

HyperVLM: Hyperbolic Space Guided Vision Language Modeling for Hierarchical Multi-Modal Understanding

Sarthak Srivastava*

Amazon

sarthasr@amazon.com

Kathy Wu*

Amazon

rhaow@amazon.com

Abstract

State-of-the-art performance has been achieved in recent years on tasks such as product search, recommendation, and classification using visuo-lingual multi-modal models. While pretrained vision-language models like CLIP have shown strong zero-shot capabilities by aligning vision and language in a shared space, they often fail to capture the natural hierarchical relationships common in real-world retail data. In this work, we propose HyperVLM: a vision-language model built on hyperbolic Poincaré geometry that learns joint image-text representations while explicitly modeling their hierarchical structure. We compare HyperVLM with CLIP on zero-shot image classification and retrieval tasks, highlighting its improved performance on tasks involving fine-grained category distinctions which is critical in large-scale retail environments. We also integrate our method into BLIP’s ITC loss module, showing enhanced retrieval accuracy. Our proposed approach holds immense value for recommendation and search systems in retail, where understanding complex product relationships and scalable retrieval is essential.

1. Introduction

Vision Language Models Large vision-language models like CLIP [34] and ALIGN [17] learn visual concepts from their natural language description via multi-modal contrastive learning. In contrastive learning [19], an anchor item representation is compared with a similar and a dissimilar item with the aim of bringing similar item representation together and pushing different ones away. The effectiveness [40] of these models results from their pretraining over a diverse large-scale image-text dataset sources from the web, allowing them to learn diverse concept from real world resulting in their impressive generalizability over

a variety of tasks in zero-shot setting like classification and retrieval. These models assume the geometry of the higher dimensional representation space as affine Euclidean [12, 29], making it harder to capture the visual-text hierarchical concepts. The entity containing more *general* concepts should be located close to the root of the hierarchy tree than the entity encapsulating a more *specific* and complex information. Hyperbolic spaces [3, 35] are natural candidate for capturing this hierarchical information about data points as their volume grows exponentially away from the origin, against polynomial growth in case of Euclidean space. Hyperbolic space can be thought of as a continuous version of a tree with its root at the origin.

Vision Language Hierarchy The saying ”A picture is worth a thousand words” conveys the information difference between an image and words describing them. For example, in Figure 1, the picture can be broken down into individual concepts consisting of ”kitty” and ”doggo”, which might be transformed in different manner to generate caption, for e.g. ’my dog’s innocence brings smile to my face’, ’a dog and a cat having fun in field’, etc. Following equivalence, a many words can be put together encapsulating complex concept to build an informative image. Injecting these inductive biases in the training of multi-modal models [34, 36] will allow them to learn a more generalizable and interpretable representation.

Hyperbolic Space Representation with HyperVLM In this work we project the image-text concepts onto a Poincaré ball model of hyperbolic space while following the state of the art contrastive methodology, to help capture the hierarchical information about the image-text pair, in addition to their semantic similarity. The contribution of this work can be described as: i. We introduce HyperVLM, a Poincaré ball based hyperbolic representation model trained using ViTs and Transformer encoder based contrastive loss using RedCap dataset containing 12M image-text pairs. ii. We introduce an embedding entropy based entailment loss to enforce the hierarchy between image-text in the

*Equal contribution.

Poincaré space. We compare the performance of the proposed method with strong baseline CLIP and MERU to demonstrate its competitiveness.

2. Related Work

2.1. Vision-Language Models

Vision-language models have seen remarkable progress in recent years, with contrastive learning emerging as a dominant paradigm for aligning visual and textual representations. CLIP [34] pioneered large-scale contrastive learning between images and text, demonstrating impressive zero-shot transfer capabilities across various downstream tasks. Following CLIP, several models have further advanced this approach: ALIGN [17] scaled up training to even larger and noisier datasets, while BLIP [22] introduced bootstrapping techniques to improve the quality of image-text pairs. More recently, models like FLAVA [38] and Florence [47] have incorporated additional pre-training objectives beyond contrastive learning to enhance multi-modal understanding.

Despite their success, these models predominantly operate in Euclidean space, which limits their ability to capture hierarchical relationships inherent in visual and textual concepts. Our work addresses this limitation by leveraging hyperbolic geometry to better represent these hierarchical structures.

2.2. Hyperbolic Representations

Hyperbolic geometry has gained attention in machine learning due to its ability to efficiently embed hierarchical structures [30, 35]. The Poincaré ball model, in particular, has been successfully applied to various domains including natural language processing [39], graph representation learning [2], and recommender systems [42].

In the context of vision, HypIR [9] applied hyperbolic geometry to image retrieval, demonstrating improved performance by capturing hierarchical relationships between images. MERU [8] extended this approach to multi-modal settings, using the Lorentz model of hyperbolic space to represent image-text pairs. However, as noted in [33], the Lorentz model has lower representation capacity compared to the Poincaré ball model, which motivates our choice of geometry in HyperVLM.

2.3. Hierarchical Multi-Modal Representations

Several approaches have attempted to capture hierarchical relationships in multi-modal data without explicitly using hyperbolic geometry. Hierarchical CLIP [48] introduced a hierarchical contrastive learning objective that considers class hierarchies during training. Similarly, HCSC [45] proposed a hierarchical contrastive learning framework for fine-grained visual categorization.

More closely related to our work, Order Embeddings

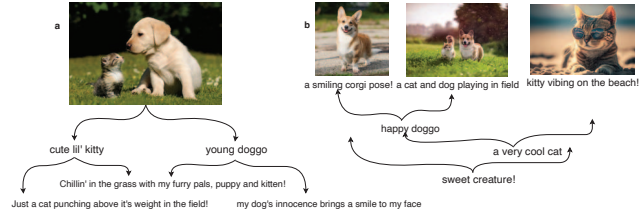


Figure 1. A picture is worth a thousand words. **Left:** Given an informative image it is possible to generate several textual concepts leveraging the visuo-lingual hierarchy. **Right:** Likewise, beginning from a simple text concept, it is possible to come up with complex visuo-lingual concepts by leveraging their hierarchical relation.

