\lambda = \max_i \lambda_i$ and $c = \sum_i c_i$.

Proof of Lemma 2. Sum the index-wise inequalities and bound λ by $\max_i \lambda_i$. □

Theorem 2 follows immediately from Theorem 1 and Lemma 2.

C Implementation Details

C.1 Lorentz Model of Hyperbolic Space

Let $\mathbb{R}^{d,1}$ be the $(d+1)$-dimensional Minkowski space, equipped with the Minkowski metric $g_{\mathbb{R}^{d,1}} = -\mathrm{d}x_0^2 + \mathrm{d}x_1^2 + \cdots + \mathrm{d}x_d^2$ in coordinates $\hat{\boldsymbol{x}} = (x_0, x_1, \ldots, x_d)$. Intuitively, x_0 denotes the time

coordinate, and the others $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ denote the space coordinates. The inner product in $\mathbb{R}^{d,1}$ is given by

$$\langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle_{\mathbb{R}^{d,1}} = -x_0 y_0 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{R}^d}. \tag{9}$$

For $\alpha > 0$, define the upper sheet of the two-sheeted hyperboloid as $\mathbb{L}_\alpha^d = \{\hat{\boldsymbol{x}} \in \mathbb{R}^{d,1} \mid \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} \rangle_{\mathbb{R}^{d,1}} = -\alpha^{-1},\ x_0 > 0\}$. Equivalently, every point satisfies $x_0 = \sqrt{\alpha^{-1} + \|\boldsymbol{x}\|_{\mathbb{R}^d}^2}$. The Riemannian metric on $\mathbb{L}_\alpha^d$ is the restriction of the Minkowski metric $g_{\mathbb{R}^{d,1}}$ to $T\mathbb{L}_\alpha^d$; with this metric, the sectional curvature is the constant $-\alpha$ (Cannon et al., 1997; Lee, 2018). The geodesic distance is

$$d_{\mathbb{L}_\alpha^d}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \alpha^{-1/2} \operatorname{arccosh}(-\alpha \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle_{\mathbb{R}^{d,1}}) \quad \text{for } \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \mathbb{L}_\alpha^d. \tag{10}$$

Then, a d-dimensional hyperbolic space $\mathbb{H}_\alpha^d$ with a curvature $-\alpha$ is isometrically embedded into $\mathbb{L}_\alpha^d$ by

$$\iota : \mathbb{H}_\alpha^d \to \mathbb{L}_\alpha^d, \boldsymbol{x} \mapsto \hat{\boldsymbol{x}} = (\sqrt{\alpha^{-1} + \|\boldsymbol{x}\|_{\mathbb{R}^d}^2}, \boldsymbol{x}), \tag{11}$$

and we denote $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}_\alpha^d} = \langle \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \rangle_{\mathbb{L}_\alpha^d}$ and $d_{\mathbb{H}_\alpha^d}(\boldsymbol{x}, \boldsymbol{y}) = d_{\mathbb{L}_\alpha^d}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ in the main body.

When feature extractors (such as encoders) operate in the Euclidean space $\mathbb{R}^d$, their output cannot be treated directly as an embedding $\boldsymbol{x}$ in a hyperbolic space due to the mismatch in geometry. Instead, the output $\boldsymbol{v} = (v_1, \ldots, v_d)$ is treated as a tangent vector in the tangent space $T_{\hat{\boldsymbol{o}}}\mathbb{L}_\alpha^d \simeq \mathbb{R}^d$ at the base point $\hat{\boldsymbol{o}} = (\alpha^{-1/2}, 0, \ldots, 0)$ of $\mathbb{L}_\alpha^d$ and mapped to a point in $\mathbb{L}_\alpha^d$ via the exponential map

$$\exp_{\hat{\boldsymbol{o}}}^\alpha : T_{\hat{\boldsymbol{o}}}\mathbb{L}_\alpha^d \to \mathbb{L}_\alpha^d, \boldsymbol{v} \mapsto \hat{\boldsymbol{x}} = \exp_{\hat{\boldsymbol{o}}}^\alpha(\boldsymbol{v}) = \cosh(\sqrt{\alpha}\|\boldsymbol{v}\|_{\mathbb{R}^d})\hat{\boldsymbol{o}} + \frac{\sinh(\sqrt{\alpha}\|\boldsymbol{v}\|_{\mathbb{R}^d})}{\sqrt{\alpha}\|\boldsymbol{v}\|_{\mathbb{R}^d}}\boldsymbol{v}. \tag{12}$$

