Uncertainty-guided Compositional Alignment with Part-to-Whole Semantic Representativeness in Hyperbolic Vision-Language Models

Hayeon Kim[1,*] Ji Ha Jang[1,*] Junghun James Kim[2] Se Young Chun[1,2,†]
[1] Dept. of Electrical and Computer Engineering, [2] INMC & IPAI
Seoul National University, Republic of Korea
{khy5630, jeeit17, jonghean12, sychun}@snu.ac.kr
*Authors contributed equally. †Corresponding author.

Abstract

While Vision-Language Models (VLMs) have achieved remarkable performance, their Euclidean embeddings remain limited in capturing hierarchical relationships such as part-to-whole or parent-child structures, and often face challenges in multi-object compositional scenarios. Hyperbolic VLMs mitigate this issue by better preserving hierarchical structures and modeling part-whole relations (i.e., whole scene and its part images) through entailment. However, existing approaches do not model that each part has a different level of semantic representativeness to the whole. We propose UNcertainty-guided Compositional Hyperbolic Alignment (UNCHA) for enhancing hyperbolic VLMs. UNCHA models part-to-whole semantic representativeness with hyperbolic uncertainty, by assigning lower uncertainty to more representative parts and higher uncertainty to less representative ones for the whole scene. This representativeness is then incorporated into the contrastive objective with uncertainty-guided weights. Finally, the uncertainty is further calibrated with an entailment loss regularized by entropy-based term. With the proposed losses, UNCHA learns hyperbolic embeddings with more accurate part-whole ordering, capturing the underlying compositional structure in an image and improving its understanding of complex multi-object scenes. UNCHA achieves state-of-the-art performance on zero-shot classification, retrieval, and multi-label classification benchmarks. Our code and models are available at: https://github.com/jeeit17/UNCHA.git.

1. Introduction

Understanding hierarchical structures is essential for capturing complex compositional information efficiently. As well established in cognitive science, human perception relies on part-whole hierarchies [18, 19], enabling generalization by interpreting new inputs through known relational structures [19, 22, 51]. Such hierarchical representations also improve information compression, classification, and inference efficiency [5, 11, 35, 53]. Vision-Language Models (VLMs) such as CLIP [40], ALIGN [23], and ALBEF [29] have demonstrated remarkable performance in image-text matching and shown strong versatility across various downstream tasks. However, owing to their reliance on Euclidean geometry, these models often face distortion of hierarchical structure and dimensionality trade-offs in capturing hierarchical or complex relational structures [14, 35, 50]. Moreover, CLIP has been reported to exhibit bias and difficulty with compositional relations in complex multi-object scenes [1], which is partly due to the lack of modeling part-whole relations.

Hyperbolic space, characterized by constant negative curvature and exponential volume growth, provides an efficient geometric foundation for embedding hierarchical and fine-grained relational structures. Motivated by these properties, recent studies [4, 7, 8, 26, 36, 41, 45] have explored hyperbolic geometry in vision-language learning. MERU [7] extended contrastive vision-language learning into hyperbolic space by explicitly modeling entailment relations between text and image pairs. ATMG [41] later demonstrated that proximity-based contrastive losses can hinder hierarchical structure learning and proposed an angle-based alternative. HyCoCLIP [36] extended entailment modeling beyond inter-modal image-text relations by including intra-modal part-whole relationships.

Although hyperbolic approaches have demonstrated improved performance in hierarchy-aware representation learning, they do not model that each part has a different level of semantic representativeness to the whole. In other words, they do not account for the varying degree to which each part is semantically representative of the whole. As illustrated in Fig. 1, part images differ substantially in how well they represent the whole scene. When all parts

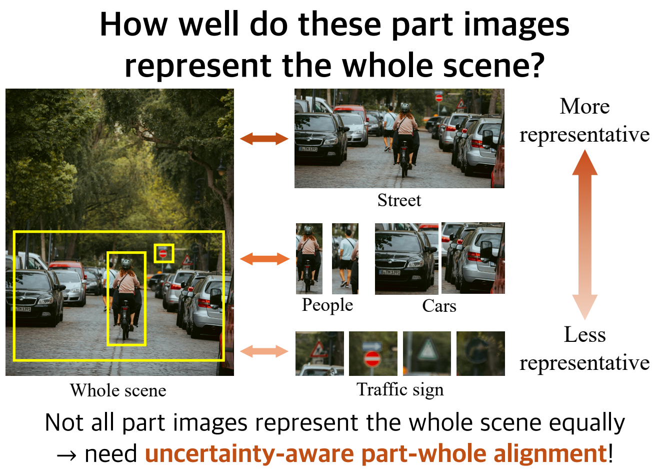


Figure 1. **Varying representativeness of part images to whole scene.** The relationship between each part image and the whole scene varies with its representativeness. We model this varying representativeness as uncertainty, enabling uncertainty-guided part–whole alignment in hyperbolic space.

are treated equally, the model may not appropriately distinguish more representative parts from less representative ones for the whole scene, often leading to degraded multi-object alignment and inefficient utilization of the embedding space [36, 41].

We propose UNcertainty-guided Compositional Hyperbolic Alignment (UNCHA) for enhancing hyperbolic VLMs. UNCHA models part-to-whole semantic representativeness by assigning lower uncertainty to more representative parts and higher uncertainty to less representative ones for the whole scene. This design is grounded in prior findings [2, 10, 34, 54] showing that hyperbolic radius correlates with factors such as abstractness or uncertainty. Then, we incorporate uncertainty as part-to-whole semantic representativeness into both contrastive and entailment loss. Specifically, we incorporate uncertainty into the contrastive objective by assigning part-dependent temperature or uncertainty-guided weights, thereby modulating the strength of each part's alignment with the whole. For the entailment loss, uncertainty is further calibrated based on the degree of part-to-whole entailment, and the entropy-based regularizer is also adapted to stabilize uncertainty estimates and promote richer use of the embedding space. By continually training with the proposed losses, UNCHA progressively strengthens the semantic relationship across parts and wholes, leading to more accurate part-whole ordering in hyperbolic embeddings, capturing the underlying compositional structure in an image and improving its understanding of complex multi-object scenes.