[41] and Hyperbolic Entailment Cones [13] proposed methods to model partial order relationships in embedding spaces. Our approach builds upon these ideas by introducing an entropy-based method to dynamically determine the entailment direction between image and text pairs, rather than making a fixed assumption about their hierarchical relationship.

3. Hyperbolic Geometry

In this section, we will walk through some key concepts of hyperbolic geometry which are relevant to our approach. Hyperbolic geometry, also known as Lobachevskian geometry is a non-Euclidean geometry where the Euclid's fifth postulate of parallels don't hold true and the space has a constant negative curvature. Hyperbolic spaces can be thought of as a continuous versions of tree data structure where the number of nodes until level h grow exponentially with the value l as $((b + 1)b^h - 2)/(b - 1)$ where b is the branching factor. This tree grows from origin where h is 0 and it grows in terms of nodes exponentially away from origin. Such a structural arrangement is not possible in \mathbb{R}^2 Euclidean space as the area and circumference of the hypercircle only grows quadratically and linearly respectively against an exponential growth in case of hyperbolic space. An intuitive overview on the key concepts of hyperbolic geometry is discussed in supplementary material.

3.1. Manifold

A manifold is a topological space that locally resembles Euclidean space. A precise definition from topology is that an n -dimensional manifold M is a topological Hausdorff space with a countable base which is locally homeomorphic to \mathbb{R}^n . For every point p in M , there exists an open neighbourhood U and a homeomorphism $h: U \rightarrow V$ which maps the set U onto an open set $V \subset \mathbb{R}^n$. Thus the point is either an isolated point (when $n = 0$), or it has a neighborhood which is homeomorphic to the open ball

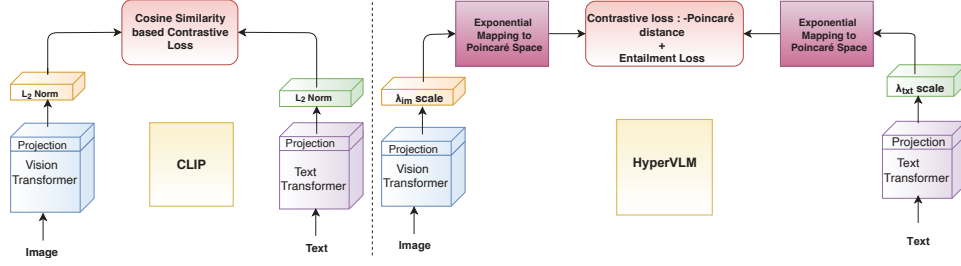


Figure 2. Overall Model Architecture. **Left:** Describes the baseline CLIP architecture based on which we have defined HyperVLM. Image and text are encoded by Vision and Text Transformers respectively before being normalized and compared for contrastive loss calculation **Right:** Describes HyperVLM architecture. It differs from CLIP in aspect that encoder output is scaled and projected onto Poincaré space before computing contrastive loss and entailment loss for optimization.

$$\mathbf{D}^n = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n : x_1^2 + x_2^2 + \dots + x_n^2 < 1\}$$

Riemannian manifold refers to real and smooth manifold with Riemannian tensor, which is metric tensor and can be defined by a family a inner products as follow:

Suppose p is a point on the curve of manifold M with $p \in M$ and denote the tangent space by $T_p(M) \in \mathbb{R}^n$, for any two tangent vectors $X(p)$ and $Y(p)$, $q : T_p M \times T_p M \rightarrow \mathbf{R}$ defines a smooth function for the point $p \in M$

3.2. Curvature

In simple terms, curvature of a curve is its measure of deviation from a straight line and that of a surface is the measure of its deviation from a plane. In terms of space, a curved space refers to spatial geometry which shows some finite curvature w.r.t a plane surface.

3.3. Hyperbolic Space

Hyperbolic n-space, denoted \mathbb{H}^n , is the unique simply connected, n-dimensional Riemannian manifold which has constantly negative sectional curvature. Let (H, d) denote a metric space, it is said to be a hyperbolic metric space if the following conditions are satisfied:

- for any points $p, q \in H$ that are the endpoints of a unique metric segment is isometric to the interval of real line $[0, d(p, q))$
- let the unique point $t = \alpha p \oplus (1 - \alpha)q$ where $\alpha \in [0, 1]$, it satisfies $dpt = 1 - \alpha dpq, dtq = \alpha dpq$ 3) for all $x, y, p, y \in H$ and $\beta \in [0, 1]$ we have $d\beta x \oplus 1 - \beta p, \beta y \oplus 1 - \beta q \leq \beta dx y + 1 - \beta dpq$

3.4. Poincaré Disk

A Poincaré disk is a hyperbolic geometric model in which we represent a line as an arc of a circle whose ends are perpendicular to the disk's diameter. It's a useful model that uses hyperbolic geometry to discover continuous hierarchical relations among data pairs by embedding them into n dimensional Poincaré hypersphere. Mathematically, we can

define an n-dimensional Poincaré ball in constant negative curvature value of $K = -1$ as:

$$\mathbb{P}_{K=-1}^n = \{x \in \mathbb{R}^n : \|x\|^2 < 1\} \quad (1)$$

where $\|\cdot\|$ represents the Euclidean norm of a data point. The metric tensor for a Poincaré ball is represented as $g_x^{K=-1} = (\gamma_x^{K=-1})^2 g_x^E$ where $\gamma_x^{K=-1} = \frac{1}{1-\|x\|^2}$ is the conformity factor and g_x^E is the metric tensor for Euclidean space represented as $g_x^E = \text{diag}([1, 1, \dots, 1])$. The distance $d_h(p_1, p_2)$ between two samples p_1 and p_2 in the Poincaré space $\mathbb{P}_{K=-k}^n$ is calculated as:

$$d_h(p_1, p_2) = \frac{2}{\sqrt{k}} \tanh^{-1}(\sqrt{k} \|(-p_1) \oplus_k p_2\|_2) \quad (2)$$

