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

Vision-Language Models (VLMs) in Euclidean space struggle to capture hierarchical part-to-whole relationships, especially in multi-object compositional scenarios. Hyperbolic VLMs address this by modeling entailment relations, but existing approaches treat all parts equally without considering their varying semantic representativeness to the whole scene. We propose UNcertainty-guided Compositional Hyperbolic Alignment (UNCHA), which models part-to-whole semantic representativeness via hyperbolic uncertainty by assigning lower uncertainty to more representative parts and higher uncertainty to less representative ones. This representativeness is incorporated into the contrastive objective with uncertainty-guided weights, and the uncertainty is further calibrated with an entailment loss regularized by an entropy-based term. UNCHA achieves state-of-the-art performance on zero-shot classification, retrieval, and multi-label classification benchmarks.

1. Motivation

Human perception relies on part-whole hierarchies [6, 7], enabling efficient generalization through known relational structures [7, 8]. VLMs such as CLIP [13] have demonstrated remarkable image-text matching performance, but their Euclidean geometry struggles to capture hierarchical structures [5, 10] and exhibits bias in complex multi-object scenes [1].

Hyperbolic space, with its constant negative curvature and exponential volume growth, provides a natural geometric foundation for hierarchical embedding. MERU [3] extended contrastive vision-language learning into hyperbolic space with entailment relations. HyCoCLIP [11] further modeled intra-modal part-whole relationships. However, these approaches do not account for the fact that *each part has a different level of semantic representativeness to the whole*. When all parts are treated equally, the model fails to distinguish representative parts from less representative ones, leading to degraded multi-object alignment.

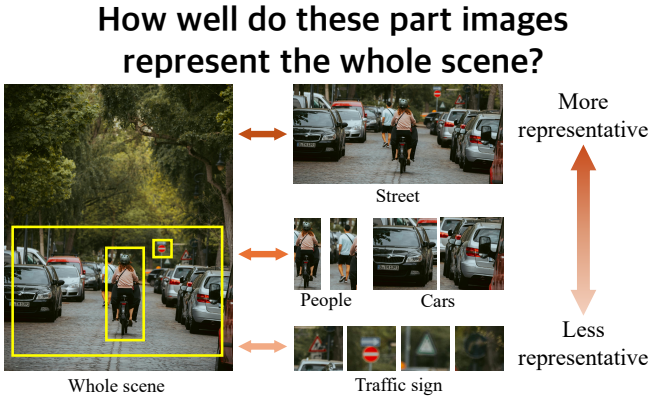

Figure 1. **Varying representativeness of part images to whole scene.** The relationship between each part image and the whole scene varies with its representativeness. We model this varying representativeness as uncertainty, enabling uncertainty-guided part–whole alignment in hyperbolic space.

We propose UNCHA, which models part-to-whole semantic representativeness as hyperbolic uncertainty, assigning lower uncertainty to more representative parts and higher uncertainty to less representative ones. This uncertainty guides contrastive learning and is calibrated by entailment loss with entropy-based regularization, leading to more accurate part–whole ordering and improved compositional understanding.

2. Method

2.1. Preliminaries

We adopt the Lorentz model of hyperbolic space with constant negative curvature $-\kappa$. A point $\mathbf{p} \in \mathbb{R}^{n+1}$ is ex-