We demonstrated that UNCHA outperforms prior hyperbolic VLMs [7, 36, 41] in diverse downstream tasks such as zero-shot image classification, retrieval, and a range of compositional and multi-object benchmarks, validating UNCHA's modeling of part-to-whole semantic representativeness and capability of more faithful compositional understanding. Our embedding space analysis further confirms UNCHA's more discriminative and efficient use of part-to-whole modeling. The contributions of this work are summarized as:

- We propose UNCHA, a uncertainty-guided compositional alignment with part-to-whole semantic representativeness, enabling hierarchy-aware and compositional representation learning for hyperbolic VLMs.
- We model part-to-whole semantic representativeness with hyperbolic uncertainty, designing uncertainty-guided contrastive and entailment loss for uncertainty calibration, regularized by entropy to adaptively reflect part–whole relations.
- We performed diverse benchmarks, demonstrating that UNCHA achieves superior performance over prior arts in diverse downstream tasks such as zero-shot classification, retrieval, and multi-object classification, validating the effectiveness of our uncertainty-guided compositional alignment.

2. Related Works

2.1. Vision-language models

Vision-Language Models (VLMs) have demonstrated strong capability in aligning image and text representations within a shared semantic space, achieving remarkable performance across tasks such as image-text retrieval and zero-shot image classification. The foundations of these models trace back to early studies on vision-language representation learning such as image retrieval, image captioning, and visual grounding, where joint embedding spaces are learned under task-specific supervision to associate visual content with linguistic semantics [17, 20, 27, 32, 42, 52]. More recently, CLIP [40] introduced a contrastive objective for aligning the two modalities using paired image-text data, achieving strong zero-shot and cross-modal performance [12, 24, 39, 44, 46]. ALIGN [23] and ALBEF [29] further extend CLIP by scaling up weak supervision and incorporating enhanced alignment-fusion strategies to better exploit large-scale, noisy datasets.

However, the inherent limitations of Euclidean space make it difficult to represent hierarchical relationships effectively [21, 35, 37]. Moreover, CLIP has been shown to exhibit biases in complex multi-object scenes [1]. Its text encoder tends to emphasize the object mentioned first in the caption, while its image encoder focuses on larger objects, which hinders performance in multi-object settings. In contrast, hyperbolic space naturally provides continuous tree-like structures that support hierarchical embedding. However, when hierarchical relationships are handled without distinguishing their varying different part-to-whole representativeness, the embeddings tend to lose meaningful structural separation and collapse toward a narrow region [36, 41]. To address this, we introduce a part-to-whole uncertainty-guided alignment framework and explicitly model diverse part-whole entailment relationships within and across modalities, thereby enhancing compositional understanding.

2.2. Hyperbolic representation learning

Hyperbolic space has emerged as an intriguing alternative in representation learning for embedding hierarchies. Hyperbolic space has exponential volume growth and a tree-like geometry, enabling near distortion-free hierarchical embeddings [11, 45]. Therefore, it provides an efficient representation for hierarchical structures. Consequently, numerous studies have leveraged hyperbolic geometry for representing text [8, 28, 48], images [2, 26, 54], and graphs [3, 31, 47]. Recently, hyperbolic space has been integrated into foundation models to better capture hierarchical, compositional, and multi-modal structures at scale, enabling more expressive representations [7, 15, 16, 34, 36, 41]. MERU [7] first introduced hyperbolic vision–language models by employing an additional entailment loss [11, 28] inspired by order embeddings [50] to reflect the informativeness of different modalities. ATMG [41] addressed hierarchical distortion and modality gap caused by spatial proximity–based contrastive learning by introducing an angle-based metric for image-text alignment in hyperbolic space. HyCoCLIP [36] further incorporated intra-modal relationships by considering box images and their corresponding texts.

However, it does not differentiate the varying strengths of these relationships, resulting in limited distinction among parts. Several studies have explored the use of hyperbolic radius, the distance between an embedding and the origin, as a proxy for concept abstractness or uncertainty [2, 10, 34, 54]. The hyperbolic radius naturally provides uncertainty estimation and boundary awareness in pixel-level classification [2, 10], image retrieval [54], and multi-modal language understanding [34], where it serves as an implicit indicator of confidence. Building on this property, we leverage the hyperbolic radius to better encode hierarchical structures in VLM and utilize entailment relationships for effective uncertainty calibration. An entropy-based regularizer further stabilizes the calibrated uncertainty, enabling more efficient use of the embedding space.

3. Method

3.1. Preliminaries

Hyperbolic space is a non-Euclidean geometry with a constant negative curvature $-\kappa$ where $\kappa \in \mathbb{R}^+$. Among several equivalent models, we adopt the Lorentz (or hyperboloid) model for embedding. A vector $\mathbf{p} \in \mathbb{R}^{n+1}$ can be expressed in the form $[p_{\text{time}}, \mathbf{p}_{\text{space}}]$, where $\mathbf{p}_{\text{space}} \in \mathbb{R}^n$ and $p_{\text{time}} \in \mathbb{R}$. The Lorentzian inner product between two vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^{n+1}$ is defined as:

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_{\text{time}} q_{\text{time}} + \langle \mathbf{p}_{\text{space}}, \mathbf{q}_{\text{space}} \rangle, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. The n-dimensional Lorentz manifold $\mathbb{L}^n$ is defined as the upper sheet of a two-sheeted hyperboloid in $(n+1)$-dimensional Minkowski space:

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

The geodesic distance between two points $\mathbf{p}, \mathbf{q}$ on the n-dimensional Lorentz manifold $\mathbb{L}^n$ is:

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

The hyperbolic radius of the embedding $\mathbf{p}$ is defined as the geodesic distance from the origin of the hyperboloid $\mathbf{o}$, *i.e.,* $d_{\mathbb{L}}(\mathbf{p}, \mathbf{o})$.