Where $\|\cdot\|$ represents the Euclidean norm of a data point. We map Euclidean feature into hyperbolic Poincaré ball manifold via $h_i = \text{exp}_0^{K=-1}(x_i^{Euc})$ where h_i represents the transformed x_i value in the hyperbolic space. The exponential map value exp_x^k for a vector p in a space having curvature value K is calculated as:

$$\text{exp}_x^K(p) = x \oplus_K \left(\tanh \left(\frac{\sqrt{-K} \gamma_x^K \|p\|}{2} \right) \frac{p}{\sqrt{-K} \|p\|} \right) \quad (3)$$

To reverse map a vector p from hyperbolic space of curvature value K to Euclidean space, we apply logarithmic mapping as following:

$$\log_x^K(p) = \frac{2}{\sqrt{-K} \gamma_x^K} \text{arctanh} \left(\sqrt{-K} \|v\| \right) \frac{v}{\|v\|} \quad (4)$$

Where v is calculated as $-x \oplus_K p$ and \oplus_K represents the Möbius addition defined as follow:

$$x \oplus_K y = \frac{(1 - 2K \langle x, y \rangle - K \|y\|^2) x + (1 + K \|x\|^2) y}{1 - 2K \langle x, y \rangle + K \|x\|^2 \|y\|^2} \quad (5)$$

Where $\langle x, y \rangle$ represents the inner product between x and y in hyperbolic space.

3.5. Why This Matters for Vision-Language Models

The key insight of our work is that vision-language relationships are inherently hierarchical. For example, the concept "animal" encompasses "dog," which encompasses "golden retriever." Similarly, an image might contain multiple concepts at different levels of specificity. By embedding both images and text in hyperbolic space, we can:

- Preserve hierarchical relationships between concepts
- Represent both general and specific concepts effectively
- Model the partial order relationships between modalities
- Capture the fact that some concepts entail others

This geometric approach allows our model to learn more nuanced relationships between visual and textual concepts than is possible in Euclidean space, leading to improved performance on a variety of tasks as demonstrated in our experiments.

4. Methodology

In this section we discuss the learning objective and modelling details of HyperVLM to learn the hierarchy aware representations for input text and images. HyperVLM is based on CLIP methodology consisting of a vision transformer based image encoder and a text transformer based text encoder using byte pair encoding. Both encoders generate image and text representations for input image and text respectively, which are then passed into a projection layer to obtain embeddings of a fixed size n . Additionally, we:

Transfer of embeddings onto the Poincaré Space

While training, the image and text samples are passed to ViT and Text Transformer encoders respectively followed by a projection layer as shown in Figure 2. This is followed by transformation of the embeddings (ν_{im}, ν_{txt}) from Euclidean geometry to hyperbolic Poincaré geometry as (h_{im}, h_{txt}) following the eq. 3 w.r.t the origin.

Numerical Overflow Prevention Since transfer from Euclidean space to hyperbolic space to calculate (h_{im}, h_{txt}) requires an exponential operation, the norm of embeddings changes from order of \sqrt{n} to $e^{\sqrt{n}}$, potentially causing numerical overflow. To fix this, embedding scaling is applied before exponential mapping via two learnable parameters λ_{im} and λ_{txt} initialized to $1/\sqrt{n}$ to prevent the norm of the embedding from numerical overflow in the Poincaré space.

Training Objectives Our training objective is to enforce semantic similarity and structural partial order relation between given image-text pairs to improve the generalization capability of vision-language models. To this end, we optimize for image-text contrastive loss and entailment loss.

4.1. Contrastive Loss

We have implemented same multi-class N-pair version of the contrastive loss as used in CLIP [34] with an important difference that we calculate the similarity via distances in Poincaré space from eq. 3 instead of cosine similarity. For a given batch size N we use the negative Poincaré space distance to compute contrastive loss between 1 positive and $N - 1$ negative pair per image and per text. The average of image wise and text wise loss is used as overall contrastive loss \mathcal{L}_{cont} to enforce image-text semantic similarity.

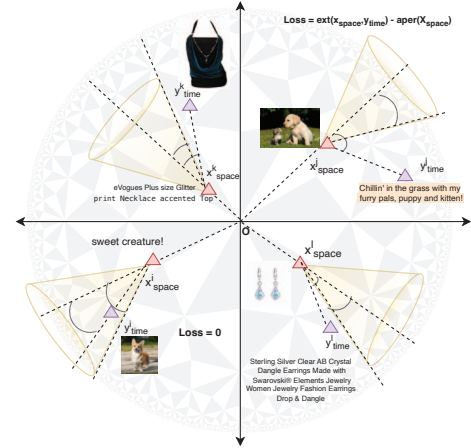


Figure 3. Entailment Cone (projection from Poincaré space on Euclidean Space). Loss pushes y_{time} embedding inside an entailment cone projected by embedding x and is defined as the difference between exterior angle $\angle OXY$, and half aperture of the cone. Loss is zero if the y_{time} is already inside the cone. Indices i and j in superscripts represent two different instances of image-text pairs.

4.2. Entailment Loss

We apply an additional entailment loss from [8] with modification to enforce partial order relationship between image-text pairs. In [8], the assumption is that text always entails the image within the entailment cone. In contrast, we adopt an entropy based strategy to determine correct entailment order between text and image per instance. In Physics, the structure of space-time is knitted together by the causal connections represented by the causal graph, the analog of entailment cone. An entailment cone is essentially a structure representing the “time evolution” from a particular initial condition [43]. Keeping this view in perspective and given that image-text embeddings from respective transformers are learned in same latent space, we can determine the relative position in entailment cone comparing the entropy of embeddings with the assumption that entropy increases with evolution of time along the entailment cone. For a given image-text pair, the simpler concept with lower entropy should be entailing more complex concept with higher entropy with time. We calculate the information entropy [16] of embeddings as:

$$H(x_{emb}) = - \sum_{i=1}^n x_i \log_2 x_i \quad (6)$$

where H is the entropy of embedding x_{emb} and x_i represents the content of size n embedding for i^{th} dimension.

