GHVL: GEOMETRY-GROUNDED HYPERBOLIC VISION–LANGUAGE MODELS FOR HIERARCHICAL MULTIMODAL REPRESENTATION LEARNING

Sarthak Srivastava*
Amazon
sarthasr@amazon.com

Kathy Wu*
Amazon
rhaow@amazon.com

ABSTRACT

Vision–language models (VLMs) have achieved remarkable performance by aligning visual and textual representations in a shared Euclidean space. However, Euclidean representations inherently fail to capture hierarchical semantic structures present in multimodal data, such as fine-grained categories or conceptual hierarchies. We propose **GHVL**, a *geometry-grounded hyperbolic VLM* that maps images and text into a Poincaré manifold to induce hierarchy-aware representations. By leveraging the exponential capacity of hyperbolic space, GHVL preserves semantic distances across multiple hierarchy levels, enabling faithful modeling of fine-grained concepts. We introduce an adaptive, entropy-driven entailment loss to enforce hierarchical ordering between modalities and integrate it into contrastive objectives for cross-modal alignment. Evaluation on zero-shot classification and image–text retrieval benchmarks demonstrates consistent improvements over Euclidean baselines such as CLIP and Lorentz-based MERU, particularly in hierarchy-sensitive scenarios. These results highlight the importance of respecting geometric structure and demonstrate that hyperbolic representations provide a principled foundation for hierarchical multimodal understanding.

1 INTRODUCTION

Hierarchy in Vision–Language Data. Multimodal data often exhibits natural hierarchical structures, from broad categories to fine-grained concepts. For example, an image of a dog" can correspond to textual descriptions ranging from the general animal" to the specific "puppy playing with a cat". Standard Euclidean embeddings struggle to capture these relationships: distances between general and specific concepts become distorted, and tree-like semantic relationships are not preserved. Hyperbolic geometry, particularly the Poincaré ball model, provides a natural solution: general concepts are placed near the origin, while increasingly specific concepts lie near the boundary, maintaining hierarchical distances. Figure 1 illustrates how GHVL leverages this geometry to jointly embed images and text, capturing both semantic similarity and hierarchical structure in a unified space.

Vision–Language Models and Limitations. Large-scale models such as CLIP (30) and ALIGN (15) learn aligned visual and textual representations via contrastive objectives, demonstrating strong zero-shot generalization across diverse datasets. Despite their success, these models operate in Euclidean space, which limits their ability to represent hierarchical relationships. As a result, distinctions between general and fine-grained concepts may be blurred, which is particularly problematic in applications requiring precise, structured understanding of multimodal data, such as product catalogs or biological image datasets.

Hyperbolic Geometry as an Inductive Bias. Hyperbolic spaces are well-suited to embedding hierarchical structures due to their exponential volume growth (31; 3). In the Poincaré ball, the geometry naturally reflects tree-like relationships, with general concepts near the center and specific concepts near the boundary. This structure allows embeddings to capture both semantic similarity and the

*Equal contribution.

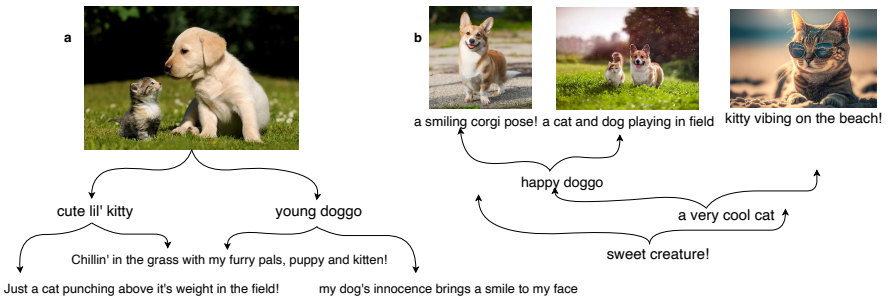

Figure 1: A picture is worth a thousand words. **Left:** Given an informative image it is possible to generate several textual concepts leveraging the visuo-lingual hierarchy. **Right:** Likewise, beginning from a simple text concept, it is possible to come up with complex visuo-lingual concepts by leveraging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

relative specificity of concepts, providing a more faithful representation of multimodal hierarchies than Euclidean spaces.

Geometry-Grounded Hyperbolic Vision–Language Modeling. Motivated by these insights, we introduce **GHVL**, which projects image and text embeddings into the Poincaré ball and enforces hierarchy-aware relationships using an adaptive, entropy-driven entailment loss. GHVL aligns cross-modal representations while maintaining general-to-specific ordering, enabling improved performance on tasks requiring fine-grained hierarchical understanding.