The tangent space at the point $\mathbf{z} \in \mathbb{L}^n$ is defined as:

$$T_{\mathbf{z}}\mathbb{L}^n = \left\{ \mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{v} \rangle_{\mathbb{L}} = 0 \right\}, \tag{4}$$

which consists of Euclidean vectors $\mathbf{v}$ orthogonal to $\mathbf{z}$ under the Lorentzian inner product. The exponential map projects a tangent vector $\mathbf{v} \in T_{\mathbf{z}}\mathbb{L}^n$ onto the manifold as below:

$$\exp_{\mathbf{z}}^{\kappa}(\mathbf{v}) = \cosh(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}})\mathbf{z} + \frac{\sinh(\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{\kappa}\, \|\mathbf{v}\|_{\mathbb{L}}}\mathbf{v}. \tag{5}$$

Conversely, the logarithmic map sends a point $\mathbf{p} \in \mathbb{L}^n$ back to the tangent space at $\mathbf{z}$ as below:

$$\log_{\mathbf{z}}^{\kappa}(\mathbf{p}) = \frac{\cosh^{-1}(-\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})}{\sqrt{(\kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}})^2 - 1}}\, \mathrm{proj}_{\mathbf{z}}(\mathbf{p}) \tag{6}$$

where $\mathrm{proj}_{\mathbf{z}}(\mathbf{p}) = \mathbf{p} + \kappa \langle \mathbf{z}, \mathbf{p} \rangle_{\mathbb{L}} \mathbf{z}$. Here, we consider the case where $\mathbf{z}$ corresponds to the origin of the hyperboloid, $\mathbf{o} = [\sqrt{1/\kappa}, \mathbf{0}]$. In this setting, the time component of vectors in the tangent Euclidean space can be treated as zero, allowing us to parameterize the space component only, which is consistent with the design of prior works [7, 36, 41].

3.2. Uncertainty-guided hyperbolic alignment

Revisiting prior arts in hyperbolic alignment. Prior hyperbolic VLMs [7, 36, 41] extend contrastive vision-language learning by defining entailment relationships. In this hyperbolic geometry, abstract concepts tend to lie closer to the origin and specific ones farther out, with each specific concept constrained to its parent's entailment cone (see Sec 3.2.3 for details). As illustrated in Fig. 2, MERU [7] incorporates an image-text entailment objective following partial-order embeddings [50], where text is considered more abstract than image. HyCoCLIP [36] extends this idea by modeling intra-modal alignment, assuming that part image is more abstract than its corresponding whole scene.

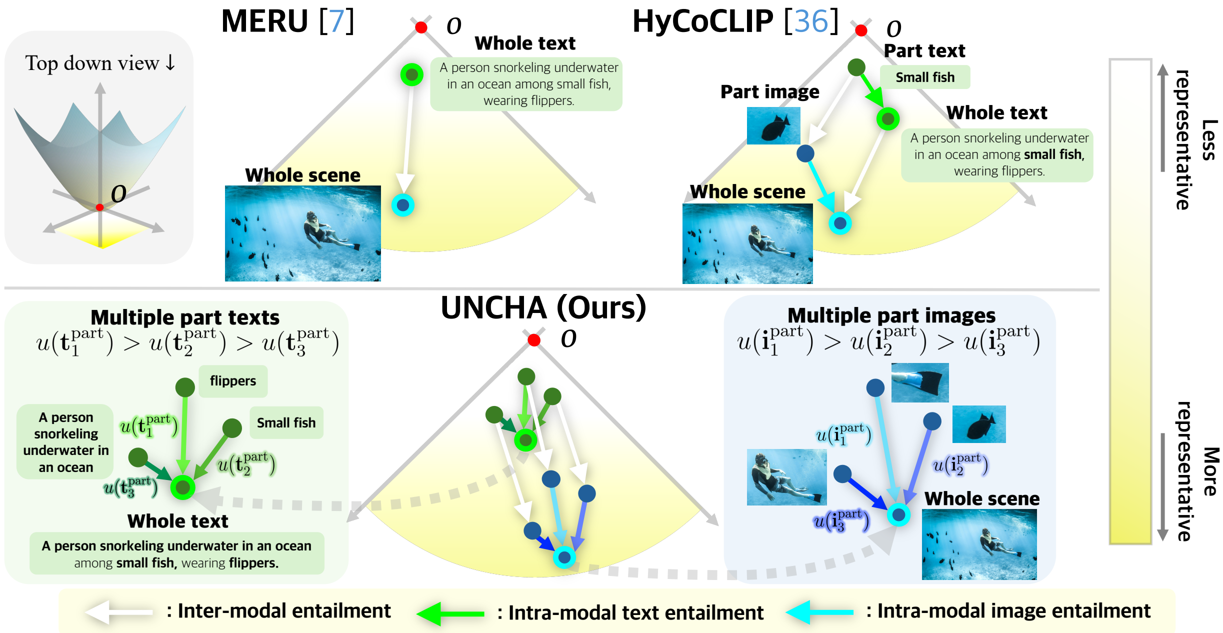


Figure 2. **Comparison of UNcertainty-guided Compositional Hyperbolic Alignment (UNCHA, Ours) with prior works.** MERU [7] models inter-modal entailment between whole scene image and text representations. HyCoCLIP [36] extends this to include intra-modal entailment between part and whole scene representations. UNCHA (Ours) further incorporates *uncertainty to quantify the semantic representativeness* of each part, enabling uncertainty-guided part–whole alignment via adaptive weighting in the contrastive objectives and uncertainty calibration through the entailment loss. In addition, entropy regularization is applied in uncertainty calibration to ensure consistent and balanced utilization of the hyperbolic embedding space across varying uncertainty levels and modalities.

Method overview.

3.2.1. Uncertainty model of semantic representativeness

We leverage the geodesic distance from the origin (radius) in hyperbolic space [2, 10, 34, 54] to quantify the part-to-whole semantic representativeness using hyperbolic uncertainty. Since more abstract concepts are typically located near the origin and more specific ones farther away, this measure naturally reflects representativeness. Thus, we design the hyperbolic uncertainty to assign lower uncertainty to parts that are more representative of the whole scene, and high uncertainty otherwise (*e.g.*, part images). As shown in Fig. 4, our estimated uncertainty well aligns with semantic representativeness, indicating that the model effectively captures the varying part-to-whole relationships.

Specifically, for a point $\mathbf{x} \in \mathbb{L}^n$, the Euclidean norm of $\mathbf{x}$ is monotonically related to its hyperbolic radius (see the supplementary material Sec. S.2.3.1). Accordingly, we define the uncertainty u as follows:

$$u(\mathbf{x}) = \log(1 + \exp(-\|\mathbf{x}\|_2)) \,. \tag{7}$$

Since points near the origin correspond to higher semantic uncertainty, the hyperbolic radius is inversely monotonically related to uncertainty. Eq. 7 is a smooth monotonic transformation of the hyperbolic radius, which is a differentiable, well-behaved uncertainty measure for numerical stability.