C.2 Hyperbolic Entailment Cones in the Lorentz Model

Hyperbolic entailment cones capture the hierarchical relationships (Ganea et al., 2018a). For every point $\boldsymbol{y}$ in each hyperbolic factor $\mathbb{H}^d$, we define a geodesic conical region $C(\boldsymbol{y})$, where all points $\boldsymbol{x} \in C(\boldsymbol{y})$ are considered more specific than $\boldsymbol{y}$ (i.e., $\boldsymbol{x} \preceq \boldsymbol{y}$). The size of this conical region is determined by its half-aperture $\omega(\boldsymbol{y})$, which is inversely proportional to the norm:

$$\omega(\boldsymbol{y}) = \sin^{-1}\left(\min\left\{1, \frac{2K}{\sqrt{\alpha}\|\boldsymbol{y}\|_{\mathbb{R}^d}}\right\}\right), \tag{13}$$

where K is set to 0.1. Then, $\boldsymbol{x} \in C(\boldsymbol{y})$ iff $\phi(\boldsymbol{x}, \boldsymbol{y}) < \omega(\boldsymbol{y})$ for the exterior angle

$$\phi(\boldsymbol{x}, \boldsymbol{y}) = \cos^{-1}\left(\frac{x_0 + y_0 \alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}_\alpha^d}}{\|\boldsymbol{y}\|_{\mathbb{R}^d}\sqrt{(\alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}_\alpha^d})^2 - 1}}\right) \tag{14}$$

C.3 Model Architecture and Hyperparameters

We introduce the details of our implementation and hyperparameters, which follow Desai et al. (2023); Pal et al. (2025) unless specified otherwise.

As an image encoder, we employ the Vision Transformer (Dosovitskiy et al., 2021; Chen et al., 2021; Touvron et al., 2021) with a patch size of 16. Each image is randomly resized by a scale from 0.5 to 1.0 and randomly cropped to 224×224 pixels, resulting in 196 tokens, concatenated with 2-D sine–cosine position embeddings. We employ the text encoder used by the original CLIP (Radford et al., 2021), which consists of a 12-layer Transformer architecture (Vaswani et al., 2017) with embeddings of 512 dimensions.

The outputs of image and text encoders are scaled by learnable scalars c_{img} and c_{txt}, respectively, before being mapped by the exponential map. These scalars are initialized to $c_{\text{img}} = c_{\text{txt}} = 1/\sqrt{512}$. The negative curvature α_i for factor i is initialized at 1.0 and clamped in $[0.1, 10.0]$. For the contrastive loss $\mathcal{L}_{\text{cont}}$ in Eq. (3), the temperature τ is initialized to 0.07 and clipped at a minimum value of 0.01. For the entailment loss $\mathcal{L}_{\text{ent}}$ in Eq. (5), the hyperparameter η is set to $\eta = 0.7$ for inter-modality entailments ($I \preceq T$ and $I^{\text{box}} \preceq T^{\text{box}}$) and $\eta = 1.2$ for intra-modality entailments ($T \preceq I^{\text{box}}$ and $T \preceq T^{\text{box}}$). These scalars are learned on a logarithmic scale.

The hyperparameter γ for the overall loss in Eq. (1) is set to $\gamma = 0.2$. We trained each model on 4 A100 GPUs for 500,000 iterations with a batch size of 768. For the large Vision Transformer, we used 8 A100 GPUs. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with hyperparameters $\beta_1 = 0.9, \beta_2 = 0.98$. We applied weight decay of 0.2 to model parameters but not to scalar parameters. We used a cosine learning rate scheduler (Loshchilov & Hutter, 2017) with a maximum learning rate of 5×10^{-4} and a warm-up of 4,000 steps.