Unlike previous approaches that assume a fixed direction of entailment, our method dynamically determines the

		CIFAR-10[20]	CIFAR-100[20]	CUB[44]	SUN397[46]	Aircraft[25]	DTD[5]	Pets[32]	Flowers[31]	STL-10[6]	EuroSAT[15]	RESISC45[4]	Country211[34]	MNIST[21]	PCAM1[10]	SST2[34]
ViT-S/16	CLIP	60.1	24.4	33.8	27.5	1.4	15.0	73.7	47.0	88.2	18.6	31.4	5.2	10.0	50.2	50.1
	MERU	52.0	24.7	33.7	28.0	1.3	16.2	72.3	49.2	91.1	30.4	32.0	4.8	7.5	51.0	50.0
	HyperVLM	53.6	27.7	35.1	27.6	1.6	17.6	71.9	47.9	90.9	30.8	32.1	5.1	10.4	53.8	50.8
ViT-B/16	CLIP	65.5	33.4	33.3	29.8	1.4	17.0	77.9	50.9	92.2	25.6	31.0	5.8	10.4	54.1	51.5
	MERU	67.7	32.7	34.8	30.9	1.7	17.2	79.3	52.1	92.5	30.2	34.5	5.6	13.0	49.8	49.9
	HyperVLM	70.4	35.4	34.9	31.3	2.1	17.9	78.5	51.3	91.9	31.7	33.5	5.5	12.1	49.6	50.0
ViT-L/16	CLIP	72.0	36.4	36.3	32.0	1.1	16.5	78.8	48.6	93.7	26.7	35.4	6.1	14.8	51.2	51.1
	MERU	68.7	35.5	37.2	33.0	2.2	17.2	80.0	52.1	93.7	28.1	36.5	6.2	11.8	52.7	49.3
	HyperVLM	74.3	38.8	37.5	33.3	2.6	18.5	80.1	51.3	93.8	27.9	37.2	6.5	12.0	55.7	50.0

Table 1. Comparison of Proposed Method HyperVLM vs Baseline Methods on different datasets. The metrics in color represent the best performance metric for particular dataset. We observe that HyperVLM outperforms all methods in 13 out of 18 datasets.

		<i>text</i> → <i>image</i>		<i>image</i> → <i>text</i>	
		R5	R10	R5	R10
ViT-S/16	CLIP	29.9	40.1	37.5	48.1
	MERU	30.5	40.9	39.0	50.5
	HyperVLM	30.5	40.2	40.4	50.7
ViT-B/16	CLIP	32.9	43.3	41.4	52.7
	MERU	33.2	44.0	41.8	52.9
	HyperVLM	33.3	43.7	42.1	53.4
ViT-L/16	CLIP	31.7	42.2	40.6	51.3
	MERU	32.6	43.0	41.9	53.3
	HyperVLM	32.6	42.7	43.2	53.8

Table 2. Zero Shot Image and Text Retrieval on COCO Dataset. Metric in color represent best performance for the task.

direction based on the relative entropy:

$$x = \begin{cases} x_{img}, & \text{if } H(x_{img}) < H(x_{txt}) \\ x_{txt}, & \text{otherwise} \end{cases} \quad (7)$$

$$y = \begin{cases} x_{txt}, & \text{if } H(x_{img}) < H(x_{txt}) \\ x_{img}, & \text{otherwise} \end{cases} \quad (8)$$

This adaptive approach allows our model to capture more nuanced hierarchical relationships that better reflect the true information content of each modality, rather than imposing a predetermined structure. Figure 3 gives an overview of the entailment loss as projected in Euclidean space. Exterior angle $\angle Oxy$ is defined as:

$$ext(\angle Oxy) = \arccos \left(\frac{\langle x, y \rangle (1 + \|x\|^2) - \|x\|^2 (1 + \|y\|^2)}{\|x\| \|y\| \sqrt{1 + \|x\|^2} \sqrt{1 + \|y\|^2} - 2\langle x, y \rangle} \right) \quad (9)$$

While the aperture of the entailment cone is defined as:

$$aper(x) = \arcsin \left(K \frac{1 - \|x\|^2}{\|x\|} \right) \quad (10)$$

We calculate the entailment loss as:

$$\mathcal{L}_{entail}(x, y) = \max(0, ext(\angle Oxy) - aper(x)) - \lambda_{reg} ext(\angle Oxy) \quad (11)$$

where λ_{reg} is the regularization coefficient. Hence, the overall loss to be optimized becomes $\mathcal{L} = \mathcal{L}_{cont} + \lambda \mathcal{L}_{entail}$ is λ is entailment regularization factor.

5. Experiments

Model Architecture We implement HyperVLM using different size versions of Vision Transformers (S/B/L) as vision encoders with a patch size of 16, freezing the positional encoding layer of the model. The text encoder follows the CLIP architecture with a 12-layer, 512-dimensional Transformer with 77 maximum length byte pair encoding. For the hyperbolic space representation, we use a Poincaré ball of 512 dimensions with learnable curvature K for Poincaré space transformation after embedding scaling.

Optimization We use the AdamW Optimizer [24] with weight decay of 0.2 and $(\beta_1, \beta_2) = (0.9, 0.98)$. Weight decay is disabled for all gains, biases, and learnable scalars. The model is trained for 120K iterations with a batch size of 1024 (≈ 10 epochs). The maximum learning rate is 5×10^{-4} , which increases linearly for the first 4K iterations, followed by cosine decay to 0 [23].

Evaluation Protocol We evaluate HyperVLM against CLIP and MERU on 18 diverse datasets for zero-shot classification and on the COCO dataset for retrieval tasks. Additionally, we evaluate image-text and text-image retrieval accuracy using the BLIP architecture [22] by implementing our ITC loss calculation module in Poincaré hyperbolic space with entropy-based image-text entailment order, comparing it with the Euclidean space BLIP architecture on the COCO dataset and Amazon Product Recommendation Clothes dataset [27] (see Table 6).