2 Related Work

Vision–Language Models. Large-scale contrastive learning models such as CLIP (30), ALIGN (15), BLIP (19), FLAVA (32), and Florence (39) have demonstrated strong alignment of visual and textual representations and robust zero-shot transfer across downstream tasks. These models typically operate in Euclidean space, which limits their ability to capture hierarchical semantic relationships present in multimodal data.

Hyperbolic Representations. Hyperbolic geometry efficiently embeds hierarchical structures (26; 31), and has been applied to NLP (33), graph representation learning (2), and vision (9; 8). MERU (8) extends hyperbolic embeddings to image-text pairs using the Lorentz model, which has lower representation capacity than the Poincaré ball, motivating our choice of geometry in GHVL.

Hierarchical Multi-Modal Representations. Prior approaches have sought to model hierarchical relationships without explicitly using hyperbolic space, e.g., hierarchical CLIP (40) and HCSC (37). Methods such as Order Embeddings (35) and Hyperbolic Entailment Cones (12) capture partial order relationships in embedding spaces. GHVL extends these ideas by dynamically inferring the entailment direction between images and text via embedding entropy, enabling flexible hierarchy-aware multimodal representations.

3 Hyperbolic Geometry for Vision-Language Models

Many real-world concepts exhibit hierarchical structures: “animal” branches into “dog”, which further branches into “golden retriever”. Capturing such relationships in embeddings is challenging. Standard Euclidean spaces cannot faithfully preserve tree-like hierarchies without large distortions, especially as hierarchy depth increases (26; 3).

Hyperbolic space provides a natural solution. Unlike Euclidean space, hyperbolic space grows exponentially with distance from the origin, allowing more room to represent hierarchical structures compactly. In the *Poincaré ball model*, general concepts lie near the center, while increasingly specific concepts are placed near the boundary. Distances in this space correspond more closely to semantic dissimilarity along hierarchies, which is critical for vision-language tasks involving fine-grained categories.

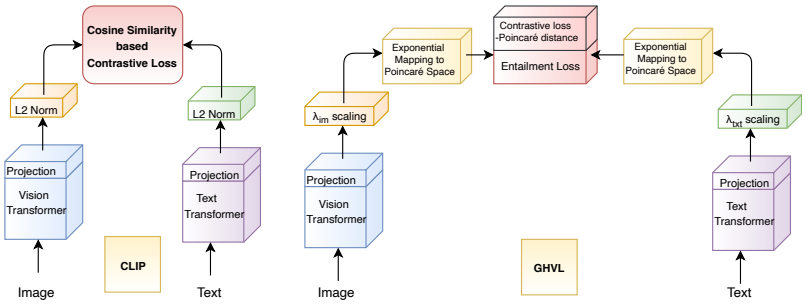

Figure 2: Overall Model Architecture. **Left:** Baseline CLIP. Image and text embeddings are compared via cosine similarity. **Right:** GHVL. Encoder outputs are scaled and projected into hyperbolic Poincaré space, then optimized with contrastive and entailment losses.

Let $\mathbb{P}^n$ denote the n-dimensional Poincaré ball. The hyperbolic distance between points $p_1, p_2 \in \mathbb{P}^n$ is:

$$d_h(p_1, p_2) = 2 \tanh^{-1} \|(-p_1) \oplus p_2\|, \tag{1}$$

where $\oplus$ denotes Möbius addition (34), which generalizes vector addition to the hyperbolic setting:

$$x \oplus y = \frac{(1 + 2\langle x, y\rangle + \|y\|^2)x + (1 - \|x\|^2)y}{1 + 2\langle x, y\rangle + \|x\|^2\|y\|^2}. \tag{2}$$

Hyperbolic embeddings can be computed from Euclidean vectors x via the *exponential map* at the origin:

$$\exp_0(x) = 0 \oplus \left(\tanh(\|x\|)\frac{x}{\|x\|}\right), \tag{3}$$

with the inverse logarithmic map allowing retrieval of Euclidean representations for downstream tasks:

$$\log_0(h) = \tanh^{-1}(\|h\|)\frac{h}{\|h\|}. \tag{4}$$

In GHVL, both image and text embeddings are projected into this hyperbolic space. This ensures that:

1. **Hierarchical consistency:** embeddings of general concepts remain near the origin, and specific concepts expand toward the boundary.
2. **Cross-modal alignment:** images and text describing the same concept are embedded close to each other in hyperbolic distance.
3. **Semantic discrimination:** subtle distinctions between fine-grained categories are preserved in the curvature of the space.