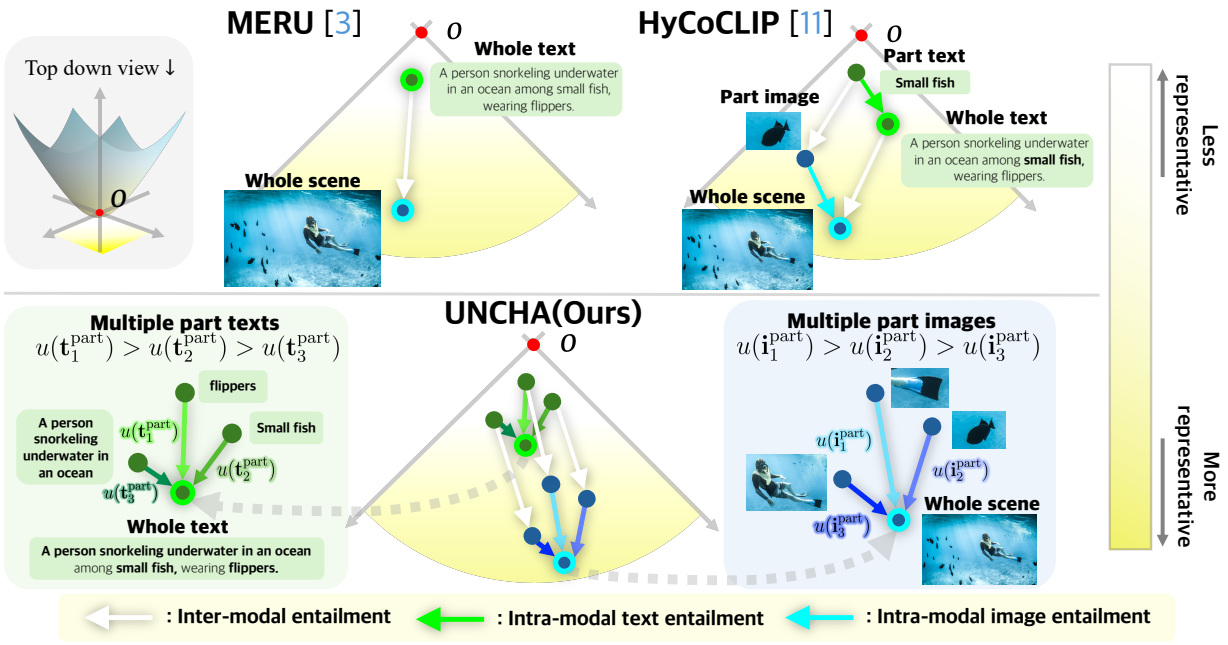

Figure 2. **Comparison of UNCHA (Ours) with prior works.** MERU [3] models inter-modal entailment between whole scene image and text. HyCoCLIP [11] extends this to include intra-modal part-whole entailment. UNCHA further incorporates *uncertainty to quantify the semantic representativeness* of each part, enabling uncertainty-guided part–whole alignment via adaptive weighting in the contrastive objectives and uncertainty calibration through the entailment loss with entropy regularization.

pressed as $[p_{\text{time}}, \mathbf{p}_{\text{space}}]$, and the Lorentzian inner product is $\langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}} = -p_{\text{time}} q_{\text{time}} + \langle \mathbf{p}_{\text{space}}, \mathbf{q}_{\text{space}} \rangle$. The geodesic distance is $d_{\mathbb{L}}(\mathbf{p}, \mathbf{q}) = \sqrt{1/\kappa}\, \cosh^{-1}(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathbb{L}})$. The hyperbolic radius of $\mathbf{p}$ is defined as $d_{\mathbb{L}}(\mathbf{p}, \mathbf{o})$, where $\mathbf{o}$ is the origin. Embeddings are parameterized in the tangent space at the origin and projected onto the manifold via the exponential map, consistent with prior works [3, 11, 14].

2.2. Uncertainty model of semantic representativeness

We leverage the hyperbolic radius to quantify part-to-whole semantic representativeness [2, 4, 9, 17]. Since abstract concepts lie near the origin and specific ones farther out, the radius naturally reflects representativeness. We define the uncertainty u for a point $\mathbf{x} \in \mathbb{L}^n$ as:

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

which is a smooth, differentiable, monotonically decreasing function of the hyperbolic radius: lower uncertainty for more representative parts (farther from origin), higher uncertainty for less representative ones.

2.3. Uncertainty-guided contrastive loss

We first define the basic contrastive loss using the negative Lorentzian distance as the similarity measure:

$$L_{\text{c}}^{*}(\mathbf{i}, \mathbf{t}; \tau) = -\sum_i \log \frac{\exp\left(-d_{\mathbb{L}}(\mathbf{i}_i, \mathbf{t}_i)/\tau\right)}{\sum_{k \neq i} \exp\left(-d_{\mathbb{L}}(\mathbf{i}_i, \mathbf{t}_k)/\tau\right)}. \tag{2}$$

Here, $\mathbf{i}$ and $\mathbf{t}$ denote the image and text embeddings, respectively, obtained from their corresponding encoders.

