ARGENT: Adaptive Hierarchical Image-Text Representations

Chuong Huynh[1†], Hossein Souri[2], Abhinav Kumar[2†], Vitali Petsiuk[2], Deen Dayal Mohan[2], and Suren Kumar[2†]

[1] University of Maryland, College Park
[2] Samsung Research America, AI Center – Mountain View
Project Page: `https://hmchuong.github.io/argent`

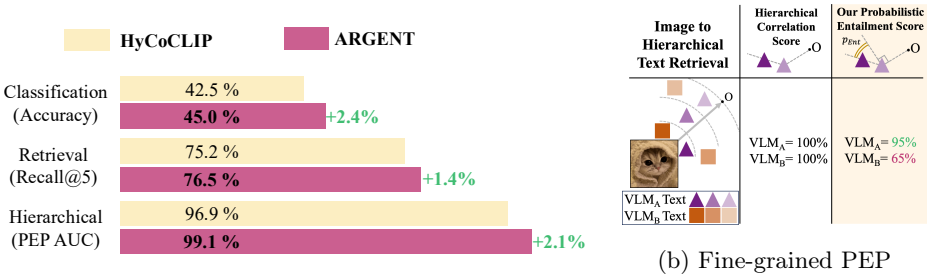
Abstract. Large-scale Vision–Language Models (VLMs) such as CLIP learn powerful semantic representations but operate in Euclidean space, which fails to capture the inherent hierarchical structure of visual and linguistic concepts. Hyperbolic geometry, with its exponential volume growth, offers a principled alternative for embedding such hierarchies with low distortion. However, existing hyperbolic VLMs use entailment losses that are unstable: as parent embeddings contract toward the origin, their entailment cones widen toward a half-space, causing catastrophic cone collapse that destroys the intended hierarchy. Additionally, hierarchical evaluation of these models remains unreliable, being largely retrieval-based and correlation-based metrics and prone to taxonomy dependence and ambiguous negatives. To address these limitations, we propose an adaptive entailment loss paired with a norm regularizer that prevents cone collapse without heuristic aperture clipping. We further introduce an angle-based probabilistic entailment protocol (PEP) for evaluating hierarchical understanding, scored with AUC-ROC and Average Precision. This paper introduces a stronger hyperbolic VLM baseline **ARGENT**, **A**daptive hie**R**archical ima**G**e-t**E**xt represe**NT**ation. ARGENT improves the SOTA hyperbolic VLM by 0.7, 1.1, and 0.8 absolute points on image classification, text-to-image retrieval, and proposed hierarchical metrics, respectively.

Keywords: Hierarchical Learning · Image-Text Representation · Vision Language Model

1 Introduction

Large-scale Vision-Language Models (VLMs) like CLIP [14, 29] have revolutionized the field by learning semantically aligned representations from web-scale image-text pairs. Their success, however, rests on a "flat-world" assumption: concepts are represented as points in Euclidean space, implicitly treating all

[†] Work done while at Samsung Research America.



(a) ARGENT outperforms SOTA HyCoCLIP.

(b) Fine-grained PEP evaluation.

Fig. 1: ARGENT improves both hierarchical training and evaluation. (a) ARGENT outperforms the HyCoCLIP baseline on downstream tasks and our hierarchical benchmark (PEP AUC). **(b)** Our new **Probabilistic Entailment Score** (p_{Ent}) offers a more discriminative evaluation; while both models may achieve 100% correlation, our metric correctly identifies VLM_A as superior.

semantic relationships as uniform distances. This clashes with the deeply hierarchical structure of both the visual world and human language [9], where a *"poodle"* is a specific type of *"dog"*, which is a class of *"animal"*. While recent work reports traces of emergent hierarchical structure in CLIP-like models [1, 6, 7], these effects are inconsistent, dataset-dependent, and difficult to control. The primary bottleneck often appears to be the text encoder, which flattens structured language into a single vector. This approach struggles to explicitly model linguistic hierarchy, hindering its alignment with structured visual information. For example, the text encoder may represent *"a poodle"* and *"a dog"* as closely related, yet fail to encode the explicit 'is-a' relationship, treating it similarly to a purely associative link (like *"dog"* and *"leash"*).

The core challenge in visual-language modeling (VLM) is geometrically accommodating the inherent exponential complexity of the conceptual and compositional hierarchies (*e.g.*, *"dog"* $\rightarrow$ *"mammal"* $\rightarrow$ *"animal"*). Embedding this structure into a traditional Euclidean space, which exhibits only polynomial volume growth, is not only mathematically inefficient but fundamentally induces high distortion. To resolve this, recent work [9, 25, 30] has leveraged hyperbolic geometry. This non-Euclidean space is defined by its constant negative curvature and, crucially, its exponential volume growth with respect to the radius. This property makes it the continuous analog of a discrete tree and a natural, low-distortion substrate for hierarchical data.

The application of hyperbolic geometry to VLMs has progressed rapidly. Initial explorations, such as MERU [9], pioneered contrastive learning in hyperbolic space. Subsequent efforts, like HyCoCLIP [25], advanced this by modeling compositional relationships using hyperbolic entailment cones. While foundational, existing hyperbolic VLMs face two significant challenges that impede further progress. On the methodological front, their core loss function suffers from **Entailment Loss Instability**, making training brittle and prone to collapse. Sepa-

rately, the evaluation framework for these models exhibits a critical shortcoming: current hierarchical reasoning benchmarks are often unreliable and sensitive to false negatives, failing to accurately measure true hierarchical understanding.

We first address the critical challenge of **Entailment Loss Instability**. The entailment loss in prior methods is inherently unstable due to a geometric vulnerability. The hierarchical relationship is defined by a cone whose aperture is inversely coupled to the norm of the parent embedding. The model can exploit this coupling to find a pathological solution: by minimizing the parent norms, the cone's aperture widens until it degenerates into a flat half-space. This collapse destroys the geometric notion of entailment, leading to a collapsed representation space and catastrophic training instability. HyCoCLIP [25] attempted to mitigate this with a heuristic regularizer, η, to manually restrict the cone's aperture. However, this is an ad-hoc fix that fails to address the fundamental geometric cause.