C.4 BENCHMARKS

Zero-shot Image Classification. We follow the protocol in Desai et al. (2023). Each class is accompanied by a set of short text templates, such as "a photo of a {class name}". The prediction is made by selecting the class whose text templates are closest on average to the image in the embedding space. We summarize the datasets below.

- **ImageNet** (Russakovsky et al., 2015): A large-scale dataset of diverse, everyday object categories.
- **Food-101** (Bossard et al., 2014): A fine-grained dataset of 101 different types of food dishes.
- **CIFAR-10** (Krizhevsky & Hinton, 2009): A dataset of low-resolution natural images across 10 general object classes.
- **CIFAR-100** (Krizhevsky & Hinton, 2009): Similar to CIFAR-10, but with 100 fine-grained object classes.
- **CUB-2011** (Wah et al., 2011): A fine-grained dataset for the identification of 200 bird species.
- **SUN397** (Xiao et al., 2010): A large-scale scene recognition dataset with 397 scene categories.
- **Stanford Cars** (Krause et al., 2013): A fine-grained dataset of cars, annotated with make, model, and year.
- **FGVC-Aircraft** (Maji et al., 2013): A fine-grained dataset for aircraft model recognition.
- **DTD** (Cimpoi et al., 2014): The Describable Textures Dataset for texture recognition.
- **Oxford-IIIT Pets** (Parkhi et al., 2012): A fine-grained dataset of 37 different pet breeds.
- **Caltech-101** (Fei-Fei et al., 2004): One of the classic object recognition datasets with 101 categories.
- **Flowers-102** (Nilsback & Zisserman, 2008): A fine-grained dataset for the classification of 102 flower categories.
- **STL-10** (Coates et al., 2011): An image recognition dataset inspired by CIFAR-10, but with higher resolution.
- **EuroSAT** (Helber et al., 2019): A dataset of Sentinel-2 satellite images for land use and land cover classification.
- **RESISC45** (Cheng et al., 2017): A benchmark for Remote Sensing Image Scene Classification (RESISC).
- **Country211** (Radford et al., 2021): A dataset for predicting the country of origin from photographs.

Zero-shot Image and Text Retrieval In text-to-image retrieval, given a text query, the model retrieves the nearest images in the embedding space, and vice versa in image-to-text retrieval. Please refer to the detailed protocol in Desai et al. (2023). We summarize the datasets used as follows.

- **COCO** (Lin et al., 2014): A large-scale dataset of complex everyday scenes with rich annotations.
- **Flickr30K** (Young et al., 2014; Karpathy & Fei-Fei, 2015): A dataset of images from the Flickr website, each paired with five descriptive captions.

Hierarchical Classification. This task was introduced in Russakovsky et al. (2015), and we used the implementation in Pal et al. (2025). The class labels are enriched by WordNet (Miller, 1995), and the embeddings of class labels are obtained in the same way as the zero-shot image classification task. Errors between predicted and true classes are measured using the WordNet graph with unit-length edges. Tree Induced Error (TIE) is the distance between the nodes corresponding to predicted and true

Table 6: Results with different model sizes.