3.2.2. Uncertainty-guided contrastive loss

In image–text pretraining, contrastive objectives are commonly employed to align multi-modal representations. Following prior works [7, 36], we adopt the negative Lorentzian distance as the similarity measure as below:

$$L_{\text{c}}^{*}(\mathbf{i}, \mathbf{t}; \tau) = -\sum_{i} \log \frac{\exp\left(-d_{\mathbb{L}}(\mathbf{i}_i, \mathbf{t}_i)/\tau\right)}{\sum_{k \neq i} \exp\left(-d_{\mathbb{L}}(\mathbf{i}_i, \mathbf{t}_k)/\tau\right)} \tag{8}$$

where the i-th image embedding $\mathbf{i}_i$ and its corresponding text embedding $\mathbf{t}_i$ form a *positive* pair while all other text embeddings $\mathbf{t}_i$ with $k \neq i$ are treated as *negatives* in the batch of size B and the temperature parameter τ controls the scaling of similarities.

Prior work [36] introduces a global–local contrastive loss $\mathcal{L}_{\text{con}}^{\text{orig}}$ that aligns part-level text features $\mathbf{t}^{\text{part}}$ with whole image embeddings, and part-level image features $\mathbf{i}^{\text{part}}$ with whole text embeddings as below:

$$\underbrace{L_{\text{c}}^{*}\left(\mathbf{i}^{\text{part}}, \mathbf{t}; \tau\right) + L_{\text{c}}^{*}\left(\mathbf{t}^{\text{part}}, \mathbf{i}; \tau\right)}_{\text{global-local contrastive loss}} + \underbrace{L_{\text{c}}^{*}(\mathbf{i}, \mathbf{t}; \tau) + L_{\text{c}}^{*}(\mathbf{t}, \mathbf{i}; \tau)}_{\text{global contrastive loss}} . \tag{9}$$

Our contrastive loss additionally includes a local contrastive loss that explicitly aligns each part image with its corresponding text on top of Eq. 9. Since whole and part images differ in information levels and occupy distinct regions in hyperbolic space, we design to assign separate temperature parameters, τ_g, τ_l, and τ_{gl} to global, local and global-local contrastive losses, respectively, to better model these relationships.

We propose uncertainty-guided contrastive loss unlike the aforementioned prior contrastive losses with fixed temperature. Our approach incorporates uncertainty into the global-local contrastive loss by considering the varying semantic representativeness of multiple parts. We modulate the temperature in an element-wise manner through an uncertainty-guided global-local contrastive loss, where the temperature is adaptively scaled according to the estimated uncertainty of each part image and text. The adaptive temperatures $\boldsymbol{\tau}^I_{\text{un},i}$ and $\boldsymbol{\tau}^T_{\text{un},i}$ are designed as below:

$$\boldsymbol{\tau}^I_{\text{un},i} = \exp\big(u(\mathbf{i}^{\text{part}}_i)/2\big)\ \tau_{gl}, \boldsymbol{\tau}^T_{\text{un},i} = \exp\big(u(\mathbf{t}^{\text{part}}_i)/2\big)\ \tau_{gl} \tag{10}$$

where higher uncertainty leads to a larger temperature and a smaller contribution to the contrastive loss. The formulation of our proposed contrastive loss is shown as below:

$$\begin{aligned}\mathcal{L}^{\text{un}}_{\text{con}} \;=\;& \underbrace{L^*_{\text{c}}\big(\mathbf{i}^{\text{part}}, \mathbf{t}; \boldsymbol{\tau}^I_{\text{un}}\big) + L^*_{\text{c}}\big(\mathbf{t}^{\text{part}}, \mathbf{i}; \boldsymbol{\tau}^T_{\text{un}}\big)}_{\text{uncertainty-guided global-local contrastive loss}} \\ &+ \underbrace{L^*_{\text{c}}(\mathbf{i}, \mathbf{t}; \tau_g) + L^*_{\text{c}}(\mathbf{t}, \mathbf{i}; \tau_g)}_{\text{global contrastive loss}} \\ &+ \underbrace{L^*_{\text{c}}(\mathbf{i}^{\text{part}}, \mathbf{t}^{\text{part}}; \tau_l) + L^*_{\text{c}}(\mathbf{t}^{\text{part}}, \mathbf{i}^{\text{part}}; \tau_l)}_{\text{local contrastive loss}}.\end{aligned} \tag{11}$$

Unlike the one-to-one correspondence between matched image-text pairs, the relationship between a part image and its whole scene or text may not be a perfect correspondence. For instance, a single scene text may correspond to multiple part images. If all embeddings within a whole scene are pushed apart with the same temperature, both highly representative and less representative regions are equally repelled, breaking semantic structure. Our proposed contrastive loss in Eq. 11 is designed to mitigate these undesirable cases.

3.2.3. Entailment loss for uncertainty calibration

Piecewise-continuous entailment loss. Building upon the hyperbolic entailment formulation in [7, 28], prior work [36] defines the entailment loss as:

$$\mathcal{L}_{\text{orig}} = \max(0, \phi(\mathbf{p}, \mathbf{q}) - \eta\omega(\mathbf{p})) \tag{12}$$

where $\phi(\mathbf{p}, \mathbf{q})$ denotes the angular distance between the embeddings $\mathbf{p}$ and $\mathbf{q}$, η and K are hyperparameters, and $\omega(\mathbf{p})$ defines the aperture of the entailment cone centered at $\mathbf{p}$ as below:

$$\omega(\mathbf{p}) = \sin^{-1}\big(2K/(\sqrt{-\kappa}\|\mathbf{p}\|)\big), \tag{13}$$

which is also illustrated in Fig. 3. The $\mathcal{L}_{\text{orig}}$ in Eq. 12 enforces entailment by constraining $\mathbf{q}$ to lie within the cone of $\mathbf{p}$. However, once $\mathbf{q}$ is fully contained in the cone, the loss becomes zero, preventing further fine-grained alignment.