Building on this formulation, we introduce an uncertainty-aware temperature that modulates the contribution of each part. Specifically, we scale the global-local temperature in an element-wise manner based on the uncertainty of each part embedding:

$$\boldsymbol{\tau}_{\text{un},i}^{I} = \exp\big(u(\mathbf{i}_i^{\text{part}})/2\big)\ \tau_{gl}, \quad \boldsymbol{\tau}_{\text{un},i}^{T} = \exp\big(u(\mathbf{t}_i^{\text{part}})/2\big)\ \tau_{gl}, \tag{3}$$

where $\mathbf{i}^{\text{part}}$ and $\mathbf{t}^{\text{part}}$ denote part-level image and text embeddings that capture local regions and fine-grained semantic components. Higher uncertainty results in a larger temperature, thereby reducing the contribution of less reliable parts to the loss. Finally, our full uncertainty-guided contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{con}}^{\text{un}} = & \underbrace{L_{\text{c}}^{*}\big(\mathbf{i}^{\text{part}}, \mathbf{t}; \boldsymbol{\tau}_{\text{un}}^{I}\big) + L_{\text{c}}^{*}\big(\mathbf{t}^{\text{part}}, \mathbf{i}; \boldsymbol{\tau}_{\text{un}}^{T}\big)}_{\text{uncertainty-guided global-local}} \\ & + \underbrace{L_{\text{c}}^{*}(\mathbf{i}, \mathbf{t}; \tau_g) + L_{\text{c}}^{*}(\mathbf{t}, \mathbf{i}; \tau_g)}_{\text{global}} \\ & + \underbrace{L_{\text{c}}^{*}(\mathbf{i}^{\text{part}}, \mathbf{t}^{\text{part}}; \tau_l) + L_{\text{c}}^{*}(\mathbf{t}^{\text{part}}, \mathbf{i}^{\text{part}}; \tau_l)}_{\text{local}}. \end{aligned} \tag{4}$$

2.4. Entailment loss for uncertainty calibration

Piecewise-continuous entailment loss. The standard entailment loss [3, 11] enforces that $\mathbf{q}$ lies within the entail-

Table 1. **Zero-shot image classification and retrieval (R@1) evaluation.** UNCHA consistently demonstrates strong performance across both architectures. Bold numbers denote the best within each architecture. †: ATMG trained on GRIT [12].

		Classification												Retrieval			
		General datasets							Fine-grained datasets					Text		Image	
	Model	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	COCO	Flickr	COCO	Flickr
ViT-B/16	CLIP [13]	40.6	78.9	48.3	43.0	70.7	92.4	48.3	10.4	9.3	3.4	45.9	21.3	71.4	**93.6**	57.4	83.5
	MERU [3]	40.1	78.6	49.3	43.0	73.0	92.8	48.5	11.0	5.3	3.7	48.5	21.6	72.3	93.5	57.4	84.0
	ATMG† [14]	34.3	68.8	42.1	48.2	68.5	91.2	43.2	14.3	6.0	2.4	42.2	15.0	62.9	85.1	51.2	78.0
	HyCoCLIP [11]	45.8	88.8	60.1	57.2	81.3	95.0	59.2	**16.4**	11.6	3.7	56.8	23.9	72.0	92.6	58.4	**84.9**
	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**	**72.7**	91.4	**60.0**	**84.9**

Table 2. **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.