5.1. Results

From Table 1, we compare HyperVLM’s performance for zero shot classification and observe it performing better than the Euclidean space CLIP for 14 out of 18 datasets and

than Lorentz model based MERU for 15 out of 18 datasets and on 13 out of 18 datasets overall. The results demonstrate that HyperVLM consistently outperforms both baseline methods:

- **Compared to Euclidean CLIP:** HyperVLM achieves superior performance on 14 out of 18 datasets, with particularly notable improvements on CIFAR-100 (+3.3%), CUB (+1.3%), and DTD (+2.6%) for the ViT-S/16 model. For the largest model (ViT-L/16), HyperVLM shows even more substantial gains, outperforming CLIP by +2.3% on CIFAR-10, +2.4% on CIFAR-100, and +1.5% on Aircraft.
- **Compared to Lorentz MERU:** HyperVLM demonstrates better performance on 15 out of 18 datasets, highlighting the advantages of our Poincaré ball model and entropy-based entailment approach over the Lorentz model used in MERU.
- **Overall Dominance:** HyperVLM achieves the best performance on 13 out of 18 datasets when compared against both CLIP and MERU simultaneously, demonstrating the robust generalization capabilities of our approach across diverse visual domains.

The consistent pattern of improvements across such a wide range of datasets—spanning fine-grained recognition (CUB, Aircraft), texture classification (DTD), scene recognition (SUN397), and general object classification (CIFAR)—demonstrates the robust generalization capabilities of our approach and its ability to capture meaningful hierarchical relationships that benefit diverse visual understanding tasks. Comparing Top N retrieval recall for COCO dataset in Table 2 we see that HyperVLM performs better than CLIP on 4 out of 4 tasks while it performs better than MERU on 3 out of 4 tasks. Overall, HyperVLM performs better than all methods on 3 out of 4 tasks demonstrating the competitiveness of the proposed method. Ablation results for regularization terms λ_{reg} and λ can be referred in the Table 4. In Table 6 we observe that the proposed method outperforms the Euclidean space representation in BLIP architecture for image-text and text-image retrieval task.

5.2. Ablation Study

To isolate the contribution of our novel entropy-based entailment direction determination, we conducted ablation studies comparing HyperVLM with a variant that uses Poincaré embeddings but assumes a fixed text→image entailment direction (as in MERU).

In Table 3 and 5, we observe the difference between the proposed Poincaré embedding with the entropy inferred text-image order entailment loss compared with Poincaré embedding without entropy inferred text-image order entailment where text always entails image in entailment loss. The zero shot classification and retrieval experiments have been conducted for ViT S/16 model for 120000 iterations

using RedCaps dataset, same optimizer and learning rate as the proposed method. As can be observed, the addition of entailment leads to improvement in 14 out of 18 datasets in zero shot classification setting while improving performance in all 4 zero shot retrieval tasks for COCO dataset.

In Table 4 we run the ablation study for λ_{reg} and λ by training ViT-S/16 HyperVLM model for 1 epoch (6k iterations) and compare the average zero shot retrieval accuracy for COCO dataset. We find that $\lambda_{reg} = 0.1$ and $\lambda = 0.1$ provides the best performance and was chosen as the value for our experiments. The entailment loss described in eq. 9 depends on λ_{reg} and λ for calculation of the overall entailment loss.

		<i>text</i> → <i>image</i>		<i>image</i> → <i>text</i>	
		R5	R10	R5	R10
ViT-S/16	Poincaré	30.1	40.2	39.0	50.2
	HyperVLM	30.5	40.2	40.4	50.7

Table 3. Zero Shot Image and Text Retrieval on COCO Dataset. Metric in color represent best performance for the task. Row corresponding to Poincaré represents the case where no entropy derived entailment order is enforced in the entailment loss and we assume that text always entail the image as assumed in MERU. The row corresponding to HyperVLM represent the case where embedding entropy derived image-text entailment order is applied in entailment loss.

		λ				
		0	0.01	0.1	0.5	1
λ_{reg}	0	20.2	17.5	18.6	16.5	18.7
	0.01	20.2	15.4	22.1	16.7	16.0
	0.1	20.2	21.1	22.3	18.9	19.0
	0.5	20.2	19.2	18.6	16.8	15.7
	1	20.2	18.6	18.1	15.4	19.5

Table 4. To select proper values of λ and λ_{reg} we run a grid search for different values and compare the average of average zero shot retrieval accuracy for different retrieval tasks for COCO dataset and zero shot classification accuracy for CIFAR 100 dataset by training ViT-S/16 model for 1 epoch (6K iterations). We find the best performance at $\lambda = 0.1$ and $\lambda_{reg} = 0.1$. The best performance metric in color.

6. Theoretical Insights

6.1. Advantages of Poincaré Ball over Lorentz Hyperboloid

The choice of Poincaré ball model over Lorentz hyperboloid for vision-language representation offers several theoretical and practical advantages [33]. In the Poincaré ball model, the representation capacity scales more effectively with dimension compared to the Lorentz model $\mathbb{L}^n = \{x \in \mathbb{R}^{n+1} : \langle x, x \rangle_{\mathbb{L}} = -1, x_0 > 0\}$ [35]. This superior scaling property emerges from three key aspects:

		Food-101[1]	CIFAR-10[20]	CIFAR-100[20]	CUB[44]	SUN397[46]	Aircraft[25]	DTD[5]	Pets[32]	Caltech-101[11]	Flowers[31]	STL-10[6]	EuroSAT[15]	RESISC45[4]	Country211[34]	MNIST[21]	CLEVR[18]	PCAM[10]	SST2[34]
ViT-S/16	Poincaré	74.9	55.3	27.5	34.1	28.3	1.5	16.4	72.9	60.0	48.4	90.7	28.3	30.6	4.9	8.3	14.4	48.9	50.2
	HyperVLM	75.1	53.6	27.7	35.1	27.6	1.6	17.6	71.9	62.1	47.9	90.9	30.8	32.1	5.1	10.4	14.8	53.8	50.8

Table 5. Comparison of proposed method HyperVLM implementing entropy inferred image-text entailment order, with HyperVLM without entropy inferred image-text entailment order where text always entail image on different datasets. The metrics in color represent best performance metric for particular dataset. We observe that HyperVLM outperforms the Poincaré method where we always assume text to be entailing image, in 14 out of 18 datasets.