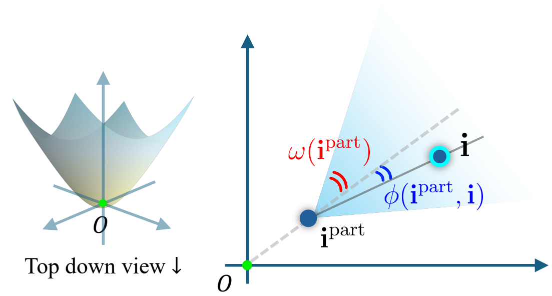


Figure 3. **Entailment geometry in hyperbolic space.** The term $\omega(\mathbf{i}^{\text{part}})$ denotes the aperture of the entailment cone centered at $\mathbf{i}^{\text{part}}$. The angle $\phi(\mathbf{i}^{\text{part}}, \mathbf{i})$ measures the geodesic angle between the embeddings $\mathbf{i}^{\text{part}}$ and $\mathbf{i}$, which is used to determine whether $\mathbf{i}$ lies within the entailment region of $\mathbf{i}^{\text{part}}$.

Here, we propose adding an angular term $\phi(\mathbf{p}, \mathbf{q})$ in Eq. 12 to encourage fine-grained alignment while maintaining smooth optimization continuity as below:

$$L^*_{\text{ent}}(\mathbf{p}, \mathbf{q}) = \max(0,\ \phi(\mathbf{p}, \mathbf{q}) - \eta\,\omega(\mathbf{p})) + \alpha\,\phi(\mathbf{p}, \mathbf{q}) \tag{14}$$

where α is a hyperparameter. This formulation can be viewed as a Leaky-ReLU-like [33] relaxation of the original hinge-based entailment loss, with the additional term preserving a small gradient even when $\mathbf{q}$ is inside the cone.

Uncertainty calibration loss. Prior studies have reported that hyperbolic embeddings often accumulate around narrow regions, leading to collapse [41]. Moreover, local and global image representations exhibit similar radii, making their separation less distinct [36]. To clearly distinct global and local representations, we propose the uncertainty calibration loss as follows:

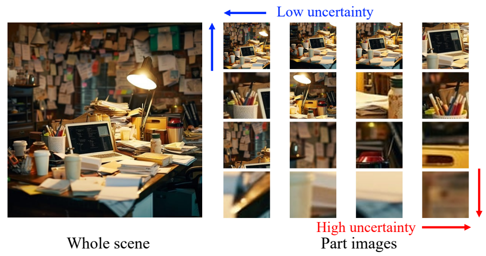
$$L^{\text{cal}}_{\text{ent}}(\mathbf{p}, \mathbf{q}) = \lfloor L^*_{\text{ent}}(\mathbf{p}, \mathbf{q}) \rfloor\, e^{-u(\mathbf{p})} + u(\mathbf{p}) + \mathcal{H}(\tilde{u}(\mathbf{p})) \tag{15}$$

where $\lfloor . \rfloor$ denotes the stop-gradient operator and $\mathcal{H}$ represents the entropy term as follows:

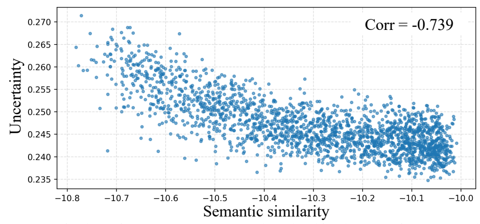
$$\mathcal{H}(\tilde{u}(\mathbf{p})) = -\sum\nolimits_i \tilde{u}(\mathbf{p}_i) \log(\tilde{u}(\mathbf{p}_i)) \tag{16}$$

where $\tilde{u}(\mathbf{p}_i) = \exp(u(\mathbf{p}_i))/\sum_j \exp(u(\mathbf{p}_j))$. When the entailment relation between $\mathbf{p}$ and $\mathbf{q}$ is weak, the term $e^{-u(\mathbf{p})}$ encourages the model to increase uncertainty. The term $u(\mathbf{p})$ prevents the model from assigning excessively high uncertainty just to reduce the loss. Thus, $\mathcal{H}(\tilde{u}(\mathbf{p}))$ regularizes the uncertainty distribution to remain diverse and informative, avoiding a collapse toward uniform or constant uncertainty, analogous to [13].

With the entropy regularizer, the proposed formulation



(a) Part images sorted by UNCHA's calibrated uncertainty



(b) Correlation between semantic similarity and uncertainty

Figure 4. **Analysis of uncertainty modeling.** (a) Randomly cropped parts are sorted by uncertainty (low→high). Semantically representative parts show low uncertainty, while blurred or less representative crops show high uncertainty. (b) On an ImageNet [43] subset, part-to-whole similarity vs. uncertainty shows a strong negative correlation ($r = -0.739$), indicating that less representative parts have higher uncertainty.

of our entailment loss is as follows:

$$\mathcal{L}_{\text{ent}}^{\text{un}} = \underbrace{L_{\text{ent}}^{*}(\mathbf{t}^{\text{part}}, \mathbf{i}^{\text{part}}) + L_{\text{ent}}^{*}(\mathbf{t}, \mathbf{i})}_{\text{inter-modal entailment}} \tag{17}$$
$$+\lambda_1 \underbrace{(L_{\text{ent}}^{*}(\mathbf{t}^{\text{part}}, \mathbf{t}) + L_{\text{ent}}^{*}(\mathbf{i}^{\text{part}}, \mathbf{i}))}_{\text{intra-modal entailment}}$$
$$+\lambda_2 \underbrace{(L_{\text{ent}}^{cal}(\mathbf{t}^{\text{part}}, \mathbf{t}) + L_{\text{ent}}^{cal}(\mathbf{i}^{\text{part}}, \mathbf{i}))}_{\text{uncertainty calibration}}$$

where λ_1 and λ_2 are hyperparameters. This uncertainty calibration enables semantic alignment with the representativeness of each part relative to the whole. This is a process that naturally fits the geometric properties of hyperbolic space, which is particularly beneficial for jointly aligning multiple objects simultaneously. Moreover, such calibration enhances multi-object alignment, as shown in Fig. 4. Parts with higher semantic similarity to the whole exhibit lower uncertainty, while less representative parts show higher uncertainty, resulting in a strong negative correlation between similarity and uncertainty. Further details on Fig. 4 are provided in the supplementary material Sec. S.2.3.2.