		Multi-object Representation								Multi-label Cls.	
		ComCo				SimCo				VOC	COCO
	Model	2 obj.	3 obj.	4 obj.	5 obj.	2 obj.	3 obj.	4 obj.	5 obj.		
ViT-B/16	CLIP [13]	77.55	80.31	81.41	80.22	77.15	84.58	87.40	88.48	78.56	53.94
	MERU [3]	72.90	77.25	78.15	77.34	77.82	83.91	85.79	86.90	79.50	54.39
	ATMG† [14]	45.91	45.97	45.80	45.82	65.52	65.32	65.28	65.12	72.22	46.81
	HyCoCLIP [11]	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**

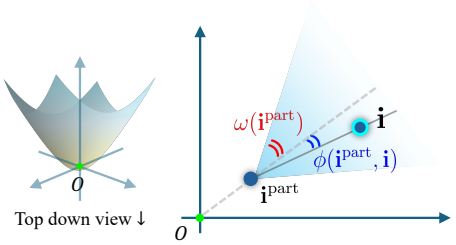

Figure 3. **Entailment geometry in hyperbolic space.** The term $\omega(\mathbf{i}^{\text{part}})$ denotes the aperture of the entailment cone centered at $\mathbf{i}^{\text{part}}$. The angle $\phi(\mathbf{i}^{\text{part}}, \mathbf{i})$ measures the geodesic angle between the embeddings $\mathbf{i}^{\text{part}}$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

ment cone of $\mathbf{p}$, but produces zero gradient once $\mathbf{q}$ is inside. To alleviate this issue, we introduce an angular term, $\phi(\mathbf{p}, \mathbf{q})$, which acts as a leaky-ReLU [16]-like relaxation by maintaining a non-zero gradient even within the cone, thereby enabling continued fine-grained alignment:

$$L^*_{\text{ent}}(\mathbf{p}, \mathbf{q}) = \max(0,\ \phi(\mathbf{p}, \mathbf{q}) - \eta\,\omega(\mathbf{p})) + \alpha\,\phi(\mathbf{p}, \mathbf{q}) \quad (5)$$

where ϕ and $\omega(\mathbf{p})$ denotes the angular distance and cone aperture, respectively.

Uncertainty calibration loss. Prior studies have reported that hyperbolic embeddings often accumulate in narrow regions, leading to representation collapse [14]. Moreover, local and global representations tend to exhibit similar radii, making their separation less distinct [11]. To explicitly enforce a meaningful separation between global and local representations while preventing collapse, we propose the following uncertainty calibration loss:

$$L^{\text{cal}}_{\text{ent}}(\mathbf{p}, \mathbf{q}) = \lfloor L^*_{\text{ent}}(\mathbf{p}, \mathbf{q}) \rfloor\, e^{-u(\mathbf{p})} + u(\mathbf{p}) + \mathcal{H}(\tilde{u}(\mathbf{p})) \quad (6)$$

where $\lfloor \cdot \rfloor$ denotes the stop-gradient operator, and $\mathcal{H}$ is the entropy term defined as:

$$\mathcal{H}(\tilde{u}(\mathbf{p})) = -\sum\nolimits_i \tilde{u}(\mathbf{p}_i) \log(\tilde{u}(\mathbf{p}_i)) \quad (7)$$

with $\tilde{u}(\mathbf{p}_i) = \exp(u(\mathbf{p}_i))/\sum_j \exp(u(\mathbf{p}_j))$. The formulation calibrates uncertainty by increasing it for weak part–whole relations, while preventing it from becoming uniformly large, and encouraging a diverse distribution across parts. With the entropy regularizer, the proposed formulation of our entailment loss is as follows:

$$\begin{aligned} \mathcal{L}^{\text{un}}_{\text{ent}} \;=\; & \underbrace{L^*_{\text{ent}}(\mathbf{t}^{\text{part}}, \mathbf{i}^{\text{part}}) + L^*_{\text{ent}}(\mathbf{t}, \mathbf{i})}_{\text{inter-modal entailment}} \\ & + \lambda_1 \underbrace{(L^*_{\text{ent}}(\mathbf{t}^{\text{part}}, \mathbf{t}) + L^*_{\text{ent}}(\mathbf{i}^{\text{part}}, \mathbf{i}))}_{\text{intra-modal entailment}} \\ & + \lambda_2 \underbrace{(L^{cal}_{\text{ent}}(\mathbf{t}^{\text{part}}, \mathbf{t}) + L^{cal}_{\text{ent}}(\mathbf{i}^{\text{part}}, \mathbf{i}))}_{\text{uncertainty calibration}} \end{aligned} \quad (8)$$

where λ_1 and λ_2 are hyperparameters. Together, these components allow uncertainty to reflect the semantic representativeness of each part while maintaining a well-structured and stable hyperbolic embedding space. The overall loss is $L = \mathcal{L}^{\text{un}}_{\text{con}} + \lambda_{\text{ent}}\mathcal{L}^{\text{un}}_{\text{ent}}$.