Figure 4 visualizes this arrangement: high-level categories cluster near the center, while low-level, specific items fan out near the boundary. This structured representation enables GHVL to achieve superior performance on zero-shot classification and image-text retrieval, particularly for hierarchical concepts.

For full derivations, curvature properties, and Riemannian metric details, see Appendix Sec. A2.

4 METHODOLOGY

In this section, we introduce **GHVL**, a novel vision-language model designed to learn hierarchy-aware, cross-modal representations. Unlike standard CLIP (30) or MERU (8), which capture semantic similarity in Euclidean space, GHVL embeds both images and text into a hyperbolic Poincaré space and enforces a dynamic, entropy-driven partial order between modalities. This combination

allows GHVL to simultaneously preserve semantic alignment and hierarchical structure, effectively modeling general-to-specific relationships across modalities.

GHVL builds upon the CLIP architecture, consisting of a vision transformer (ViT) image encoder and a text transformer encoder using byte pair encoding. Both encoders produce fixed-size embeddings of dimension n, which are projected into a latent space. GHVL extends this pipeline in two key ways: (i) embedding projection into hyperbolic Poincaré space to encode hierarchical relationships, and (ii) a novel entropy-driven entailment loss to enforce partial-order structure between modalities. As shown in Figure 4, GHVL embeddings follow a hierarchical structure with low-entropy concepts near the center.

Transfer of Embeddings onto the Poincaré Space. During training, image and text samples are passed through the respective ViT and text transformer encoders, followed by a projection layer, as shown in Figure 2. The resulting Euclidean embeddings (ν_{im}, ν_{txt}) are mapped into hyperbolic Poincaré space as (h_{im}, h_{txt}) using the exponential map (Eq. 3) with respect to the origin. This ensures that the hierarchical relationships between general and specific concepts are explicitly captured in the embedding space.

Numerical Overflow Prevention. Mapping embeddings from Euclidean space to hyperbolic space involves exponential operations, which can inflate the norm of embeddings from $\mathcal{O}(\sqrt{n})$ to $\mathcal{O}(e^{\sqrt{n}})$, causing potential numerical instability. To address this, we apply scaling via two learnable parameters λ_{im} and λ_{txt}, initialized to $1/\sqrt{n}$, before exponential mapping. This stabilizes the embeddings while preserving their geometric properties.

Training Objectives. GHVL jointly optimizes for semantic similarity and hierarchical structure between image-text pairs. To achieve this, we combine a hyperbolic contrastive loss with an adaptive, entropy-driven entailment loss.

4.1 Contrastive Loss

We adopt the multi-class N-pair contrastive loss from CLIP (30), with a key modification: similarity is computed via negative Poincaré distance(Eq. 3) instead of cosine similarity. For a batch of size N, each image-text pair has one positive and $N-1$ negative samples, and the overall contrastive loss $\mathcal{L}_{cont}$ is computed as the average of image-wise and text-wise losses. This enforces semantic alignment while respecting the hyperbolic geometry of the embedding space.

4.2 Entailment Loss

To capture hierarchical structure, we introduce an adaptive, entropy-driven entailment loss, extending prior work (8). Unlike MERU, which assumes a fixed direction of entailment (text always entails image), GHVL dynamically determines which modality is more general or specific based on the embedding entropy.

For an image-text pair, the embedding with lower entropy is considered the entailer (more general concept), while the higher-entropy embedding is the entailed (more specific). Formally, the entropy of an embedding x_{emb} is computed as:

$$H(x_{emb}) = -\sum_{i=1}^{n} x_i \log_2 x_i \tag{5}$$

The adaptive assignment is:

$$x = \begin{cases} x_{img}, & \text{if } H(x_{img}) < H(x_{txt}) \\ x_{txt}, & \text{otherwise} \end{cases} \qquad y = \begin{cases} x_{txt}, & \text{if } H(x_{img}) < H(x_{txt}) \\ x_{img}, & \text{otherwise} \end{cases} \tag{6}$$

The entailment loss is defined using the projected Euclidean cone:

$$\mathcal{L}_{entail}(x, y) = \max\left(0, ext(\angle Oxy) - aper(x)\right) - \lambda_{reg} ext(\angle Oxy) \tag{7}$$

where $ext(\angle Oxy)$ is the exterior angle between embeddings (Eq. 14) and $aper(x)$ is the half-aperture of the entailment cone (Eq. 15). The hyperparameter λ_{reg} controls regularization strength.

