PHyCLIP: ℓ_1-Product of Hyperbolic Factors Unifies Hierarchy and Compositionality in Vision-Language Representation Learning

Daiki Yoshikawa[1], Takashi Matsubara[1,2] ([1]Hokkaido University, [2]CyberAgent)

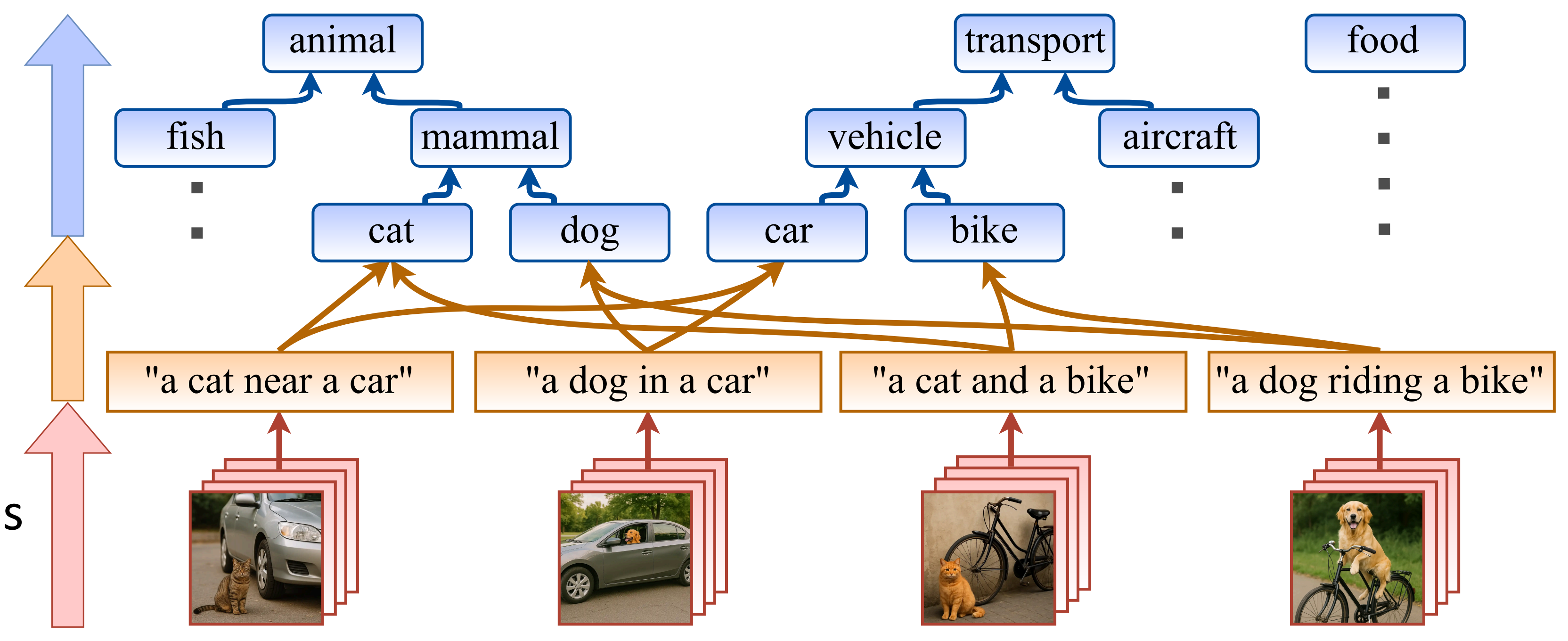
HOKKAIDO UNIVERSITY · CyberAgent AI Lab

TL;DR:

Images and texts have two aspects:

tree-like taxonomic hierarchy
by a hyperbolic space
&
Boolean-like compositionality
by an ℓ_1-product metric
⇓
a product of metric trees
by a **P**roduct of **Hy**perbolic spaces

Let's enjoy the best of both worlds!