- Geometric Properties** The Poincaré ball model provides conformal mapping that preserves angles, leading to more stable optimization. The metric tensor at point x is given by $g_x^{\mathbb{D}} = (\frac{2}{1-\|x\|^2})^2 g^E$ where g_x^E is the metric tensor for euclidean space represented as $g_x^E = \text{diag}([1, 1, \dots, 1])$, which naturally adapts to the hierarchical structure of the data [30]. In contrast, the Lorentz model’s metric tensor $g_x^{\mathbb{L}} = \text{diag}(-1, 1, \dots, 1)$ remains constant, potentially limiting its adaptability to complex hierarchical relationships.
- Numerical Stability** The bounded nature of the Poincaré ball ($\|x\| < 1$) provides inherent numerical stability during optimization [28]. The gradients in Poincaré space are naturally scaled by the conformal factor, preventing exponential explosion or vanishing issues common in Lorentz space where coordinates can grow unboundedly. This leads to more stable training dynamics: $\|\nabla_{\mathbb{D}} f(x)\| \leq \frac{2}{1-\|x\|^2} \|\nabla_E f(x)\|$
- Representation Efficiency** For hierarchical structures of depth d and branching factor b , the Poincaré ball achieves distortion $O(\log(d))$ compared to $O(\sqrt{d})$ in Lorentz space [35]. This leads to more efficient embedding of hierarchical structures, particularly for deep hierarchies: $\text{Distortion}_{\mathbb{D}}(T) < c \log(d) \ll c\sqrt{d} < \text{Distortion}_{\mathbb{L}}(T)$ where T represents a tree structure and c is a constant. This efficiency translates to
 - Better preservation of hierarchical relationships.
 - More accurate representation of fine-grained semantic differences.
 - Improved gradient flow during optimization.

Dataset	Model	Text			Image		
		R@5	R@10	Mean	R@5	R@10	Mean
COCO	BLIP	94.10	97.20	95.65	84.50	90.70	87.6
COCO	HyperVLM(BLIP)	94.52	97.32	95.92	85.12	91.32	88.22
Amazon Clothes	BLIP	2.10	3.30	2.7	6.10	10.50	8.3
Amazon Clothes	HyperVLM(BLIP)	2.74	3.83	3.28	6.41	11.60	9.0

Table 6. Comparison of HyperVLM(BLIP) and BLIP Models for COCO and Amazon Product Recommendation dataset’s Clothes dataset’s Text-Image and Image-text Retrieval

These advantages make the Poincaré ball particularly suitable for vision-language modeling where preserving both hierarchical structure and semantic similarity is crucial.

6.2. Information-Theoretic Hierarchy and Compositional Entailment

Motivation Vision-language representation learning inherently involves hierarchical structures in both modalities. For instance, visual concepts form natural hierarchies (e.g., animal \rightarrow mammal \rightarrow dog \rightarrow breed), and textual descriptions similarly exhibit hierarchical relationships. Traditional Euclidean spaces, with their polynomial volume growth [26], are suboptimal for representing such hierarchical structures. In contrast, hyperbolic geometry, characterized by exponential volume growth [14], naturally accommodates tree-like hierarchical structures.

Information Content and Hierarchical Structure in Shared Space The fundamental connection between embedding complexity and hierarchical relationships can be established through Shannon’s information theory [37]. For embeddings of different modalities projected into a common space through encoders $f_{\theta_{img}}$ and $f_{\theta_{txt}}$, the shared representation ensures that information content comparison is meaningful. This is because:

- The encoders map inputs to a common manifold where geometric and information-theoretic properties are preserved
- The contrastive learning objective ensures semantic alignment in this shared space
- The hyperbolic nature of the space maintains consistent hierarchical relationships across modalities

For an embedding vector x in this shared space, the information entropy:

$$H(x) = - \sum_{i=1}^n p_i \log p_i, \quad p_i = \frac{|x_i|}{\sum_{j=1}^n |x_j|}$$

represents the evolved information content of the con-

cept in the common space [7]. This measure provides theoretical justification for hierarchical relationships because:

1. **The Common Currency Principle** When image and text embeddings are projected into the same space through encoders $f_{\theta_{img}}$ and $f_{\theta_{txt}}$, we can meaningfully compare their information content—like comparing books in the same library using the same cataloging system. This works because:
 - The encoders map inputs to a common manifold where both geometric and information-theoretic properties are preserved
 - The contrastive learning objective ensures semantic alignment in this shared space
 - The hyperbolic nature of the space maintains consistent hierarchical relationships
2. **Information Evolution:** The embedding entropy reflects how information evolves from general to specific concepts in the shared manifold. More specific concepts require additional information beyond their parent concepts. For example, "golden retriever" contains all the information of "dog" plus additional distinguishing features. In our shared embedding space, this principle is captured by:

$$H_{shared}(x) = H_{modal}(x) + I_{alignment}(x)$$

where $I_{alignment}$ represents the additional information gained through cross-modal alignment

3. **Information Content Principle:** More specific concepts require additional information to be fully specified beyond their parent concepts, leading to higher entropy values:

$$\Delta H = H(child) - H(parent) \geq 0$$

7. Discussion

We obtain better performance for HyperVLM over Euclidean space CLIP owing to hyperbolic nature of Poincaré geometry which allows the capture of partial order relation between image and text, in addition to the semantic relation for learning representation. The incremental benefit over MERU can be attributed to 2 reasons: 1. Use of Poincaré space over Lorentz space: As per the work done in [28] Poincaré geometry has a relatively larger capacity than the Lorentz model for correctly representing points 2. Entailment loss based on entropy derived relative hierarchy between image and text at instance level.