		w/ boxes	Hierarchical Classification					VL-CheckList–Object					
			WordNet					Location			Size		
			TIE(↓)	LCA(↓)	J(↑)	P_H(↑)	R_H(↑)	Center	Mid	Margin	Large	Medium	Small
ViT S/16	CLIP		4.127	2.434	0.7526	0.8295	0.8304	64.6	65.7	61.2	66.3	63.2	62.5
	MERU		4.201	2.435	0.7479	0.8273	0.8256	63.5	60.5	59.6	63.2	61.9	61.2
	HyCoCLIP	✓	**3.637**	**2.209**	**0.7831**	**0.8528**	**0.8507**	67.2	66.1	65.3	69.1	65.0	65.8
	PHyCLIP	✓	3.715	2.241	0.7778	0.8492	0.8476	**70.4**	**69.5**	**70.8**	**72.8**	**67.0**	**69.7**
ViT B/16	CLIP		3.705	2.254	0.7805	0.8498	0.8503	67.1	65.5	64.3	69.9	63.8	64.2
	CLIP	✓	3.720	2.265	0.7797	0.8487	0.8509	66.1	61.6	64.7	67.0	64.6	63.3
	MERU		3.832	2.292	0.7720	0.8451	0.8439	63.3	60.0	60.5	66.6	57.3	58.6
	MERU	✓	3.793	2.277	0.7740	0.8462	0.8454	62.6	58.3	59.8	62.6	60.3	59.8
	HyCoCLIP	✓	3.378	2.113	0.8008	0.8653	0.8636	65.9	65.6	63.1	67.6	63.1	63.9
	PHyCLIP	✓	**3.285**	**2.088**	**0.8065**	**0.8684**	**0.8682**	**73.0**	**72.0**	**71.4**	**76.4**	**69.2**	**69.0**
ViT L/16	CLIP		3.475	2.158	0.7957	0.8605	0.8607	64.2	60.7	60.2	64.6	61.2	58.0
	MERU		3.558	2.178	0.7891	0.8574	0.8553	58.9	56.3	55.3	61.0	56.4	54.0
	HyCoCLIP	✓	3.100	2.007	0.8179	0.8770	0.8751	73.9	71.2	**70.9**	**75.3**	69.3	70.1
	PHyCLIP	✓	**3.044**	**1.993**	**0.8223**	**0.8795**	**0.8790**	**74.3**	**72.7**	70.5	75.1	**70.5**	**70.8**

Among methods with the same backbone, the best and second performances are emphasized by bold fonts and underlines, respectively.

classes. Lowest Common Ancestor (LCA) error is the maximum of the distances from predicted and true classes to their LCA. Jaccard similarity J, hierarchical precision P_H, and hierarchical recall R_H are similarities between the sets of ancestors of predicted and true classes. Intuitively, hierarchical precision P_H quantifies correctness under over-generalization: it takes value 1 if the predicted label is the ground truth or one of its ancestors in the taxonomy. Conversely, hierarchical recall R_H quantifies correctness under over-specialization: it takes value 1 if the predicted label is the ground truth or one of its descendants.

Compositional Understanding. Samples in typical multi-modal datasets are diverse enough that there are few near-duplicate image–text pairs; consequently, models insensitive to detailed semantics can still perform well on retrieval tasks. To assess whether a model truly understands the compositionality of words in a caption, hard negative captions are generated, which are almost correct but differ in a small, targeted way and evaluate whether models can select the true caption.

In VL-CheckList–Object, nouns in the caption are replaced. Because the difficulty varies with the replaced object's location (center/mid/margin) and size (small/medium/large) in image, results are reported separately for each subset.

In SugarCrepe, three operations (replace, swap, and add) are applied to objects, attributes, and relations. *Replace-Obj* is similar to VL-CheckList–Object. *Swap* exchanges roles or pairings. In *Swap-Obj*, the model must correctly resolve agent–action combinations. *Add* introduces nouns or adjectives that were absent from the original caption.

D Additional Results and Visualizations

D.1 Additional Experimental Results

We obtained results with the small and large Vision Transformers as the image encoder (Dosovitskiy et al., 2021; Chen et al., 2021; Touvron et al., 2021) in Table 6. As the model size increases, the overall performance improves in most cases. Nevertheless, PHyCLIP remains the best or at least competitive across all evaluation metrics for hierarchy and compositionality.

D.2 Additional Visualizations

In this section, we provide additional visualizations that complement Fig. 4 in Section 4.3. We embed each word using the single-concept prompt "a photo of a {word}" and the composition of two words using the conjunctive prompt "a photo of a {word 1} and a {word 2}."

Figure 5 is a larger version of Fig. 4 (b) with labels, where HoroPCA (Chami et al., 2021) projects embeddings in $d = 8$-dimensional hyperbolic factors onto 2D disks. Embeddings of mammal-related terms are spread over a wide area in factor $i = 39$. Dog-related terms cluster in the left half, cat-related terms in the right, and "chihuahua," "corgi," and "puppy" are positioned farther from the