We address this by proposing the Adaptive Entailment (AdaEnt) Loss that dynamically scales the entailment angle based on embedding similarity, removing the detrimental norm-to-cone relationship without relying on unstable heuristics. We stabilize it further by integrating AdaEnt with a norm regularizer. Fig. 1a shows that our resulting model, ARGENT, trained with the adaptive loss, outperforms the HyCoCLIP baseline by a large margin.

Second, we tackle the pervasive problem of **Inadequate Hierarchical Evaluation**. Prior benchmarks, often relying on WordNet, are limited to short, single-level phrases. While more recent datasets like HierarCaps [1] provide valuable multi-level captions, their established evaluation protocol suffers from two issues: (1) The inherent hierarchical signal within these datasets is weak due to ambiguous ordering. (2) More critically, the evaluation procedure introduces a systemic bias by incorrectly defining the ground truth. During standard image-to-text retrieval, only the single caption paired with the query image is considered a positive match. This ignores that many other captions in the candidate pool could be equally valid hierarchical parents or children, preventing the evaluation from measuring true hierarchical generalization.

To address this, we propose a new, direct, and robust Probabilistic Entailment Protocol (PEP). Leveraging HierarCaps, we redefine the metric by treating the angle between embeddings as a direct proxy for entailment probability. We then compute AUC-ROC and Average Precision (AP) across a carefully selected set of intra- and inter-sample pairs. This protocol moves beyond brittle retrieval accuracy to provide a significantly more reliable and fine-grained measure of a model's hierarchical understanding. Fig. 1b shows that our metric provides a fine-grained score that correctly distinguishes model performance where the original protocol fails.

In summary, this paper makes three key contributions to advance how VLMs represent conceptual hierarchies:

- We introduce the novel Adaptive Entailment (AdaEnt) Loss, which resolves the critical Entailment Loss Instability in hyperbolic VLMs, ensuring stable training and robust representation (Sec. 4).

- We propose the Probabilistic Entailment Protocol (PEP), a new, objective evaluation that overcomes the systematic biases of prior benchmarks to reliably measure hierarchical generalization (Sec. 5).
- Through comprehensive analysis, we validate our proposed loss and protocol, demonstrating that ARGENT sets a new state-of-the-art on hierarchical representation learning benchmarks (Sec. 6).

Together, these contributions provide a stronger foundation for building and validating the next generation of hyperbolic VLMs.

2 Related Work

This work builds upon advancements in three primary areas: large-scale vision-language models, the application of hyperbolic geometry to deep learning, and the rapidly emerging field of hyperbolic vision-language models.

Vision-Language Models. Large-scale Vision-Language Models (VLMs), pre-trained on extensive web data, have emerged as the leading approach in multi-modal research. Models such as CLIP [29] and ALIGN [14] utilize a contrastive loss on massive image-text datasets to learn a joint, semantically rich embedding space. This shared space shows remarkable zero-shot transfer abilities across various tasks. However, the inherent Euclidean geometry of these models is not optimal for representing data with implicit hierarchical structures, such as semantic taxonomies. This paper extends these foundational architectures by adopting an alternative geometry specifically tailored to model such structures more effectively.

Hyperbolic Deep Learning. Hyperbolic geometry, a non-Euclidean space with constant negative curvature, is promising for embedding hierarchical data [3,24]. Early work in natural language processing [10, 33] and graph representation learning [11, 12, 21] demonstrated the advantage of hyperbolic embeddings for representing taxonomies like WordNet [23] and large-scale graph structures. These methods commonly utilize the Poincaré ball model to adapt neural network architectures to this curved space. This work applies their geometric principles to learn representations for multi-modal data.

Hyperbolic Vision-Language Models. The success of hyperbolic geometry in unimodal tasks has led to its recent adoption in vision-language modeling. Early efforts [15, 36], such as MERU [9], extended CLIP's contrastive learning framework to the Poincaré ball, introducing hyperbolic VLMs. HyCoCLIP [25] advanced this by explicitly enforcing compositional and hierarchical relationships using hyperbolic entailment cones, moving beyond basic contrastive learning. ATMG [30] explored the modality gap in hyperbolic space, proposing mitigation strategies for this issue, often worsened by the space's curvature. Recent works [5, 22,35] applies the hyperbolic concept to vision large language models or extends to work video modality [16, 17, 34] or adapts to specific tasks [13, 28, 32]. This work complements these efforts by focusing on training stability and evaluation of hierarchical structures within the embedding space, rather than solely on the initial separation of modalities.

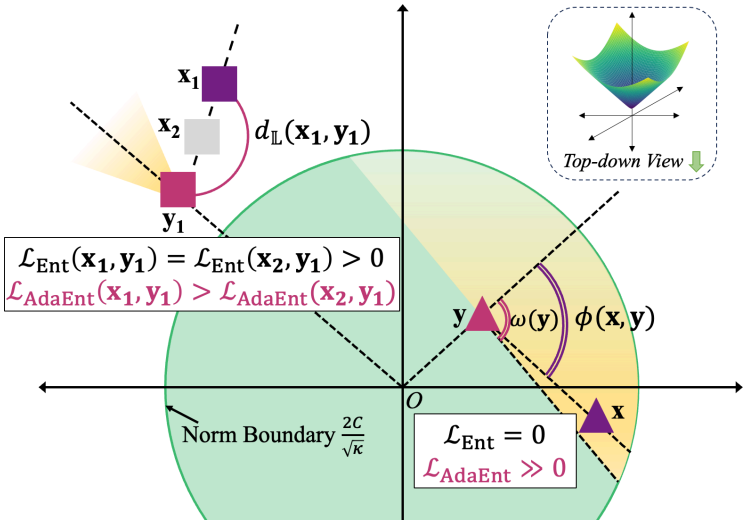


Fig. 2: Behavior of Adaptive Entailment $\mathcal{L}_{\text{AdaEnt}}$ and standard Entailment Loss $\mathcal{L}_{\text{Ent}}$. The figure highlights two cases in a top-down view of hyperbolic space: **(1) Inside the norm boundary** ($\|\tilde{\mathbf{y}}\| \leq \frac{2C}{\sqrt{\kappa}}$): The standard $\mathcal{L}_{\text{Ent}}$ collapses to zero for all $\mathbf{x}$ in the non-origin half-space, even when the exterior angle $\phi(\mathbf{x}, \mathbf{y})$ is large. Our $\mathcal{L}_{\text{AdaEnt}}$ remains active ($\mathcal{L}_{\text{AdaEnt}} \gg 0$), preventing vanishing gradients. **(2) Outside the norm boundary**: $\mathcal{L}_{\text{Ent}}$ penalizes the likely noisy positive $\mathbf{x}_2$ and the true negative $\mathbf{x}_1$ with the same value. Our $\mathcal{L}_{\text{AdaEnt}}$ adaptively assigns a lower loss to $\mathbf{x}_2$ while strongly penalizing $\mathbf{x}_1$.