7.1. Cross-Modal Generation and Understanding

The improved alignment between visual and textual modalities enables more sophisticated cross-modal generation and understanding:

- **Improved Image Captioning:** By understanding the hierarchical relationships between visual concepts, our approach can generate more accurate and contextually appropriate retail product image captions.
- **Visual Question Answering:** The ability to model entailment relationships helps in answering questions about images that require understanding hierarchical relationships (e.g., "Is there any blue furniture on this e-retail platform?").

8. Conclusion

In this work, we present HyperVLM, a Poincaré geometry-based image-text model tailored to capture the semantic and hierarchical relationships between visual and textual data. Motivated by challenges in retail—such as large-scale product categorization, fine-grained visual distinctions, and multimodal retrieval—our model leverages the properties of hyperbolic space to represent structured product information more effectively than conventional Euclidean methods.

The main contributions of this work are summarized as follows:

1. We introduce HyperVLM, a vision-language model that embeds images and text into a shared hyperbolic space, enabling the representation of both semantic similarity and hierarchical relations prevalent in retail product catalogs.
2. An embedding entropy-based method is proposed to dynamically infer partial order between image-text pairs. This allows the model to flexibly determine modality dominance per instance, rather than relying on fixed assumptions.
3. We show that HyperVLM outperforms Euclidean-based models (e.g., CLIP) and hyperbolic baselines (e.g., MERU) across multiple zero-shot classification and retrieval benchmarks, demonstrating consistent gains in robustness and accuracy.
4. Theoretical justifications are provided for using the Poincaré ball model, showing how its exponential volume growth aligns with the tree-like structure of hierarchical multimodal data.

Our results highlight the value of explicitly modeling hierarchical structure in vision-language learning, especially for real-world retail applications involving vast and evolving product inventories.

In conclusion, HyperVLM offers a scalable and generalizable framework for structure-aware multimodal representation. Its effectiveness in zero-shot settings makes it particularly well-suited for modern retail systems, where robustness, fine-grained understanding, and adaptability are essential.

References

- [1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. pages 446–461, 2014. 8
- [2] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 4868–4879, 2019. 3
- [3] Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. Trees, forests, and imperfect phylogenies: Sublinear-time inference and sample complexity. *arXiv preprint arXiv:2002.00497*, 2020. 2
- [4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 6, 8
- [5] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. pages 3606–3613, 2014. 6, 8
- [6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. pages 215–223, 2011. 6, 8
- [7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 9
- [8] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. *International Conference on Machine Learning*, pages 7694–7731, 2023. 3, 5
- [9] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khruklov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. *arXiv preprint arXiv:2203.10833*, 2022. 3
- [10] Andre Esteva, Brett Kuperl, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017. 6, 8
- [11] 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. *Computer vision and Image understanding*, 106(1):59–70, 2007. 8
- [12] Nora Frankl and János Pach. Embedding graphs in euclidean space. *Journal of the European Mathematical Society*, 22(4): 1139–1148, 2020. 2
- [13] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. *International Conference on Machine Learning*, pages 1646–1655, 2018. 3
- [14] Mikhael Gromov. Hyperbolic groups. *Essays in group theory*, pages 75–263, 1987. 8
- [15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6, 8
- [16] Edwin T Jaynes. Entropy, a measure of uncertainty. *Papers on probability, statistics and statistical physics*, pages 131–148, 1989. 5
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *International conference on machine learning*, pages 4904–4916, 2021. 2, 3
- [18] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. pages 2901–2910, 2017. 8
- [19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2021. 2
- [20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6, 8
- [21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6, 8
- [22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 3, 6
- [23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2017. 6
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019. 6
- [25] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 8
- [26] Jiří Matoušek. On geometric optimization with few violated constraints. *Discrete & Computational Geometry*, 22(4):633–650, 1999. 8
- [27] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 43–52, New York, NY, USA, 2015. Association for Computing Machinery. 6
- [28] Gal Mishne, Ines Chami, and Albert Gu. Numerical stability in hyperbolic neural networks. *arXiv preprint arXiv:2302.10190*, 2023. 8, 9
- [29] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 2
- [30] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 3, 8
- [31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. pages 722–729, 2008. 6, 8

- [32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. The oxford-iiit pet dataset. 2012. 6, 8
- [33] 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*, 2021. 3, 7
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *International conference on machine learning*, pages 8748–8763, 2021. 2, 3, 5, 6, 8
- [35] Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. *International Conference on Machine Learning*, pages 4460–4469, 2018. 2, 3, 7, 8
- [36] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2
- [37] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 8
- [38] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2022. 3
- [39] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 3
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
- [41] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 3
- [42] Tran Dang Vinh, Yi Tay, Shuai Zhang, Gao Cong, and Xiaoli Li. Hyperml: A boosting metric learning approach in hyperbolic space for recommender systems. *arXiv preprint arXiv:2002.06252*, 2020. 3
- [43] Athanasios Vrontzos, Bernhard Kainz, and Ciarán M Gilligan-Lee. Causal representation learning for out-of-distribution generalization in reinforcement learning. *arXiv preprint arXiv:2010.07273*, 2020. 5
- [44] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6, 8
- [45] Chaowei Wang, Yizhou Xu, Feng Ni, Huazhu Yu, Meng Wang, Yuehai Duan, Xiaoqiang Huang, and Mingming Xu. Hierarchical contrastive learning for pattern-generalizable image corruption detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10241–10250, 2021. 3
- [46] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. pages 3485–3492, 2010. 6, 8
- [47] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3
- [48] Xingchen Zhou, Xudong Gu, Ying Ma, Liang Han, Yuchen Guo, Jingjing Liu, and Shuicheng Yan. Learning to generalize across domains on single test samples. In *International Conference on Learning Representations*, 2022. 3

HyperVLM: Hyperbolic Space Guided Vision Language Modeling for Hierarchical Multi-Modal Understanding

Supplementary Material

Hyperbolic Geometry: An Intuitive Overview

In this section, we provide an accessible introduction to hyperbolic geometry and its relevance to our approach. We aim to build intuition for these concepts before diving into the mathematical details.