4.3 Overall Loss

The final optimization objective combines both contrastive and entailment losses:

$$\mathcal{L} = \mathcal{L}_{cont} + \lambda \mathcal{L}_{entail} \tag{8}$$

where λ is the entailment loss weight. This dual-objective framework ensures GHVL simultaneously preserves semantic similarity and hierarchical ordering across modalities.

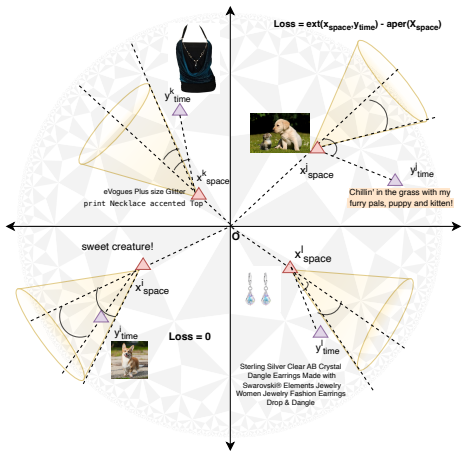

Figure 3: Entailment Cone (projection from Poincaré space onto Euclidean space). The loss pushes the higher-entropy embedding inside the cone projected by the lower-entropy embedding. Indices i and j in superscripts denote two different instances of image-text pairs.

5 Experiments

Model Architecture We implement GHVL using different size versions of Vision Transformers (S/B/L) as vision encoders with a patch size of 16, freezing the positional encoding layer. The text encoder follows the CLIP architecture with a 12-layer, 512-dimensional Transformer with 77 maximum length byte pair encoding. For hyperbolic 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 (21) with weight decay of 0.2 and (β_1, β_2) = (0.9, 0.98). Weight decay is disabled for all gains, biases, and learnable scalars. The model is trained

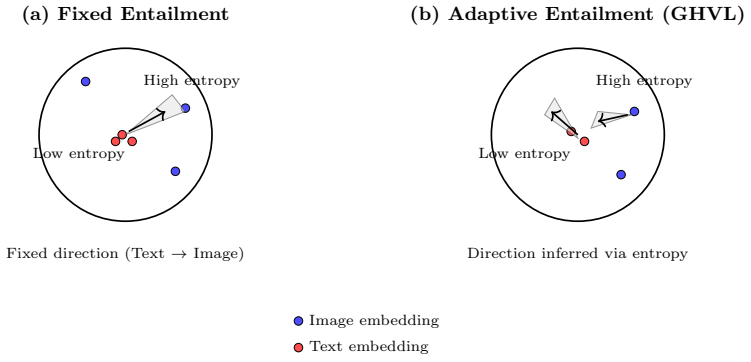

Figure 4: Illustration of image-text embeddings projected into hyperbolic Poincaré space. Lower-entropy embeddings (more general concepts) are closer to the center, while higher-entropy embeddings (more specific) are near the boundary.

for 120K iterations with batch size 1024 ($\approx$ 10 epochs). The maximum learning rate is 5×10^{-4}, increasing linearly for the first 4K iterations, followed by cosine decay to 0 (20).

Evaluation Protocol We evaluate GHVL against CLIP and MERU on 18 diverse datasets for zero-shot classification and on COCO for retrieval tasks. Additionally, we evaluate image-text and text-image retrieval using BLIP (19), integrating our entropy-based ITC loss in Poincaré space. We test BLIP on COCO and Amazon Product Recommendation – Clothes datasets (24) (Table 3).

Additional experiment results and ablation study can be found in Appendix Sec. A4.

		CIFAR-10(17)	CIFAR-100(17)	CUB(36)	SUN397(38)	Aircraft(22)	DTD(5)	Pets(28)	Flowers(27)	STL-10(6)	EuroSAT(14)	RESISC45(4)	Country211(30)	MNIST(18)	PCAM(10)	SST2(30)
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
	GHVL	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
	GHVL	**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
	GHVL	**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 GHVL with baseline methods (CLIP and MERU) across multiple image classification datasets using ViT-S/16, ViT-B/16, and ViT-L/16 backbones. GHVL leverages geometry-grounded hyperbolic embeddings to improve hierarchical multimodal representation. Best metrics for each dataset are highlighted in **color**.