Finally, our overall loss with the proposed uncertainty-guided contrastive loss in Eq. 11 and the entailment loss with uncertainty calibration in Eq. 17 is defined as follows:

$$L = \mathcal{L}_{\text{con}}^{\text{un}} + \lambda_{ent} \mathcal{L}_{\text{ent}}^{\text{un}} \tag{18}$$

where λ_{ent} is a hyperparameter. We detail all hyperparameters in the supplementary material Sec. S.1.2.

4. Experiments

4.1. Training details

To ensure a fair comparison, baseline models [7, 36, 40, 41] are reproduced under identical dataset and training configurations, while preserving the optimization settings specified in their original implementations. The batch size and total number of training iterations are fixed at 768 and 500,000, respectively. All models are trained on the Grounded Image-Text Pairs (GRIT) [38] dataset, which contains 20.5 million grounded vision–language pairs and 35.9 million part-level annotations. Detailed descriptions of the settings and hyperparameters are provided in Sec. S.1 of the supplementary material.

4.2. Downstream tasks

4.2.1. Zero-shot image classification

We conduct zero-shot classification experiments on 16 benchmark datasets as listed in Tab. 1. We report Top-1 accuracy as the evaluation metric for all results following prior works [7, 40]. To evaluate scalability, we experiment with different sizes of vision encoders, ViT-S and ViT-B. For ATMG [41], we follow the original setup, computing similarity via averaged exterior angles instead of Lorentz or Euclidean inner products. This configuration is used for all downstream tasks. As shown in Tab. 1, our method consistently outperforms prior approaches across all benchmark datasets, demonstrating generalization and robust performance on downstream tasks.

4.2.2. Zero-shot retrieval

For the retrieval task, we evaluate the model's ability to retrieve the most relevant samples across modalities. Specifically, given an input image (or text), the model retrieves the Top-K text (or image) candidates from the collection, and the retrieval accuracy is computed accordingly. All experiments are conducted under the zero-shot setting using the COCO [30] validation set and the Flickr30K [25, 55] test set. As shown in Tab. 2, our method shows steady performance, indicating its reliable cross-modal alignment capability across both benchmarks.

4.2.3. Hierarchical Classification

To evaluate how well the model embeds hierarchical relationships in hyperbolic space, we adopt the hierarchy-aware metrics introduced in HyCoCLIP [36]. As shown in Tab. 2, our model achieves consistently strong performance in hierarchical metrics, demonstrating its improved ability to preserve the structural hierarchy of the class labels within the embedding space, partly due to the uncertainty-guided alignment. More detailed explanations are in supplementary material Sec. S.2.2.3.

Table 1. **Zero-shot image classification evaluation.** UNCHA (Ours) consistently demonstrates strong zero-shot classification performance across both architectures. Bold numbers denote the best performance within each architecture. † denotes ATMG trained on the GRIT [38].

	Model	General datasets						Fine-grained datasets						Misc. datasets			
		ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
ViT-S/16	CLIP [40]	36.7	70.2	42.6	35.8	57.6	89.7	44.7	9.8	6.9	2.0	44.6	14.8	22.3	**40.7**	40.1	5.1
	MERU [7]	35.4	71.2	40.4	33.8	57.3	89.7	41.2	11.3	5.2	**4.2**	42.7	17.3	18.6	39.1	38.9	**5.3**
	ATMG† [41]	34.1	66.9	42.1	47.9	68.5	90.7	43.6	14.1	5.8	2.5	41.8	14.9	19.7	35.8	40.3	4.6
	HyCoCLIP [36]	41.7	85.0	53.4	52.5	75.7	92.5	50.2	**14.7**	8.1	**4.2**	52.0	20.5	23.3	38.3	**45.7**	5.2
	UNCHA (Ours)	**43.9**	**85.9**	**56.6**	**52.6**	**80.5**	**94.4**	**52.1**	12.5	**9.2**	2.7	**52.1**	**24.6**	**25.4**	36.2	43.4	5.2
ViT-B/16	CLIP [40]	40.6	78.9	48.3	43.0	70.7	92.4	48.3	10.4	9.3	3.4	45.9	21.3	23.4	37.1	42.7	5.7
	MERU [7]	40.1	78.6	49.3	43.0	73.0	92.8	48.5	11.0	5.3	3.7	48.5	21.6	22.1	31.7	42.6	5.4
	ATMG† [41]	34.3	68.8	42.1	48.2	68.5	91.2	43.2	14.3	6.0	2.4	42.2	15.0	19.4	35.0	40.4	4.6
	HyCoCLIP [36]	45.8	88.8	60.1	57.2	81.3	95.0	59.2	**16.4**	11.6	3.7	56.8	23.9	29.4	35.8	45.6	**6.5**
	UNCHA (Ours)	**48.8**	**90.4**	**63.2**	**57.7**	**83.9**	**95.7**	**60.3**	14.8	**14.0**	**3.8**	**57.1**	**27.0**	**30.3**	**41.3**	**52.7**	6.1

Table 2. **Zero-shot retrieval and hierarchical classification metrics on ImageNet [6].** UNCHA (Ours) consistently achieves superior performance across both retrieval and hierarchical metrics, showing the effectiveness of our uncertainty-based hyperbolic alignment.

	Model	Text retrieval				Image retrieval				Hierarchical metrics				
		COCO		Flickr		COCO		Flickr		TIE(↓)	LCA(↓)	J(↑)	$\mathbf{P_H}$(↑)	$\mathbf{R_H}$(↑)
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5					
ViT-S/16	CLIP [40]	69.3	79.1	90.2	**95.2**	53.7	65.2	81.1	87.9	4.02	2.39	0.76	0.83	0.84
	MERU [7]	68.8	78.8	89.4	94.8	53.6	65.3	80.4	87.5	4.08	2.39	0.76	0.83	0.83
	ATMG† [41]	62.6	74.2	85.5	91.6	50.3	62.1	76.9	84.6	4.26	2.50	0.75	0.82	0.83
	HyCoCLIP [36]	69.5	79.5	89.1	93.9	55.2	66.6	81.5	88.1	3.55	2.17	0.79	**0.86**	0.85
	UNCHA (Ours)	**69.9**	**79.7**	**90.8**	94.8	**56.2**	**67.6**	**82.5**	**89.3**	**3.39**	**2.14**	**0.80**	**0.86**	**0.86**
ViT-B/16	CLIP [40]	71.4	81.5	**93.6**	**96.9**	57.4	68.5	83.5	89.9	3.60	2.21	0.79	0.85	0.85
	MERU [7]	72.3	82.0	93.5	96.2	57.4	68.6	84.0	90.0	3.63	2.22	0.78	0.85	0.85
	ATMG† [41]	62.9	74.0	85.1	92.2	51.2	62.6	78.0	85.3	4.19	2.48	0.75	0.83	0.83
	HyCoCLIP [36]	72.0	82.0	92.6	95.4	58.4	69.3	**84.9**	90.3	3.17	2.05	0.81	0.87	0.87
	UNCHA (Ours)	**72.7**	**82.7**	91.4	95.9	**60.0**	**71.0**	**84.9**	**91.2**	**2.94**	**1.96**	**0.83**	**0.88**	**0.88**