Hyperbolic Geometry for Hierarchical Data

Imagine trying to draw a large tree with many branches on a flat piece of paper. As you move away from the root, the number of nodes grows exponentially, quickly running out of space. This is precisely the challenge when representing hierarchical data in Euclidean space—the available space grows only polynomially (quadratically in 2D), while hierarchical structures grow exponentially.

Hyperbolic geometry, in contrast, is a non-Euclidean geometry where the space itself expands exponentially as you move away from the origin. This property makes it naturally suited for representing hierarchical structures like trees, taxonomies, and knowledge graphs. In a tree with branching factor b , the number of nodes at level h grows as $((b+1)b^h - 2)/(b-1)$ —an exponential growth that hyperbolic space can accommodate naturally.

Key Concepts in Hyperbolic Geometry

To understand hyperbolic geometry, it helps to compare it with familiar Euclidean geometry:

- **Curvature:** While Euclidean space has zero curvature (flat), hyperbolic space has constant negative curvature. This negative curvature is what creates the exponential expansion of space.
- **Straight Lines:** In hyperbolic space, the shortest path between two points (a geodesic) appears curved when visualized in Euclidean space. This is similar to how flight paths on a globe appear curved on a flat map.
- **Distance:** Distances in hyperbolic space grow exponentially as you move away from the origin, allowing more room to separate points that would be crowded together in Euclidean space.

The Poincaré Disk Model

Among several mathematical models of hyperbolic space, we use the Poincaré disk model due to its favorable properties for deep learning. Think of the Poincaré disk as a unit disk where:

- Points near the center behave similarly to Euclidean space
- As you approach the boundary, distances stretch exponentially

- The boundary represents "infinity" and can never be reached

This model allows us to embed hierarchical structures by placing more general concepts (like "animal") near the center and more specific concepts (like "golden retriever") toward the boundary. The exponential distortion of space naturally preserves the hierarchical relationships between concepts.

8.1. Time Complexity Analysis

While hyperbolic embeddings offer significant advantages for modeling hierarchical relationships, they come with increased computational costs compared to their Euclidean counterparts. In this section, we analyze the worst-case time complexity of HyperVLM and compare it with CLIP and MERU.

8.1.1. Encoder Complexity

The Vision Transformer (ViT) and Text Transformer components have identical time complexity in all three models (CLIP, MERU, and HyperVLM):

- **Vision Transformer:** $\mathcal{O}(n^2 \cdot d)$, where n is the number of image patches and d is the embedding dimension.
- **Text Transformer:** $\mathcal{O}(l^2 \cdot d)$, where l is the sequence length of the text.

8.1.2. Projection and Hyperbolic Mapping

The projection layer in all models has a complexity of $\mathcal{O}(d \cdot n)$. However, HyperVLM and MERU require additional operations to map Euclidean embeddings to hyperbolic space:

- **Exponential Mapping:** $\mathcal{O}(d)$ per embedding for the exponential map operation $\exp_0^{K=-1}(x_i^{Euc})$.
- **Scaling Operation:** $\mathcal{O}(d)$ for the learnable scaling parameters λ_{im} and λ_{txt} .

8.1.3. Distance Calculation

The worst-case time complexity for distance calculations differs significantly:

- **CLIP (Cosine Similarity):** $\mathcal{O}(d)$ per pair of embeddings.
- **MERU (Lorentz Distance):** $\mathcal{O}(d)$ per pair, but with higher constant factors due to hyperbolic operations.
- **HyperVLM (Poincaré Distance):** $\mathcal{O}(d)$ per pair, with operations including \tanh^{-1} and Möbius addition \oplus_K , which have higher computational costs than simple dot products.

8.1.4. Entailment Loss Calculation

The entailment loss calculation in HyperVLM introduces additional complexity:

- **Entropy Calculation:** $\mathcal{O}(d)$ for computing $H(x_{emb}) = -\sum_{i=1}^n x_i \log_2 x_i$.
- **Exterior Angle Calculation:** $\mathcal{O}(d)$ for computing $ext(\angle Oxy)$, involving multiple vector operations.
- **Aperture Calculation:** $\mathcal{O}(1)$ for computing $aper(x)$.

8.1.5. Batch Processing

For a batch of size B , the contrastive loss computation requires:

- **CLIP:** $\mathcal{O}(B^2 \cdot d)$ for computing all pairwise similarities.
- **MERU:** $\mathcal{O}(B^2 \cdot d)$ with higher constant factors.
- **HyperVLM:** $\mathcal{O}(B^2 \cdot d)$ for pairwise distances, plus $\mathcal{O}(B \cdot d)$ for entropy calculations and $\mathcal{O}(B^2 \cdot d)$ for entailment loss.

8.1.6. Overall Worst-Case Time Complexity

The overall worst-case time complexity for a forward pass with batch size B is:

- **CLIP:** $\mathcal{O}(n^2 \cdot d + l^2 \cdot d + B^2 \cdot d)$
- **MERU:** $\mathcal{O}(n^2 \cdot d + l^2 \cdot d + B^2 \cdot d)$ with higher constant factors
- **HyperVLM:** $\mathcal{O}(n^2 \cdot d + l^2 \cdot d + B^2 \cdot d + B \cdot d)$

While the asymptotic complexity remains similar across models, the constant factors in HyperVLM are higher due to the more complex operations in hyperbolic space.

8.1.7. Numerical Stability Considerations

Beyond pure computational complexity, operations in hyperbolic space require careful handling to maintain numerical stability:

- The exponential mapping operation can lead to overflow if not properly scaled, necessitating the learnable scaling parameters λ_{im} and λ_{txt} .
- Operations near the boundary of the Poincaré disk (where $\|x\| \approx 1$) require higher numerical precision, potentially increasing computation time.
- The \tanh^{-1} function in the distance calculation becomes unstable as its input approaches 1, requiring additional checks and safeguards.

Despite these additional computational costs, the improved performance of HyperVLM across multiple benchmarks demonstrates that the benefits of modeling hierarchical relationships in hyperbolic space outweigh the increased computational complexity. For applications where inference speed is critical, techniques such as model distillation or quantization could potentially be applied to reduce the computational overhead while preserving the advantages of hyperbolic embeddings.