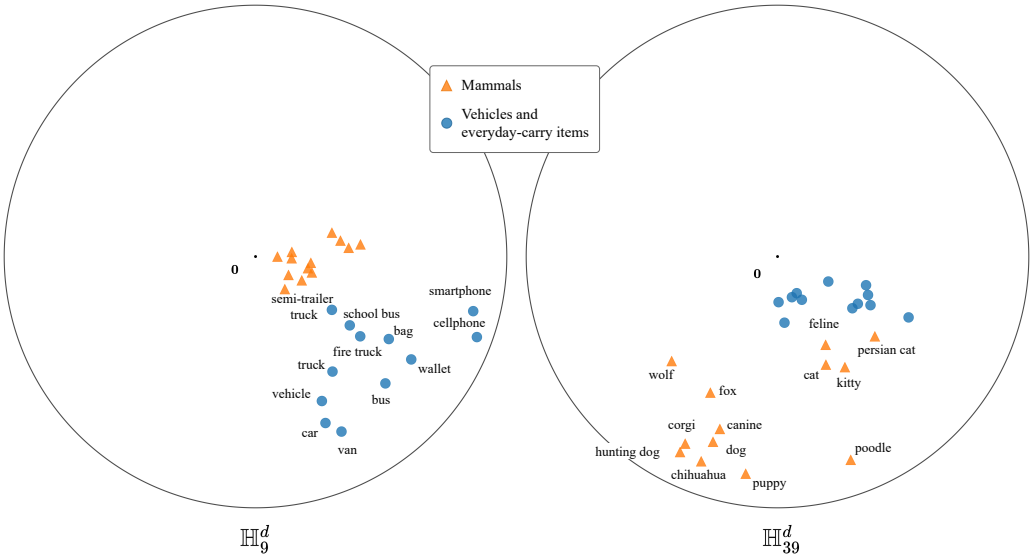
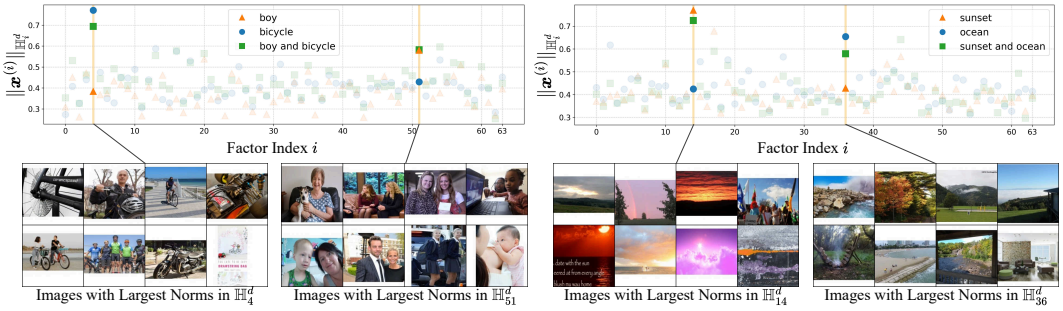


Figure 5: **Embeddings projected onto 2D disks by HoroPCA.** A set of relevant concepts (hyponyms of mammals or words related to vehicles and everyday-carry items) forms a hierarchical structure in the corresponding factor ($i = 39$ or $i = 9$), while the same concepts cluster near the origin in another factor ($i = 9$ or $i = 39$).



(a) boy and bicycle

(b) sunset and ocean

Figure 6: **Visualization of embedding norms in hyperbolic factors.** See also Fig. 4

origin than "dog." These patterns indicate that factor $i = 39$ encodes a hierarchy of a concept family of mammals (or more specifically, Carnivora). In contrast, in factor $i = 9$, the same embeddings concentrate near the origin, suggesting that this factor does not capture mammals. Conversely, terms related to vehicles and everyday-carry items form a hierarchical arrangement in factor $i = 9$ but cluster near the origin in factor $i = 39$. Together, these observations indicate that distinct hyperbolic factors capture taxonomies of different concept families.

Figure 6 visualizes factor-wise embedding norms of single concepts and conjunctive prompts, complementing Fig. 4 (a). In Fig. 6 (a), the "boy" embedding activates factor $i = 51$, which is also activated by various human images, indicating that this factor captures humans; the "bicycle" embedding activates factor $i = 4$, associated with bicycles and wheels. The conjunctive prompt "boy and bicycle" activates both factors $i = 4$ and $i = 51$. In Fig. 6 (b), the "sunset" embedding activates factor $i = 14$, which captures a family of skies, whereas the "ocean" embedding activates factor $i = 36$, which captures a family of natural landscapes. The conjunctive prompt "sunset and ocean" activates both factors $i = 14$ and $i = 36$.