Table 3. **Comparison on part-level alignment evaluation with hard negatives.** Ours achieves substantial performance gains under the most challenging scenario of [49], demonstrating its strong ability for fine-grained compositional understanding.

Model	All Pick5		All
	SCM	Neg	Hard Negs
CLIP [40]	13.10	22.94	52.89
ATMG† [41]	12.23	23.08	53.91
MERU [7]	12.59	20.69	54.56
HyCoCLIP [36]	11.65	23.52	53.33
UNCHA (Ours)	**13.53**	**23.81**	**56.51**

4.2.4. Zero-shot multi-label classification

We conduct multi-label classification experiments on the MS-COCO [30] and VOC [9] datasets, as shown in Tab. 5. The evaluation metric is mean Average Precision (mAP). To further assess performance in more complex multi-object settings, we employed the ComCo and SimCo datasets [1]. These datasets evaluate compositional understanding with images containing N objects. ComCo features realistic object compositions, whereas SimCo provides synthetic scenes with diverse geometric shapes. For evaluation, we train a lightweight classifier on the embeddings and reported test-set classification mAP. As shown in Tab. 5,

Table 4. **Ablation study on classification and retrieval benchmarks.** Removing any component leads to consistent performance drops, showing that all modules contribute meaningfully. Bold numbers indicate the best performance within each task group.

Model	Classification			Retrieval	
	General	Fine	MISC.	Text	Image
Ours (full)	**68.98**	**25.53**	**27.55**	**83.80**	**73.90**
w/o uncertainty	64.57	22.98	26.67	79.60	69.68
w/o contrastive	65.14	23.92	25.58	80.78	70.55
w/o entropy	65.61	23.09	24.78	80.60	69.95

Table 5. **Comparison across Multi-object Representation and Classification tasks.** Left: zero-shot mAP comparison across multi-object configurations on ComCo and SimCo datasets. Right: zero-shot multi-label classification (Cls.) on VOC and COCO datasets (mAP only). Our method consistently achieves higher mAP across both tasks.

	Model	**Multi-object Representation**								**Multi-label Cls.**	
		ComCo				**SimCo**				**VOC**	**COCO**
		2 obj.	**3 obj.**	**4 obj.**	**5 obj.**	**2 obj.**	**3 obj.**	**4 obj.**	**5 obj.**		
ViT-B/16	CLIP [40]	77.55	80.31	81.41	80.22	77.15	84.58	87.40	88.48	78.56	53.94
	MERU [7]	72.90	77.25	78.15	77.34	77.82	83.91	85.79	86.90	79.50	54.39
	ATMG† [41]	45.91	45.97	45.80	45.82	65.52	65.32	65.28	65.12	72.22	46.81
	HyCoCLIP [36]	72.90	73.22	73.51	72.90	75.71	81.13	82.41	82.85	80.43	58.12
	UNCHA (Ours)	**77.92**	**80.96**	**81.83**	**81.18**	**79.72**	**86.93**	**89.75**	**90.65**	**82.14**	**59.43**

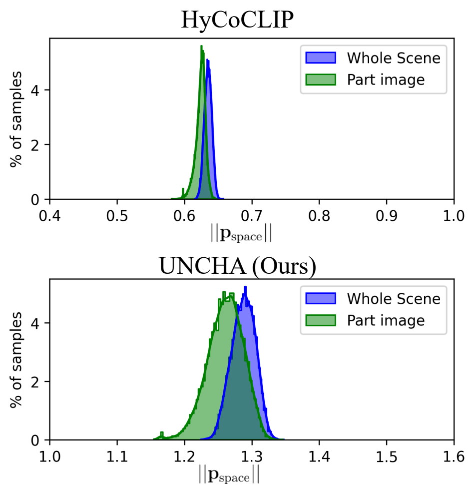


Figure 5. **Analysis of hyperbolic embedding.** Compared to HyCoCLIP [36], whose hyperbolic embeddings exhibit a narrower range, UNCHA yields a more dispersed and structured distribution, reflecting richer use of the hyperbolic space.

UNCHA outperforms all baselines across both multi-label classification and multi-object representation benchmarks which indicate that our uncertainty-aware modeling provides a substantially stronger compositional understanding. These results highlight UNCHA's ability to better disentangle object-level semantics and maintain robust alignment in complex multi-object scenes.

4.2.5. Part-level alignment with hard negatives

We evaluate part-level text–image matching using the benchmark derived from the densely annotated Densely Captioned Images [49]. The benchmark pairs cropped parts with their corresponding texts and introduces region-specific hard negatives to test fine-grained alignment. We report results on the 'All Pick5' and 'All Hard negs' in Tab. 3, which require the model not only to identify the correct pair among hard negative captions but also to produce a correct ordering between matching and non-matching pairs. UNCHA (Ours) achieves the highest performance among baselines, exhibiting substantial improvements in the 'All Pick5' setting. This shows that our model effectively captures fine-grained part-whole distinctions, yielding better region-level visual-semantic alignment.

4.3. Analysis about hyperbolic space

We visualize the radii of hyperbolic embedding for $10,000$ ImageNet [43] images and their randomly cropped parts, shown in Fig. 5. As noted in HyCoCLIP [36], the embeddings of image and their parts often collapse into a narrowly concentrated region, yielding minimal separation between part and whole. In contrast, UNCHA produces a more distinctive and semantically structured geometry: part embeddings consistently lie closer to the origin than whole-scene embeddings, and the two distributions become clearly separated. This behavior results from the application of our uncertainty calibration and entropy regularizer. A more detailed analysis of hyperbolic space is provided in Sec. S.2.5 of the supplementary material.