3. Results

3.1. Setup

All models are trained on the GRIT [12] dataset (20.5M grounded pairs, 35.9M part-level annotations) with batch size 768 for 500K iterations. Baselines [3, 11, 13, 14] are reproduced under identical configurations.

3.2. Zero-shot image classification and zero-shot retrieval

Tab. 1 reports Top-1 accuracy on 16 benchmarks with ViT-B/16. UNCHA consistently outperforms prior hyperbolic VLMs across general, fine-grained, and miscellaneous datasets. On ViT-B/16, UNCHA achieves 48.8% on ImageNet, 90.4% on CIFAR-10, and 83.9% on Caltech-101, demonstrating strong generalization. On COCO and Flickr30K retrieval, UNCHA achieves the best image retrieval performance and competitive text retrieval, as shown in Tab. 1. These results suggest that our uncertainty-guided alignment improves cross-modal matching performance.

3.3. Multi-object and compositional benchmarks

UNCHA demonstrates substantial gains on multi-object tasks. On ComCo and SimCo benchmarks with ViT-B/16, UNCHA outperforms all baselines across all object configurations (2–5 objects). On multi-label classification, UNCHA achieves 82.14 mAP on VOC and 59.43 on COCO, validating that uncertainty-aware modeling provides stronger compositional understanding.

3.4. Analysis about hyperbolic space

We visualize the radii of hyperbolic embedding for $10,000$ ImageNet [15] images and their randomly cropped parts, shown in Fig. 4. As noted in HyCoCLIP [11], 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.

3.5. Ablation study

To assess the contribution of each component in our framework, we performed ablation experiments, each removing a distinct component. In Tab. 3, '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.

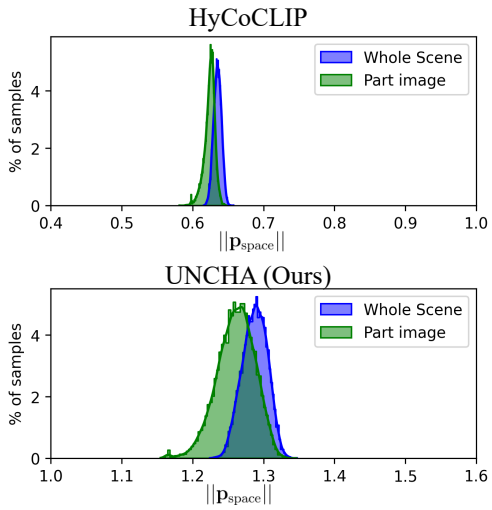

Figure 4. **Analysis of hyperbolic embedding.** Compared to HyCoCLIP [11], whose hyperbolic embeddings exhibit a narrower range, UNCHA yields a more dispersed and structured distribution, reflecting richer use of the hyperbolic space.

Table 3. **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.

	Classification			Retrieval	
Model	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

4. Conclusion

We proposed UNCHA, a hyperbolic VLM that models part-to-whole semantic representativeness as hyperbolic uncertainty, integrating it into both contrastive and entailment learning with entropy-based regularization. UNCHA achieves state-of-the-art on zero-shot classification, retrieval, and multi-label benchmarks, demonstrating the importance of uncertainty-guided alignment for compositional understanding.

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

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

[4] 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

[5] 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

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

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

[8] 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

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

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

[11] 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

[12] 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. 3, 4

[13] 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, 3, 4

[14] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *CVPR*, 2024. 2, 3, 4

[15] 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. 4

[16] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 3

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