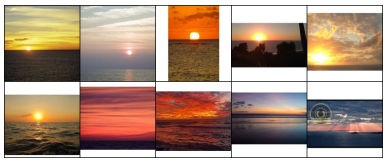
Figure 7 shows top-10 GRIT images retrieved using conjunctive prompts and the factor-wise "max" of single-concept prompts. Specifically, we embed two single-concept prompts (e.g., "a photo of

(a) conjunctive prompt “a dog and a car”

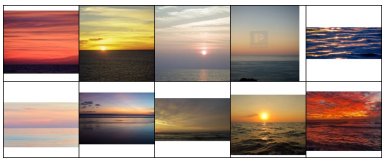
(b) “max” of single-concept prompts “a dog” and “a car”

(c) conjunctive prompt “a boy and a bicycle”

(d) “max” of single-concept prompts “a boy” and “a bicycle”



(e) conjunctive prompt “a sunset and an ocean”



(f) “max” of single-concept prompts “a sunset” and “an ocean”

Figure 7: **Retrieval results by conjunctive prompts and factor-wise “max” of single-concept prompts.** A new embedding is constructed by taking, for each factor, the factor-wise embedding with the larger norm between two single-concept prompts. The retrieval results are appropriate in both cases.

a dog” and “a photo of a car”) as $\boldsymbol{X}_a = (\boldsymbol{x}_a^{(1)}, \ldots, \boldsymbol{x}_a^{(k)})$ and $\boldsymbol{X}_b = (\boldsymbol{x}_b^{(1)}, \ldots, \boldsymbol{x}_b^{(k)})$, and then we construct a new embedding $\boldsymbol{X}_{\max\{a,b\}}$ by selecting, for each factor, the factor-wise embedding with the larger norm between two single-concept prompts, i.e., we take

$$\boldsymbol{X}_{\max\{a,b\}} = (\boldsymbol{x}_{\max\{a,b\}}^{(1)}, \ldots, \boldsymbol{x}_{\max\{a,b\}}^{(k)}) \text{ with } \boldsymbol{x}_{\max\{a,b\}}^{(i)} = \operatorname*{arg\,max}_{\boldsymbol{x} \in \{\boldsymbol{x}_a^{(i)}, \boldsymbol{x}_b^{(i)}\}} \|\boldsymbol{x}\|_{\mathbb{H}_i^d} \text{ for } i = 1, \ldots, k.$$

Then, the factor-wise norms satisfy $\|\boldsymbol{x}_{\max\{a,b\}}^{(i)}\|_{\mathbb{H}_i^d} = \max\{\|\boldsymbol{x}_a^{(i)}\|_{\mathbb{H}_i^d}, \|\boldsymbol{x}_b^{(i)}\|_{\mathbb{H}_i^d}\}$. If each factor were a bit $\{0, 1\}$, this operation would reduce to the union operation or the logical OR for a Boolean algebra. If each factor were a real number $\mathbb{R}$, it coincides with an element-wise $\max$, examined in order embeddings (Vendrov et al., 2016). The retrieval results by both methods are appropriate in most cases and often overlap. Concepts specified in the prompt are embedded with large norms in factors that capture their corresponding concept families, whereas unspecified concepts are represented with small norms. Consequently, by retaining only the high-norm factors, we can compose concepts without corrupting the semantics of the original prompts. These results suggest that PHyCLIP expresses cross-family composition in a manner analogous to Boolean algebra and order embeddings.

In conclusion, in PHyCLIP, different hyperbolic factors capture distinct concept families, and the ℓ_1-product metric represents cross-family composition through the simultaneous activation of multiple factors.

The Use of Large Language Models.

We used ChatGPT and GitHub Copilot as assistance tools for polishing the manuscript and implementing the experimental code. We did not use large language models for research ideation or for proofs.