6 Theoretical Insights

6.1 Advantages of Poincaré Ball over Lorentz Hyperboloid

Choosing the Poincaré ball model over the Lorentz hyperboloid for vision-language representations provides both theoretical and practical benefits (29). In the Poincaré ball, representation capacity scales more effectively with dimension than in the Lorentz model $\mathbb{L}^n = \{x \in \mathbb{R}^{n+1} : \langle x, x \rangle_{\mathbb{L}} = -1, x_0 > 0\}$ (31). This advantage arises from three key aspects:

		$text \rightarrow image$		$image \rightarrow text$	
		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
	GHVL	**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
	GHVL	**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
	GHVL	**32.6**	42.7	**43.2**	**53.8**

Table 2: Zero-shot image-text retrieval results on the COCO dataset for ViT backbones. GHVL demonstrates improved cross-modal alignment over CLIP and MERU. Best metrics in each column are highlighted in **color**.

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	GHVL	**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	GHVL	**2.74**	**3.83**	**3.28**	**6.41**	**11.60**	**9.0**

Table 3: Text-image and image-text retrieval performance of GHVL versus BLIP on COCO and Amazon Clothes datasets. GHVL leverages geometry-grounded embeddings for improved cross-modal alignment. Best metrics in each column are highlighted in **color**.

- **Geometric Properties:** The Poincaré ball provides a conformal mapping that preserves angles, yielding more stable optimization. Its metric tensor at point x is

$$g_x^{\mathbb{D}} = \left(\frac{2}{1 - \|x\|^2}\right)^2 g^E,$$

  where $g^E = \text{diag}([1, 1, \ldots, 1])$ is the Euclidean metric tensor. This scaling adapts naturally to hierarchical structures (26). In contrast, the Lorentz metric $g_x^{\mathbb{L}} = \text{diag}(-1, 1, \ldots, 1)$ is constant, limiting its flexibility in representing complex hierarchies.
- **Numerical Stability:** The bounded Poincaré ball ($\|x\| < 1$) ensures numerical stability during training (25). Gradients are scaled by the conformal factor:

$$\|\nabla_{\mathbb{D}} f(x)\| \leq \frac{2}{1 - \|x\|^2} \|\nabla_E f(x)\|,$$

  preventing exponential explosion or vanishing that can occur in Lorentz space with unbounded coordinates. This results in smoother and more reliable optimization.
- **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 (31):

$$\text{Distortion}_{\mathbb{D}}(T) < c \log(d) \ll c\sqrt{d} < \text{Distortion}_{\mathbb{L}}(T),$$

  where T is a tree and c is a constant. This efficiency translates to:
  1. Better preservation of hierarchical relationships,
  2. More accurate representation of fine-grained semantic differences,
  3. Improved gradient flow during optimization.

These properties make the Poincaré ball especially suitable for vision-language modeling, where preserving hierarchy and semantic similarity is critical.

6.2 Information-Theoretic Hierarchy and Compositional Entailment

Motivation: Vision-language representations inherently involve hierarchical structures. Visual concepts often form natural hierarchies (e.g., animal → mammal → dog → breed), and textual descriptions mirror these structures. Euclidean spaces, with polynomial volume growth (23), are suboptimal for representing such hierarchies. Hyperbolic geometry, with exponential volume growth (13), naturally accommodates tree-like structures.

Information Content in a Shared Space: Embedding vectors from different modalities, projected into a common space through encoders $f_{\theta_{img}}$ and $f_{\theta_{txt}}$, allow meaningful comparisons of information content. For an embedding x in this shared space, the entropy:

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

quantifies the information content of a concept (7). This provides a theoretical basis for capturing hierarchical relationships:

1. **Common Currency Principle:** By projecting image and text embeddings into the same space, we can compare their information content directly. This is analogous to using a shared cataloging system for books in a library. The alignment works because:

- Encoders map inputs to a shared manifold preserving geometric and information-theoretic properties,
- Contrastive learning ensures semantic alignment,
- The hyperbolic geometry maintains consistent hierarchical relationships.

2. **Information Evolution:** Embedding entropy reflects how information evolves from general to specific concepts. For instance, "golden retriever" contains all information of "dog" plus additional details. In the shared embedding space:

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

where $I_{\text{alignment}}$ represents additional information gained through cross-modal alignment.

3. **Information Content Principle:** More specific concepts require extra information beyond their parent concepts:

$$\Delta H = H(\text{child}) - H(\text{parent}) \geq 0.$$

This principle guides the entropy-based entailment mechanism used in GHVL to enforce hierarchical consistency across modalities.