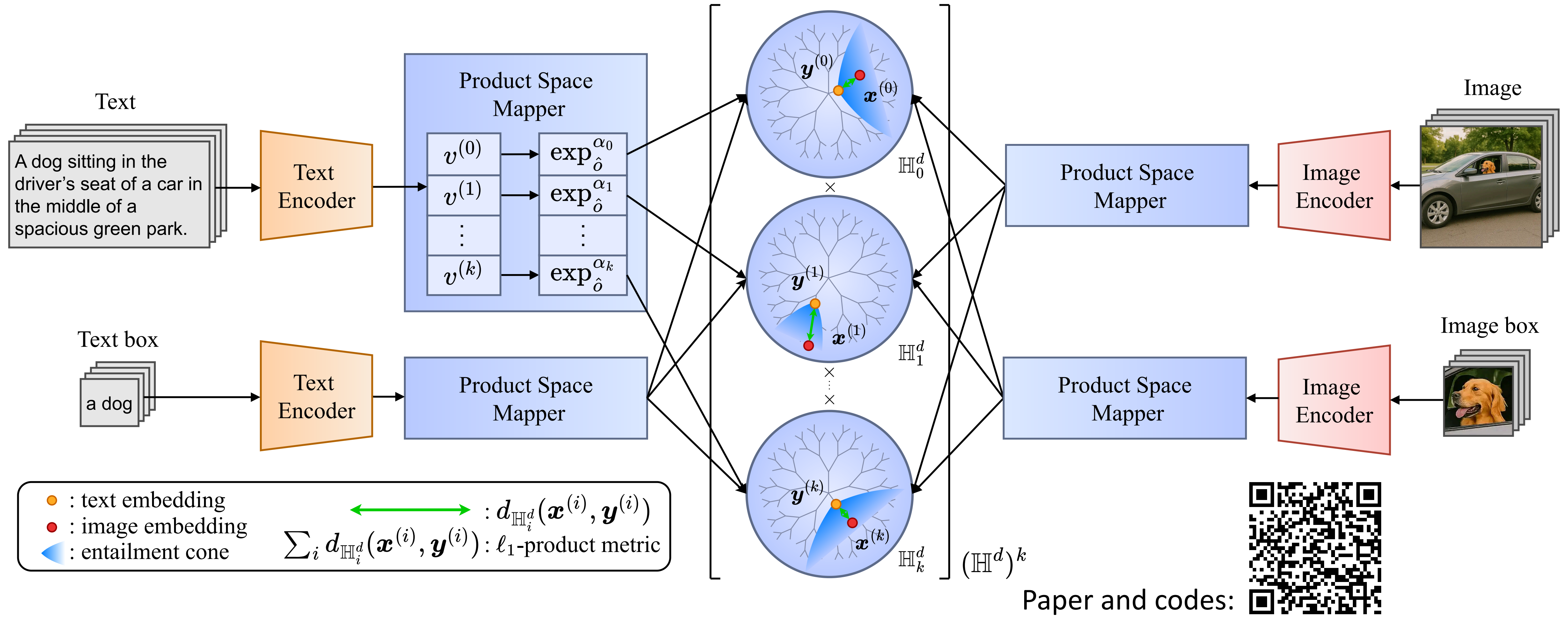
Methods:

Theorem 1 (Sarkar, 2011):
A metric tree T is quasi-isometrically embedded into a 2D hyperbolic space $\mathbb{H}^2$.

Proposition 1:
A Boolean algebra with indicators and the Hamming distance, $(\{0,1\}^k, d_{\mathrm{Ham}})$ is isometrically embedded into an ℓ_1-product metric space $(\mathbb{R}^k, d_1)$, but not into a hyperbolic space $(\mathbb{H}^k, d_{\mathbb{H}^k}$.

PHyCLIP:
A CLIP-type vision-language representation learning that embeds instances into $((\mathbb{H}^d)^k, d_1)$,
a Cartesian product of k-copies of hyperbolic spaces $\mathbb{H}^d$ equipped with an ℓ_1-product metric d_1.

Theorem 2:
PHyCLIP quasi-isometrically embeds a product of metric trees,
capturing intra-family taxonomic hierarchies by hyperbolic factors
and cross-family Boolean-like compositionality by an ℓ_1-product metric.