Evaluation of Hierarchical Reasoning. Existing methods for evaluating hierarchical reasoning often rely on suboptimal benchmarks. WordNet-based evaluations [25] are typically restricted to simple, single-level relationships and depend on a single, manually curated taxonomy. While more recent datasets like HierarCaps [1] offer multi-level captions, their evaluation protocol has fundamental flaws. Specifically, conventional retrieval-based metrics fail to account for valid hierarchical relationships with items outside of the query's ground-truth pairs, thus hindering the accurate measurement of a model's true generalization capabilities. This work directly addresses these limitations by introducing a novel and more robust evaluation protocol designed to precisely assess a model's comprehension of hierarchical entailment.

3 Background

This section provides the necessary background on the Lorentz model of hyperbolic geometry and its application to hierarchical representation learning.

Hyperbolic Representations in the Lorentz Model. To effectively embed data with a tree-like structure, MERU [9] utilize hyperbolic geometry, which offers a powerful inductive bias due to its constant negative curvature and ex-

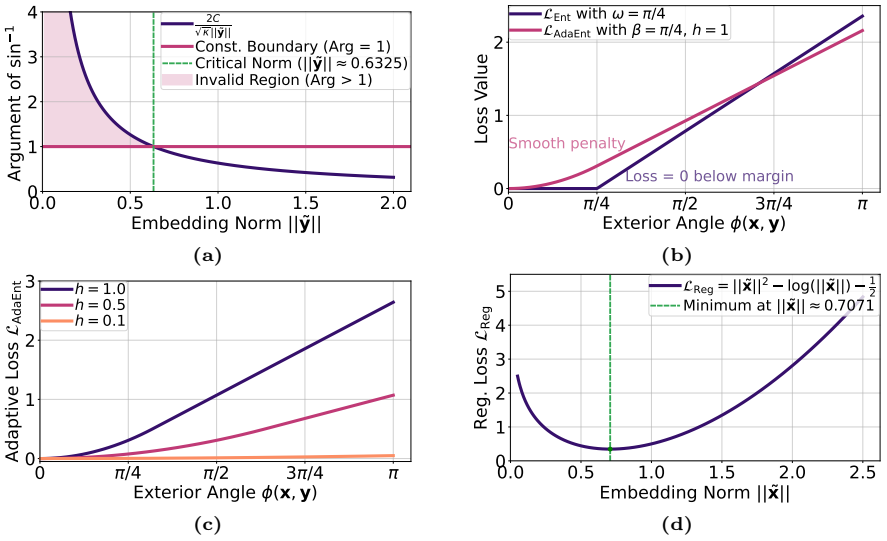


Fig. 3: Analysis of the vanilla and proposed adaptive entailment loss. (a) The norm constraint required by the standard loss. (b) A comparison between loss functions. (c) The effect of our adaptive weight. (d) The behavior of our norm regularizer.

ponential volume growth. MERU selects the Lorentz model (or the hyperboloid model), a well-established framework for hyperbolic geometry. This model defines the n-dimensional hyperbolic space $\mathbb{L}^n$ as a specific manifold within the $(n+1)$-dimensional Minkowski spacetime, $\mathbb{R}^{n+1}$. The Lorentzian inner product governs the geometry of this spacetime, defined for vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n+1}$ as:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}} = -x_0 y_0 + \langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle_{\mathbb{E}} \tag{1}$$

where x_0, y_0 are the first *time* coordinates and $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}$ are the remaining n *spatial* coordinates. The hyperbolic manifold $\mathbb{L}^n$ is then the set of points on the upper sheet of a two-sheeted hyperboloid:

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

Here, $-\kappa \in \mathbb{R}$ represents the curvature of the space. For any point $\mathbf{x}$ on this manifold, its time coordinate is fully determined by its spatial coordinates via the relation $x_0 = \sqrt{1/\kappa + \|\tilde{\mathbf{x}}\|^2}$. The distance between two points is measured along the geodesic connecting them, calculated as:

$$d_{\mathbb{L}}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\kappa}} \cosh^{-1}\left(-\kappa \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbb{L}}\right). \tag{3}$$

Vanilla deep learning encoders produce vectors in Euclidean space. To map these vectors onto the hyperbolic manifold, MERU predicts in the tangent space at the origin of $\mathbb{L}^n$ and projects them using the exponential map, $\exp_{\mathbf{o}}^{\kappa}$.

Learning Alignment and Hierarchy. Prior works [9,25] focus on two primary objectives: aligning related concepts (*e.g.* an image of a cat and the text "a photo of a cat") and hierarchical ordering (*e.g.* "cat" is a type of "animal", the crop of cat region is more generic than the full image of a cat and a dog). These methods align image-text via contrastive learning. Given a batch B of image-text pairs (I, T), a standard contrastive loss function $\mathcal{L}_{\text{Cont}}$ pulls corresponding representations together while pushing non-corresponding pairs apart:

$$\mathcal{L}_{\text{Cont}}(I,T) = -\sum_{i \in B} \log \frac{\exp\left(s(I_i, T_i)/\tau\right)}{\sum\limits_{j \in B} \exp(s(I_i, T_j)/\tau)}, \tag{4}$$

where the similarity score $s(\mathbf{x}, \mathbf{y})$ is the negative geodesic distance, $s(\mathbf{x}, \mathbf{y}) = -d_{\mathbb{L}}(\mathbf{x}, \mathbf{y})$ and τ is a learnable temperature parameter.