4.4. Ablation study

To assess the contribution of each component in our framework, we performed ablation experiments, each removing a distinct component. In Tab. 4, 'w/o contrastive' removes the uncertainty-aware scaling from the global-local contrastive loss, while 'w/o uncertainty' disables the uncertainty calibration in uncertainty-guided entailment loss. Finally, 'w/o entropy' removes the entropy regularization from the uncertainty calibration module. The results demonstrate that all components of our method are essential. All experiments were conducted with ViT-S/16 architecture.

5. Conclusion

We propose UNCHA, a hyperbolic VLM that integrates part-to-whole representativeness, quantified as hyperbolic uncertainty, into both contrastive and entailment learning for hierarchy-aware compositional modeling. By further calibrating uncertainty using part-to-whole entailment relationships and an entropy based regularization term, our method enables efficient use of hyperbolic space and yields well-calibrated part-whole orderings. Extensive experiments on zero-shot classification, retrieval, and multi-label benchmarks, including complex multi-object scenes, demonstrate state-of-the-art performance, highlighting the importance of uncertainty-guided alignment for compositional understanding in vision-language learning.

Acknowledgements This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government(MSIT) [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) / No.RS-2025-02314125, Effective Human-Machine Teaming With Multimodal Hazy Oracle Models], the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (Nos. RS-2022-NR067592, RS-2025-02263628), the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (No. RQT-25-120066), the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University and AI-Bio Research Grant through Seoul National University.

References

[1] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *CVPR*, 2025. 1, 2, 7

[2] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, 2022. 2, 3, 4

[3] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *NeurIPS*, 32, 2019. 3

[4] Ines Chami, Albert Gu, Vaggos Chatziafratis, and Christopher Ré. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *NeurIPS*, 2020. 1

[5] Minshuo Chen, Yu Bai, Jason D Lee, Tuo Zhao, Huan Wang, Caiming Xiong, and Richard Socher. Towards understanding hierarchical learning: Benefits of neural representations. *NeurIPS*, 2020. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7

[7] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[8] Bhuwan Dhingra, Christopher Shallue, Mohammad Norouzi, Andrew Dai, and George Dahl. Embedding text in hyperbolic spaces. In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, 2018. 1, 3

[9] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 7

[10] Luca Franco, Paolo Mandica, Konstantinos Kallidromitis, Devin Guillory, Yu-Teng Li, Trevor Darrell, and Fabio Galasso. Hyperbolic active learning for semantic segmentation under domain shift. *ICML*, 2023. 2, 3, 4

[11] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pages 1646–1655. PMLR, 2018. 1, 3

[12] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *CVPR*, 2023. 2

[13] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 2004. 5

[14] Neil He, Jiahong Liu, Buze Zhang, Ngoc Bui, Ali Maatouk, Menglin Yang, Irwin King, Melanie Weber, and Rex Ying. Position: Beyond euclidean–foundation models should embrace non-euclidean geometries. *arXiv preprint arXiv:2504.08896*, 2025. 1

[15] Neil He, Hiren Madhu, Ngoc Bui, Menglin Yang, and Rex Ying. Hyperbolic deep learning for foundation models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025. 3

[16] Neil He, Menglin Yang, and Rex Ying. Hypercore: The core framework for building hyperbolic foundation models with comprehensive modules. *arXiv preprint arXiv:2504.08912*, 2025. 3

[17] Xiangteng He and Yuxin Peng. Fine-grained image classification via combining vision and language. In *CVPR*, 2017. 2

[18] Geoffrey Hinton. Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, 1979. 1

[19] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *Neural Computation*, 2023. 1

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[21] Sarah Ibrahimi, Mina Ghadimi Atigh, Nanne Van Noord, Pascal Mettes, and Marcel Worring. Intriguing properties of hyperbolic embeddings in vision-language models. *TMLR*, 2024. 2

[22] Takuya Ito, Tim Klinger, Doug Schultz, John Murray, Michael Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. *NeurIPS*, 2022. 1

[23] 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. In *ICML*, pages 4904–4916. PMLR, 2021. 1, 2

[24] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *CVPR*, 2023. 2

[25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 6

[26] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, 2020. 1, 3

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[28] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019. 3, 5

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6, 7

[31] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *NeurIPS*, 32, 2019. 3

[32] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 2

[33] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 5

[34] Paolo Mandica, Luca Franco, Konstantinos Kallidromitis, Suzanne Petryk, and Fabio Galasso. Hyperbolic learning with multimodal large language models. In *ECCV*, 2024. 2, 3, 4

[35] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 2017. 1, 2

[36] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *ICLR*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[37] 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. 2

[38] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 6, 7

[39] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *CVPR*, 2023. 2

[40] 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. In *ICML*, 2021. 1, 2, 6, 7, 8

[41] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *CVPR*, 2024. 1, 2, 3, 5, 6, 7, 8

[42] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Joint image-text representation by gaussian visual-semantic embedding. In *Proceedings of the 24th ACM international conference on Multimedia*, 2016. 2

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6, 8

[44] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, 2023. 2

[45] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, 2018. 1, 3

[46] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *CVPR*, 2025. 2

[47] Aditya Sinha, Siqi Zeng, Makoto Yamada, and Han Zhao. Learning structured representations with hyperbolic embeddings. *NeurIPS*, 2024. 3

[48] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 3

[49] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *CVPR*, 2024. 7, 8

[50] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 1, 3

[51] James Whittington, Timothy Muller, Shirely Mark, Caswell Barry, and Tim Behrens. Generalisation of structural knowledge in the hippocampal-entorhinal system. *NeurIPS*, 2018. 1

[52] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*, 2019. 2

[53] Shuchen Wu, Noémi Élteto, Ishita Dasgupta, and Eric Schulz. Learning structure from the ground up—hierarchical representation learning by chunking. *NeurIPS*, 2022. 1

[54] Shiyang Yan, Zongxuan Liu, and Lin Xu. Hyp-uml: Hyperbolic image retrieval with uncertainty-aware metric learning. *arXiv preprint arXiv:2310.08390*, 2023. 2, 3, 4

[55] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2014. 6