7 Discussion

GHVL achieves superior performance over Euclidean CLIP models by leveraging the hyperbolic nature of the Poincaré geometry, which effectively encodes both semantic similarity and hierarchical relationships between images and text. Its improvements over Lorentz-based MERU stem from two complementary design choices. First, representing embeddings in Poincaré space provides greater capacity for modeling hierarchical structures, naturally accommodating the tree-like relationships inherent in multimodal data. Second, the entropy-based entailment loss dynamically infers instance-specific partial orders, enforcing more precise cross-modal alignment than fixed-order assumptions. Together, these components allow GHVL to preserve hierarchical consistency while improving semantic alignment, which is especially beneficial in domains with complex, structured data such as retail catalogs, scientific imagery, or knowledge graphs.

The enhanced multimodal alignment also translates into more sophisticated reasoning capabilities. By explicitly capturing hierarchical relationships, GHVL supports more accurate, context-aware image captioning that reflects both visual content and conceptual structure. Hierarchical visual question answering similarly benefits, enabling the model to handle queries that require layered or conditional reasoning, e.g., "Are there any blue leather chairs available in this catalog?" Beyond practical applications, GHVL illustrates the value of geometry-grounded representation learning in multimodal AI, showing how hyperbolic embeddings can reconcile semantic similarity with structural constraints. Overall, the combination of Poincaré embeddings and entropy-driven entailment demonstrates that structure-aware reasoning across vision and language modalities is both feasible and effective, providing a foundation for future research in geometry-aware multimodal systems.

8 Conclusion

We introduced **GHVL (Geometry-Grounded Hyperbolic Vision–Language models)**, a Poincaré geometry-based framework for hierarchical multimodal representation learning. Our approach embeds images and text into a shared hyperbolic space, enabling simultaneous modeling of semantic similarity and hierarchical relationships. Motivated by challenges in retail—large-scale product catalogs, fine-grained visual distinctions, and multimodal retrieval—GHVL provides a scalable and generalizable solution.

Key contributions include:

1. **Hyperbolic multimodal embeddings**: GHVL represents images and text in a shared Poincaré ball, preserving tree-like hierarchical relationships inherent in multimodal data.
2. **Entropy-based partial order inference**: We propose a method to dynamically determine image-text entailment order at the instance level, providing flexible and adaptive cross-modal alignment.

3. **Empirical validation**: GHVL consistently outperforms Euclidean-based models (CLIP) and Lorentz-based hyperbolic models (MERU) across multiple zero-shot classification and retrieval benchmarks, demonstrating robustness and practical applicability.
4. **Theoretical grounding**: We provide justification for Poincaré embeddings, showing how their exponential volume growth naturally aligns with hierarchical multimodal structures.

By explicitly modeling hierarchical structure in vision-language learning, GHVL enhances cross-modal understanding and reasoning, making it particularly well-suited for large-scale, structured multimodal systems. This work highlights the advantages of geometry-grounded representations in capturing hierarchical structure for structured multimodal AI.

REFERENCES

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. pp. 446–461, 2014.

[2] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 4868–4879, 2019.

[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.

[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.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. pp. 3606–3613, 2014.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. pp. 215–223, 2011.

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[8] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. *International Conference on Machine Learning*, pp. 7694–7731, 2023.

[9] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. *arXiv preprint arXiv:2203.10833*, 2022.

[10] Andre Esteva, Brett Kuprel, 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.

[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.

[12] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. *International Conference on Machine Learning*, pp. 1646–1655, 2018.

[13] Mikhael Gromov. Hyperbolic groups. *Essays in group theory*, pp. 75–263, 1987.

[14] 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.

[15] 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*, pp. 4904–4916, 2021.

[16] 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. pp. 2901–2910, 2017.

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[18] 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.

[19] 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.

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2017.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2019.

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

[23] Jiří Matoušek. On geometric optimization with few violated constraints. *Discrete & Computational Geometry*, 22(4):633–650, 1999.

[24] 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*, SIGIR '15, pp. 43–52, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336215. doi: 10.1145/2766462.2767755. URL https://doi.org/10.1145/2766462.2767755.

[25] Gal Mishne, Ines Chami, and Albert Gu. Numerical stability in hyperbolic neural networks. *arXiv preprint arXiv:2302.10190*, 2023.

[26] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[27] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. pp. 722–729, 2008.

[28] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. The oxford-iiit pet dataset. 2012.

[29] 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.

[30] 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*, pp. 8748–8763, 2021.

[31] Frederic Sala, Christopher De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. *International Conference on Machine Learning*, pp. 4460–4469, 2018.

[32] 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.

[33] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

[34] Abraham A. Ungar. Möbius gyrovector spaces in quantum information and computation. *Commentationes Mathematicae Universitatis Carolinae*, 49(2):341–356, 2008. URL `http://eudml.org/doc/250484`. Discusses Möbius addition and its generalization in gyrovector spaces.