To explicitly model "is-a" relationships, HyCoCLIP [25] utilizes a geometric construct of an entailment cone. For any given concept embedding $\mathbf{y}$, its entailment cone Λ_q defines a region in the space. The learning objective is to position embeddings of more specific concepts (*e.g. "poodle"*, $\mathbf{x}$) inside the cones of more general concepts (*e.g. "dog"*, $\mathbf{y}$). The size of this cone is determined by its half-aperture, $\omega(\mathbf{y})$:

$$\omega(\mathbf{y}) = \sin^{-1}\left(\frac{2C}{\sqrt{\kappa}\|\tilde{\mathbf{y}}\|}\right), \tag{5}$$

where the constant C=0.1 [25] keeps the embedding closer to the origin. This formulation is position-dependent; semantically more general concepts that lie closer to the origin (smaller $\|\tilde{\mathbf{y}}\|$) are assigned wider cones. An entailment loss, $\mathcal{L}_{\text{ent}}$, enforces this structure by penalizing cases where a specific concept $\mathbf{x}$ lies outside the cone of a general concept $\mathbf{y}$:

$$\mathcal{L}_{\text{Ent}}(\mathbf{x}, \mathbf{y}) = \max(0, \phi(\mathbf{x}, \mathbf{y}) - \eta\omega(\mathbf{y})) \tag{6}$$

where $\phi(\mathbf{x}, \mathbf{y}) = \pi - \angle\mathbf{Oyx}$ is the exterior angle of $\mathbf{x}$ and η is a scaling factor [25]. Prior works [25] apply this loss to various pairs, such as *(image, text)*, *(image crop, full image)*, and *(text phrase, full caption)*. We show many of these concepts pictorially in Fig. 2.

4 ARGENT

We propose a novel adaptive entailment learning method to improve the alignment of latent representations in hyperbolic space. Specifically, we introduce our adaptive entailment loss, which is designed to overcome the issues of cone collapse and instability in prior work, thereby facilitating significantly more robust hierarchical learning.

A core issue with the standard entailment loss $\mathcal{L}_{\text{Ent}}$ in Eq. (6), which hinders learning meaningful representations, stems from the definition of the half-aperture in Eq. (5). The $\sin^{-1}$ function is only defined if its argument is ≤ 1, which imposes the constraint $\|\tilde{\mathbf{y}}\| \geq \frac{2C}{\sqrt{\kappa}}$. With typical converged values of C=0.1 and κ=0.1 [25], this requires $\|\tilde{\mathbf{y}}\| \geq 0.6325$. Fig. 3a illustrates this invalid domain

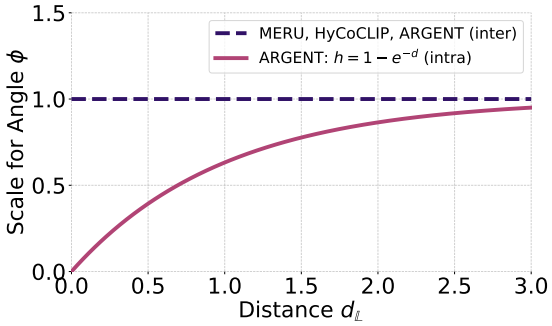


Fig. 4: Weighing factor $h(\mathbf{x}, \mathbf{y})$ (scaling) w.r.t. the distance (negative of similarity) for intra and inter-modality samples in MERU, HyCoCLIP and ARGENT. **ARGENT uses adaptive weights** compared to constant weights in MERU and HyCoCLIP.

(shaded region). However, many embeddings in previous works [9, 25] fall into the invalid region, violating this constraint. In such cases, the aperture is clipped to $\pi/2$, transforming the cone into a half-space and reducing the entailment objective to a simple norm constraint: $\|\tilde{\mathbf{y}}\| \leq \|\tilde{\mathbf{x}}\|$. The hyperparameter η in Eq. (6) of HyCoCLIP [25] thus serves as an ad-hoc fix, manually tuning cone sizes to manage these geometric inconsistencies.

To address this issue, we introduce a more direct and robust ***Adaptive Entailment Loss***. For a specific concept $\mathbf{x}$ and a general concept $\mathbf{y}$, we formulate the loss $\mathcal{L}_{\text{AdaEnt}}$ as:

$$\mathcal{L}_{\text{AdaEnt}} = \text{Huber}\left(h(\mathbf{x}, \mathbf{y}) \cdot \phi(\mathbf{x}, \mathbf{y}), \beta\right) \tag{7}$$

where β is the Huber loss delta. Fig. 3b shows that this formulation directly minimizes the exterior angle and provides a smoother loss surface than $\mathcal{L}_{\text{ent}}$, without relying on the problematic aperture calculation.

The adaptive weight term $h(\mathbf{x}, \mathbf{y})$ modulates the loss strength, with its effect visualized in Fig. 3c. For inter-modality pairs (*e.g.* image-text), we set $h(\mathbf{x}, \mathbf{y}) = 1$. For intra-modality pairs (*e.g.* a crop and its full image), we define:

$$h(\mathbf{x}, \mathbf{y}) = 1 - \exp(s(\mathbf{x}, \mathbf{y})) = 1 - \exp(-d_{\mathbb{L}}(\mathbf{x}, \mathbf{y})) \tag{8}$$

This adaptive weighting prevents the entailment loss from pushing near-identical pairs apart, such as when a crop is almost identical to the full image. Thus, the adaptive loss removes explicit geometric constraints of the entailment cone. Fig. 4 visualizes the adaptive term $h(\mathbf{x}, \mathbf{y})$ against the distance $d_{\mathbb{L}}(\mathbf{x}, \mathbf{y})$ across models.

While improving stability, it drifts the embedding away from the origin, creating an overly sparse representation space. To counteract this, we introduce a regularization term: $\mathcal{L}_{\text{Reg}}(\mathbf{x}) = \|\tilde{\mathbf{x}}\|^2 - \log(\|\tilde{\mathbf{x}}\|)$. Fig. 3d shows that this regularizer encourages embedding norms to remain within a stable range by penalizing values that are either too small or too large. Our training objective is a

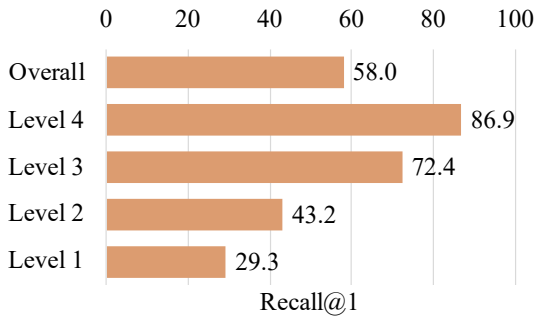


Fig. 5: Recall@1 of the BLIP-L model in image-to-text retrieval task on HierarCaps. Level 1 and 4 denotes the most generic and the most specific captions. We restrict the candidate pool to a local set containing ground-truth captions and the top-5 most similar (presumed false negative) captions. **The model performance degrades when the text is more generic (Level 1).**

combination of the contrastive, adaptive entailment, and regularization losses: $\mathcal{L} = \mathcal{L}_{\text{Cont}} + \gamma_1 \mathcal{L}_{\text{AdaEnt}} + \gamma_2 \mathcal{L}_{\text{Reg}}$ where, γ_1, γ_2 are the weighing terms.

5 Revisiting Hierarchical Entailment Evaluation

To our knowledge, HierarCaps [1] is the only benchmark for evaluating multi-level, hierarchical image-text representations. It provides semi-automatically generated captions at varying detail levels to evaluate entailment. However, we identify two critical, previously undiscussed limitations in its evaluation protocols:

Insensitivity of Ranking-Based Metrics. The primary ranking metric, Kendall's Correlation, is insensitive to the magnitude of ordering errors. For example, consider four captions ordered from coarse to fine: *"people"; "a group of people", "a group of people milling about inside a large space."; "a group of people milling about inside a large space in the winter"*. This metric assigns the same penalty for a swap between the two coarsest levels (*e.g.* *"people"* and *"a group of people"*) as it does for a swap between the two finest levels. Intuitively, the latter error is more severe as the captions differ by a greater amount of semantic detail. Furthermore, the metric fails to capture the degree of entailment. If two models produce the same correct ordering, Kendall's correlation cannot differentiate which model better represents the semantic hierarchy.

Ambiguity in Retrieval-Based Evaluation. Image-to-text retrieval protocols use a global candidate pool that mixes true negatives with many false negatives. We observe cases where captions from visually similar images provide (near) valid descriptions, yet evaluation counts them as distractors and penalizes them. We perform two analyses to quantify this impact.

In our first analysis, we use a BLIP-L-ITM [18] model finetuned on COCO [20] to retrieve the top-5 most similar images for each image in the benchmark. For

Table 1: HierarCaps image-to-text retrieval performance of $\mathbf{CLIP}^{\mathbf{L}}_{\mathbf{FT}}$ increases as false negative captions are progressively removed from the candidate pool. This shows that the current evaluation protocol is flawed.

Remove top-k	Precision (%) ↑	Recall (%) ↑
None	15.36	43.23
k=5	17.10 (+1.74)	46.45 (+3.22)
k=10	18.05 (+2.69)	48.10 (+4.87)

every image, we construct a local candidate pool consisting of its ground-truth captions and the captions from these 5 similar images. We then re-evaluate the BLIP model on image-to-text retrieval at each hierarchical level using only this local pool. Fig. 5 shows the Recall@1 from the most generic (Level 1) to the most specific (Level 4). At the most generic level, performance drops to 30%. Although we expect some overlap for generic concepts, this sharp degradation within a small, highly confounding local pool indicates that the original retrieval protocol defines an unreliable evaluation task.

In our second analysis, we evaluate $\text{CLIP}^{\text{L}}_{\text{FT}}$ [1], a model fine-tuned on hierarchical data, using cleaner global candidate pools. This avoids potential artifacts from the BLIP model, which was not trained on hierarchical data. We follow the original HierarCaps protocol but progressively remove captions from the top-k similar images (identified by BLIP) from the global pool. Tab. 1 shows that retrieval performance steadily increases as we remove more of these ambiguous false negatives. This confirms the highly significant and negative impact of dataset ambiguity on the retrieval evaluation.

To reduce the impact of noisy and ambiguous annotations, we construct a cleaned benchmark by filtering out ambiguous image–caption pairs and retaining only high-confidence matches. We apply a model-assisted cleaning pipeline based on Qwen3-VL [2] (verification), Qwen3-VL-Embed and Qwen3-VL-Reranker [19] (similarity); details and thresholds are provided in the supplementary. From 1,000 samples, we retain 522 image–caption pairs with strong alignment and low ambiguity. We also build a negative caption pool by selecting 100 captions with the lowest retrieval degree (i.e., rarely retrieved for other images).

Proposed Probabilistic Entailment Protocol. We propose a new protocol to evaluate hierarchical entailment on the HierarCaps dataset. Our metric quantifies the entailment probability p_{Ent} via the angle ϕ between embeddings:

$$p_{\text{Ent}}(\mathbf{x}, \mathbf{y}) = \max\left(1 - \frac{2\phi(\mathbf{x}, \mathbf{y})}{\pi}, 0\right) \tag{9}$$

This function maps a perfect alignment ($\phi = 0$) to $p_{\text{Ent}} = 1$ and an orthogonal or opposed alignment ($\phi \geq \pi/2$) to $p_{\text{ent}} = 0$. For each image, we compute p_{Ent} using all corresponding hierarchical captions as positives, and use fine-grained captions from the non-ambiguous negative pool as negatives. We then report the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and Average Precision (AP) as the final metrics for quantifying entailment quality.

Table 2: Main Results. ARGENT outperforms baselines on almost all metrics, often by a large margin. The adaptive approach not only improves performance on downstream tasks (classification and retrieval) but also reduces hierarchical errors. Results of all 16 image classification datasets are in supplementary material. [Keys: **Best**, Gain and drop relative to HyCoCLIP baseline [25], Cls= Classifcation, T2I= Text-to-Image, I2T= Image-to-Text, Ret= Retrieval]