[35] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[37] 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*, pp. 10241–10250, 2021.

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. pp. 3485–3492, 2010.

[39] 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.

[40] 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.

A Appendix

A1. Hyperbolic Geometry: An Intuitive Overview

This section provides an accessible introduction to hyperbolic geometry and its relevance to GHVL, aiming to build intuition before presenting mathematical details.

A1.1 Hyperbolic Geometry for Hierarchical Data

Representing hierarchical data in Euclidean space is challenging: as tree depth grows, the number of nodes expands exponentially, quickly exceeding available space. Hyperbolic geometry, in contrast, expands exponentially with distance from the origin, naturally accommodating tree-like structures, taxonomies, and knowledge graphs. For a tree with branching factor b, the number of nodes at level h grows as $((b+1)b^h - 2)/(b-1)$—a growth pattern that hyperbolic space handles efficiently.

A1.2 Key Concepts in Hyperbolic Geometry

- **Curvature**: Euclidean space is flat (zero curvature), whereas hyperbolic space has constant negative curvature, producing exponential expansion.
- **Geodesics**: Shortest paths in hyperbolic space appear curved in Euclidean representations, similar to great-circle flight paths on a globe.
- **Distance**: Points diverge exponentially from the origin, providing ample space for hierarchical separation.

A1.3 The Poincaré Disk Model

GHVL uses the Poincaré disk model for its favorable properties in deep learning. The model is a unit disk where:

- Central points behave nearly Euclidean
- Distances stretch exponentially toward the boundary
- The boundary represents "infinity" and cannot be reached

This allows general concepts (e.g., "animal") to be placed near the center and specific concepts (e.g., "golden retriever") near the boundary, preserving hierarchical relationships naturally.

A2. Hyperbolic Geometry: Detailed Formulation

In this section, we provide a complete mathematical description of the hyperbolic space used in GHVL. We focus on the n-dimensional Poincaré ball model $\mathbb{P}^n$ and the operations needed for embedding and optimization.

A2.1 Poincaré Ball Model

The n-dimensional Poincaré ball of radius $1/\sqrt{c}$ (with curvature $-c$, $c > 0$) is defined as

$$\mathbb{P}_c^n = \{x \in \mathbb{R}^n : \|x\| < 1/\sqrt{c}\}. \tag{9}$$

The hyperbolic distance between two points $p, q \in \mathbb{P}_c^n$ is:

$$d_{\mathbb{P}}(p,q) = \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c}\|(-p) \oplus_c q\|\right), \tag{10}$$

where $\oplus_c$ is the Möbius addition defined below.

A2.2 MÖBIUS ADDITION AND SCALAR MULTIPLICATION

Möbius addition generalizes Euclidean vector addition to hyperbolic space:

$$x \oplus_c y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2}. \tag{11}$$

Möbius scalar multiplication (stretching along geodesics) is defined for $r \in \mathbb{R}$ and $x \in \mathbb{P}_c^n$ as:

$$r \otimes_c x = \tanh\left(r \tanh^{-1}(\sqrt{c}\|x\|)\right)\frac{x}{\|x\|\sqrt{c}}. \tag{12}$$

A2.3 RIEMANNIAN METRIC AND CURVATURE

The Poincaré ball is a Riemannian manifold with metric:

$$g_x = \left(\frac{2}{1 - c\|x\|^2}\right)^2 g_{\text{Euclidean}}, \tag{13}$$

where $g_{\text{Euclidean}}$ is the standard Euclidean metric. This metric induces constant negative curvature $-c$, allowing the embedding space to expand exponentially and represent hierarchies efficiently.

A2.4 EXPONENTIAL AND LOGARITHMIC MAPS

To move between Euclidean and hyperbolic representations:

Exponential map at the origin:

$$\exp_0^c(x) = \tanh(\sqrt{c}\|x\|)\frac{x}{\|x\|\sqrt{c}} \in \mathbb{P}_c^n. \tag{14}$$

Logarithmic map at the origin:

$$\log_0^c(y) = \frac{1}{\sqrt{c}} \tanh^{-1}(\sqrt{c}\|y\|)\frac{y}{\|y\|} \in \mathbb{R}^n. \tag{15}$$

These maps allow standard Euclidean optimization techniques to be applied in the hyperbolic space by pulling vectors to the tangent space at the origin and pushing back after updates.