Model	Cls (Acc ↑)		T2I Ret (R@k ↑)				I2T Ret (R@k ↑)				Hierarchical Metrics				
			COCO		Flickr		COCO		Flickr		WordNet			PEP	
	INet	Avg.	k=5	k=10	k=5	k=10	k=5	k=10	k=5	k=10	TIE ↓	LCA ↓	Jac ↑	AUC ↑	AP ↑
CLIP-S [29]	32.8	35.5	51.4	63.0	78.3	85.7	66.0	76.5	88.5	93.3	4.39	2.50	73.6	–	–
MERU-S [9]	32.6	34.3	51.2	63.3	77.7	86.3	65.8	76.7	88.2	94.3	4.35	2.48	73.7	58.4	19.9
HyCoCLIP-S [25]	37.3	**39.2**	52.0	63.5	78.5	85.9	65.2	75.9	88.7	92.8	3.96	2.33	76.2	96.8	87.7
ARGENT-S	**38.8**	38.4	**53.2**	**64.7**	**79.9**	**87.5**	**66.9**	**77.8**	**90.4**	**95.6**	**3.76**	**2.26**	**77.7**	**99.3**	**91.2**
Δ	+1.5	-0.8	+1.2	+1.2	+1.4	+1.6	+1.7	+1.9	+0.7	+2.8	-0.20	-0.07	+1.5	+2.5	+3.5
CLIP-B [29]	36.9	37.6	54.8	66.3	81.9	88.6	70.4	79.3	92.1	95.3	3.87	2.30	**84.0**	–	–
MERU-B [9]	36.2	37.0	54.6	66.3	81.1	88.4	68.6	79.1	91.0	95.4	3.98	2.35	**84.0**	55.5	14.4
HyCoCLIP-B [25]	43.1	42.6	56.7	68.0	82.8	89.3	69.5	79.7	91.6	95.9	3.40	2.11	79.9	97.2	88.5
ARGENT-B	**44.0**	**42.8**	**57.9**	**69.2**	**84.4**	**90.7**	**70.8**	**80.9**	**92.9**	**96.2**	**3.36**	**2.10**	80.2	**99.4**	**90.6**
Δ	+0.9	+0.2	+1.2	+1.2	+1.6	+1.4	+1.3	+1.2	+1.3	+0.3	-0.04	-0.01	+0.3	+2.2	+2.1
CLIP-L [29]	39.9	40.6	57.7	69.2	84.6	90.3	70.5	80.5	**93.6**	96.1	3.64	2.22	78.6	–	–
MERU-L [9]	39.6	40.2	57.7	68.8	84.3	90.0	70.9	81.2	91.2	95.8	3.71	2.24	78.0	60.4	23.8
HyCoCLIP-L [25]	43.9	44.4	57.5	68.6	84.2	90.1	70.7	80.2	92.3	95.9	3.32	2.09	80.5	98.0	89.5
ARGENT-L	**45.6**	**45.1**	**58.6**	**69.7**	**84.9**	**90.6**	**71.7**	**81.7**	92.8	**96.8**	**3.15**	**2.03**	**81.6**	**99.5**	**90.3**
Δ	+1.7	+0.7	+1.1	+1.1	+0.7	+0.5	+1.0	+1.5	+0.5	+0.9	-0.17	-0.06	+1.1	+1.5	+0.8

6 Experiments

Training Datasets. Our experiments primarily utilize the GRIT dataset [26], which comprises $20.5M$ grounded vision-language pairs. We train all models with images and captions, with the exception of HyCoCLIP, which also incorporates box information. For ablations, we apply HyCoCLIP's data generation pipeline to create a smaller CC3M dataset [31].

Benchmarks and Metrics. We evaluate our models on a suite of standard downstream tasks including zero-shot image classification and image-text retrieval following previous works [9, 25]:

- **Zero-Shot Classification:** We use a set of 16 widely used classification datasets and report top-1 accuracy on ImageNet [8], along with the overall average across all 16 datasets.
- **Image-Text Retrieval:** We report standard metrics, Recall@5 and Recall@10 (R@5, R@10) for both image-to-text and text-to-image retrieval on Flickr30k [27] and COCO [20].
- **Hierarchical Metrics (WordNet):** Leveraging WordNet [23,25] to evaluate the hierarchy of ImageNet predictions, we benchmark models using set-based metrics: TIE, LCA, Precision, Recall, and Jaccard Index.
- **Hierarchical Metrics (Ours):** AUC-ROC and AP for our proposed probabilistic entailment protocol.

Model Architecture. We extended the publicly available HyCoCLIP [25] codebase, employing a dual-encoder CLIP architecture [9, 29] consistent with estab-

lished practices. The image encoder processes 224×224 inputs with a 16×16 patch size, and the text encoder handles a maximum of 77 tokens. This architecture was adapted to hyperbolic space to establish MERU [9] and HyCoCLIP [25] as baselines. While MERU learns representations for full image-caption pairs, HyCoCLIP further utilizes phrases and image crops to impose hierarchical structure. Our proposed model, ARGENT, integrates our adaptive entailment loss ($\mathcal{L}_{\text{AdaEnt}}$) into the HyCoCLIP framework. For $\mathcal{L}_{\text{AdaEnt}}$, we empirically set both weighing terms γ_1 and γ_2 to 0.1. We experimented with Small (S), Base (B), and Large (L) model sizes. More details are available in the supplementary material.

6.1 Quantitative Results

Tab. 2 compares ARGENT against all baselines trained on the GRIT dataset. ARGENT consistently achieves substantial performance improvements, including gains of up to 2.8% in average zero-shot classification and significant enhancements across all retrieval tasks. Furthermore, ARGENT shows consistent improvements in hierarchical metrics, specifically the set-based WordNet scores (TIE, LCA, Jaccard), and achieves a superior AUC on our evaluation metrics. ARGENT proves robust across various model scales (Small, Base, Large), consistently achieving top scores in nearly all configurations. Notably, increasing model size correlates with more pronounced gains in downstream classification tasks (*e.g.*, a +1.7% increase on ImageNet for the Large model), while the Average Precision (AP) on HierarCaps decreases. This observation suggests a potential trade-off between optimizing for broad downstream performance and fine-grained hierarchical representation, an area reserved for future investigation.

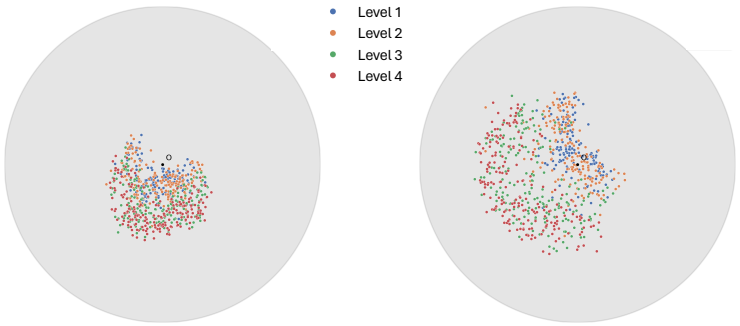
6.2 Hyperbolic Space Analysis

We next report the impact of the adaptive entailment loss by a comparative analysis of the embedding structures learned by HyCoCLIP and ARGENT. Utilizing HoroPCA [4], a specialized hyperbolic dimensionality reduction technique, we project the learned embeddings into 2D Poincaré space and visualize their norm distributions in Fig. 6.