A2.5 HYPERBOLIC EMBEDDING IN GHVL

In GHVL, image embeddings v and text embeddings t are projected via a Euclidean-to-hyperbolic map:

$$h_v = \exp_0^c(W_v v), \quad h_t = \exp_0^c(W_t t), \tag{16}$$

where W_v, W_t are learnable projection matrices. The hyperbolic distance $d_{\mathbb{P}}(h_v, h_t)$ is then used in the alignment loss, ensuring:

1. Hierarchical consistency: higher-level concepts near the origin.
2. Fine-grained discrimination: subtle semantic differences captured via distance and curvature.
3. Cross-modal alignment: images and text describing the same concept are embedded close together.

A3. TIME COMPLEXITY ANALYSIS

Hyperbolic embeddings provide hierarchical modeling advantages but introduce higher computational costs compared to Euclidean embeddings. We analyze the worst-case complexity of GHVL relative to CLIP and MERU.

A3.1 Encoder Complexity

Vision and text encoders have the same complexity across models:

- **Vision Transformer (ViT)**: $\mathcal{O}(n^2 \cdot d)$, with n patches and d embedding dimension
- **Text Transformer**: $\mathcal{O}(l^2 \cdot d)$, with l tokens

A3.2 Projection and Hyperbolic Mapping

Projection layer: $\mathcal{O}(d \cdot n)$

Additional hyperbolic operations for GHVL:

- **Exponential Mapping**: $\mathcal{O}(d)$ per embedding
- **Scaling Parameters**: $\mathcal{O}(d)$ for λ_{im} and λ_{txt}

A3.3 Distance Calculation

- **CLIP (Cosine)**: $\mathcal{O}(d)$ per pair
- **MERU (Lorentz)**: $\mathcal{O}(d)$, with higher constants
- **GHVL (Poincaré)**: $\mathcal{O}(d)$, including $tanh^{-1}$ and Möbius addition $\oplus_K$

A3.4 Entailment Loss Computation

- **Entropy**: $\mathcal{O}(d)$
- **Exterior Angle**: $\mathcal{O}(d)$
- **Aperture**: $\mathcal{O}(1)$

A3.5 Batch Processing

- **CLIP**: $\mathcal{O}(B^2 \cdot d)$
- **MERU**: $\mathcal{O}(B^2 \cdot d)$, higher constants
- **GHVL**: $\mathcal{O}(B^2 \cdot d + B \cdot d)$

A3.6 Overall Complexity

- **CLIP**: $\mathcal{O}(n^2 \cdot d + l^2 \cdot d + B^2 \cdot d)$
- **MERU**: same asymptotic order, higher constants
- **GHVL**: $\mathcal{O}(n^2 \cdot d + l^2 \cdot d + B^2 \cdot d + B \cdot d)$

A3.7 Numerical Stability Considerations

- Exponential mapping may overflow without proper scaling; λ_{im} and λ_{txt} mitigate this.
- Operations near the Poincaré boundary ($\|x\| \approx 1$) require high precision.
- $tanh^{-1}$ can become unstable near 1, requiring safeguards.

Despite the added costs, GHVL's superior hierarchical modeling justifies the complexity. For latency-sensitive scenarios, distillation or quantization can reduce overhead while preserving benefits.

A4. Additional Experiment Results and Ablation Study

		$text \rightarrow image$		$image \rightarrow text$	
		R5	R10	R5	R10
ViT-S/16	**Poincaré**	30.1	**40.2**	39.0	50.2
	GHVL	**30.5**	**40.2**	**40.4**	**50.7**

Table 4: Zero-shot image-text retrieval on COCO comparing GHVL with the Poincaré baseline, which assumes text always entails image. GHVL uses entropy-derived image-text entailment ordering, leading to superior retrieval performance. Best metrics in each column are highlighted in **color**.

		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 5: Grid search for hyperparameters λ and λ_{reg} using average zero-shot retrieval accuracy on COCO and CIFAR-100 with ViT-S/16. Best performance is achieved at $\lambda = 0.1$ and $\lambda_{reg} = 0.1$. Metrics in each column are highlighted in **color**.

		Food-101(1)	CIFAR-10(17)	CIFAR-100(17)	CUB(36)	SUN397(38)	Aircraft(22)	DTD(5)	Pets(28)	Caltech-101(11)	Flowers(27)	STL-10(6)	EuroSAT(14)	RESISC45(4)	Country211(30)	MNIST(18)	CLEVR(16)	PCAM(10)	SST2(30)
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
	GHVL	**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 6: Comparison of GHVL versus Poincaré embeddings across 18 datasets, showing the effect of entropy-inferred image-text entailment order on hierarchical multimodal representation. GHVL outperforms Poincaré in 14 out of 18 datasets. Best metrics in each column are highlighted in **color**.