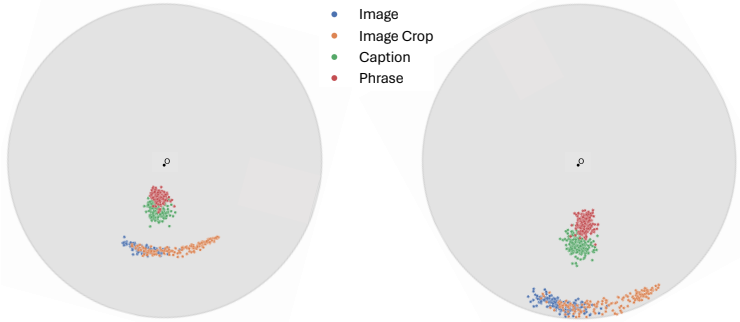
Fig. 6a shows HierarCaps text embeddings, revealing ARGENT's ability to produce a well-defined hierarchy with clear specificity level separation, unlike HyCoCLIP's collapsed embeddings near the origin. This highlights the effectiveness of our loss in structuring and stabilizing the hyperbolic embedding space.

Similarly, Fig. 6b, displaying HoroPCA projections of CC3M components (*images, crops, captions, phrases*), demonstrates ARGENT's superior geometric structure by exhibiting clearer decision boundaries between captions and phrases and a broader spatial distribution of embeddings, suggesting a higher-capacity representation space.

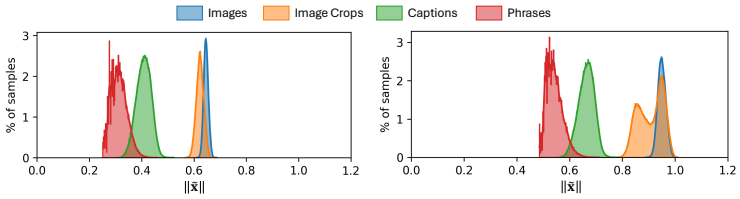
Further insights from Fig. 6c's norm distributions indicate that ARGENT increases embedding norm variance and reduces overlap between caption and phrase distributions. The most significant difference, however, appears in image



(a) HierarCaps: HoroPCA on different text embedding levels.



(b) CC3M: HoroPCA on different data components.



(c) CC3M: Embedding norm distribution.

Fig. 6: HoroPCA and norm distribution of HyCoCLIP (*left*) and ARGENT (*right*) on HierarCaps and CC3M. **ARGENT increases the variance of embedding norms and reduces the overlap between caption and phrase distributions.**

and crop distributions: while the baseline shows considerable overlap, our model generates a bimodal distribution for crops. One peak aligns with whole images (likely large, salient crops), and a distinct second peak is positioned further away from the origin, showcasing the effectiveness of our adaptive weight h.

6.3 Ablation Studies

We perform ablation studies on our smaller-scale CC3M dataset, training for $40,000$ iterations ($30M$ data points) to isolate the effect of each component. Unless otherwise specified, we keep all hyperparameters identical to the main experiments and modify only the component under study.

Table 3: Ablation studies of proposed components. Our proposed $\mathcal{L}_{\text{AdaEnt}}$ outperforms standard entailment loss in (a). The adaptive wight is the most efficient when applying on intra-modality pairs in (b). Regularization loss balances the performance of the model on tasks in (d). [Key: **Best** in each group].

Model	Module	Cls.	Ret.	WordNet		PEP	
		Acc. ↑	R@5 ↑	Err. ↓	Jacc. ↑	AUC ↑	AP ↑
(a) Adaptive Entailment loss with Different Backbones							
MERU-S	Ent	**20.8**	52.6	5.07	59.2	73.5	45.0
	AdaEnt	20.6	**53.4**	**4.98**	**59.4**	**95.5**	**60.9**
HyCoCLIP-S	Ent	**22.9**	54.8	4.73	62.2	94.6	76.2
	AdaEnt	22.8	**55.5**	**4.60**	**62.8**	**97.3**	**76.9**
HyCoCLIP-B	Ent	**25.4**	59.6	4.41	64.4	94.2	76.4
	AdaEnt	25.3	**60.2**	**4.28**	**65.7**	**97.9**	**78.2**
HyCoCLIP-L	Ent	25.8	62.1	4.37	**65.5**	93.6	76.0
	AdaEnt	**26.2**	**63.0**	**4.32**	65.2	**97.9**	**79.2**
(b) Entailment Relationship applied Adaptive Weight h							
MERU-S	None	20.6	**53.4**	**4.98**	**59.4**	**95.5**	**60.9**
	Any	**21.1**	53.1	5.07	59.0	71.1	44.4
HyCoCLIP-S	None	20.9	51.3	4.87	60.4	94.9	74.3
	Any	21.6	54.9	4.68	62.0	71.1	40.2
	Intra	**22.8**	**55.5**	**4.60**	**62.8**	**97.3**	**76.9**
(c) Regularization loss $\mathcal{L}_{Reg}$							
MERU-S	✗	**20.8**	53.2	5.11	58.8	94.8	58.7
	✓	20.6	**53.4**	**4.98**	**59.4**	**95.5**	**60.9**
HyCoCLIP-S	✗	20.8	52.6	4.85	61.0	96.2	75.6
	✓	**22.8**	**55.5**	**4.60**	**62.8**	**97.3**	**76.9**

Efficacy of Adaptive Entailment Loss. Tab. 3(a) shows that replacing the vanilla entailment loss $\mathcal{L}_{\text{Ent}}$ (Ent) with $\mathcal{L}_{\text{AdaEnt}}$ (AdaEnt) consistently improves performance on both MERU and HyCoCLIP backbones, with particularly strong gains on hierarchical metrics. AdaEnt also boosts downstream retrieval and, for larger models, improves classification performance.

Adaptive Weight is Crucial for Intra-Modality Entailment. Tab. 3(b) confirms that applying the adaptive weight h to all pairs ('Any') degrades performance, since it incorrectly weakens the crucial inter-modality constraints. Restricting h to intra-modality pairs only ('Intra') achieves the highest performance across all tasks, validating our choice.

Regularization Loss Stabilizes Performance. Tab. 3(c) shows that removing $\mathcal{L}_{\text{Reg}}$ lowers performance on both downstream tasks and WordNet-based hierarchical scores. This behavior is expected: without regularization, the embedding space becomes overly sparse, which harms generalization in zero-shot tasks.

7 Conclusion

In this work, we tackle key limitations in both the training and evaluation of hierarchical vision-language representations. We introduce the Adaptive Entailment Loss ($\mathcal{L}_{\text{AdaEnt}}$), for learning hierarchical relationships that directly minimizes angular distance, thereby avoiding the constraint violations in prior cone-based losses. Concurrently, we propose a new, more reliable evaluation protocol PEP that reformulates hierarchical assessment as a probabilistic entailment classification task. Our model, ARGENT, consistently outperforms previous baselines on most downstream tasks and hierarchical metrics. Furthermore, our qualitative analysis of the embedding space, using HoroPCA visualizations, substantiates ARGENT's ability to learn a more clearly separated and semantically meaningful hierarchical structure. Our new protocol tackles the critical ambiguities identified in the HierarCaps benchmark. By transforming the evaluation from a noisy retrieval task to a probabilistic classification, we leverage standard, robust metrics like AUC-ROC and AP to better quantify a model's true entailment quality.

Limitations and Future Work: Our adaptive loss, removes the explicit geometric constraints of prior work, which increases embedding sparsity and necessitates using $\mathcal{L}_{\text{Reg}}$ regularizer. This introduces a new, albeit manageable, trade-off in balancing the different loss components during training. Future works involve creating a large-scale, and high-quality benchmark to advance research in hierarchical modeling and retrieval with the new evaluation protocol.

References

1. Alper, M., Averbuch-Elor, H.: Emergent visual-semantic hierarchies in image-text representations. In: ECCV (2024)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Chamberlain, B., Clough, J., Deisenroth, M.: Neural embeddings of graphs in hyperbolic space. arXiv preprint arXiv:1705.10359 (2017)
4. Chami, I., Gu, A., Nguyen, D., Ré, C.: HoroPCA: Hyperbolic dimensionality reduction via horospherical projections. In: ICML (2021)
5. Chen, N., Su, X., Bao, F.: Hyperbolic representations for prompt learning. In: COLING (2024)
6. Chou, J., Alam, N.: Embedding geometries of contrastive language-image pre-training. In: ECCV (2024)
7. Chun, S., Kim, W., Park, S., Yun, S.: Probabilistic language-image pre-training. In: ICLR (2025)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
9. Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., Vedantam, R.: Hyperbolic image-text representations. In: ICML (2023)
10. Dhingra, B., Shallue, C., Norouzi, M., Dai, A., Dahl, G.: Embedding text in hyperbolic spaces. arXiv preprint arXiv:1806.04313 (2018)

11. Flaborea, A., di Melendugno, G., D'Arrigo, S., Sterpa, M., Sampieri, A., Galasso, F.: Contracting skeletal kinematics for human-related video anomaly detection. PR (2024)
12. Franco, L., Mandica, P., Munjal, B., Galasso, F.: Hyperbolic self-paced learning for self-supervised skeleton-based action representations. arXiv preprint arXiv:2303.06242 (2023)
13. Gonzalez-Jimenez, A., Lionetti, S., Amruthalingam, L., Gottfrois, P., Gröger, F., Pouly, M., Navarini, A.: Is hyperbolic space all you need for medical anomaly detection? In: MICCAI (2025)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
15. Kim, W., Chun, S., Kim, T., Han, D., Yun, S.: Hype: Hyperbolic entailment filtering for underspecified images and texts. In: ECCV (2024)
16. Li, J., Wang, J., Tan, C., Lian, N., Chen, L., Wang, Y., Zhang, M., Xia, S.T., Chen, B.: Enhancing partially relevant video retrieval with hyperbolic learning. In: ICCV (2025)
17. Li, J., Wang, J., Tan, C., Lian, N., Chen, L., Wang, Y., Zhang, M., Xia, S.T., Chen, B.: Hlformer: Enhancing partially relevant video retrieval with hyperbolic learning. arXiv preprint arXiv:2507.17402 (2025)
18. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
19. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
21. Liu, Q., Nickel, M., Kiela, D.: Hyperbolic graph neural networks. NeurIPS (2019)
22. Mandica, P., Franco, L., Kallidromitis, K., Petryk, S., Galasso, F.: Hyperbolic learning with multimodal large language models. In: ECCV (2024)
23. Miller, G.: Wordnet: a lexical database for english. Communications of the ACM (1995)
24. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. NeurIPS (2017)
25. Pal, A., Spengler, M., Melendugno, G., Flaborea, A., Galasso, F., Mettes, P.: Compositional entailment learning for hyperbolic vision-language models. In: ICLR (2025)
26. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
27. Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV (2015)
28. Qiu, Z., Liu, J., Chen, Y., King, I.: Hihpq: Hierarchical hyperbolic product quantization for unsupervised image retrieval. In: AAAI (2024)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
30. Ramasinghe, S., Shevchenko, V., Avraham, G., Thalaiyasingam, A.: Accept the modality gap: An exploration in the hyperbolic space. In: CVPR (2024)

31. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
32. Shimizu, R., Wang, Y., Kimura, M., Hirakawa, Y., Wada, T., Saito, Y., McAuley, J.: A fashion item recommendation model in hyperbolic space. In: CVPR (2024)
33. Tifrea, A., Bécigneul, G., Ganea, O.E.: Poincar\'e glove: Hyperbolic word embeddings. arXiv preprint arXiv:1810.06546 (2018)
34. Wen, J., Chen, Y., Shi, R., Ji, W., Yang, M., Gao, D., Yuan, J., Zimmermann, R.: Hover: Hyperbolic video-text retrieval. TIP (2025)
35. Yang, M., Feng, A., Xiong, B., Liu, J., King, I., Ying, R.: Hyperbolic fine-tuning for large language models. arXiv preprint arXiv:2410.04010 (2024)
36. Yang, M., Verma, H., Zhang, D.C., Liu, J., King, I., Ying, R.: Hypformer: Exploring efficient transformer fully in hyperbolic space. In: KDD (2024)