Paper and codes:
https://github.com/tksmatsubara/PHyCLIP

Experiments and Results:

PHyCLIP trained on GRIT (Peng et al., 2023) with the contrastive & entailment losses is better at hierarchical classifications and object compositions, but worse at object relations.

		General datasets						Fine-grained datasets						Specialized datasets			
	w/ boxes	ImageNet	CIFAR-10	CIFAR-100	SUN397	Caltech-101	STL-10	Food-101	CUB	Cars	Aircraft	Pets	Flowers	DTD	EuroSAT	RESISC45	Country211
CLIP		38.87	76.26	48.19	50.70	73.62	93.03	51.19	12.90	7.82	3.01	45.89	<u>21.16</u>	22.02	35.73	42.03	5.13
CLIP	✓	38.81	76.53	48.59	50.80	74.29	93.34	51.05	12.70	8.40	2.89	46.19	**21.32**	21.74	<u>37.49</u>	41.78	5.10
MERU		37.96	77.63	46.37	49.39	72.10	93.14	51.67	11.09	7.80	<u>3.53</u>	43.36	19.98	22.18	**38.81**	41.77	4.86
MERU	✓	38.08	78.14	46.80	49.59	72.69	93.28	51.92	10.70	7.77	<u>3.53</u>	43.22	18.31	22.07	37.31	41.73	5.01
HyCoCLIP	✓	<u>43.80</u>	<u>89.00</u>	<u>58.59</u>	<u>54.49</u>	<u>76.14</u>	**94.96**	<u>52.64</u>	<u>14.90</u>	<u>10.24</u>	**3.57**	<u>53.33</u>	19.41	**25.90**	36.36	<u>46.97</u>	**5.64**
PHyCLIP	✓	**44.31**	**89.33**	**59.05**	**55.32**	**76.35**	<u>94.84</u>	**57.26**	**15.90**	**10.89**	3.24	**54.18**	19.98	<u>25.50</u>	36.29	**48.22**	<u>5.56</u>

		Text → Image				Image → Text				Hierarchical Classification				
		COCO		Flickr		COCO		Flickr		WordNet				
	w/ boxes	R@5	R@10	R@5	R@10	R@5	R@10	R@5	R@10	TIE(↓)	LCA(↓)	J(↑)	P_H(↑)	R_H(↑)
CLIP		56.29	67.53	<u>83.15</u>	89.58	70.32	80.09	**91.60**	95.60	3.750	2.276	0.7774	0.8471	0.8483
CLIP	✓	56.20	67.50	82.75	89.42	<u>70.35</u>	<u>80.19</u>	91.10	95.63	3.736	2.279	0.7784	0.8473	0.8501
MERU		55.73	67.02	82.15	89.05	69.57	79.33	90.77	**95.83**	3.815	2.294	0.7733	0.8454	0.8450
MERU	✓	55.87	67.21	81.96	88.89	69.70	79.69	91.20	**95.83**	3.802	2.289	0.7740	0.8457	0.8455
HyCoCLIP	✓	<u>57.11</u>	<u>68.32</u>	83.06	<u>89.63</u>	69.51	<u>79.73</u>	<u>91.47</u>	95.63	<u>3.319</u>	<u>2.092</u>	<u>0.8043</u>	<u>0.8676</u>	<u>0.8661</u>
PHyCLIP	✓	**58.03**	**69.05**	**83.39**	**89.93**	**70.94**	**80.86**	91.20	95.53	**3.294**	**2.083**	**0.8059**	**0.8684**	**0.8672**

		VL-CheckList–Object						SugarCrepe							
		Location			Size			Replace			Swap		Add		
	w/ boxes	Center	Mid	Margin	Large	Medium	Small	Obj	Att	Rel	Obj	Att	Obj	Att	Overall
CLIP		61.9	60.3	60.4	63.9	60.2	58.2	89.37	79.95	<u>69.54</u>	60.54	66.02	80.39	73.36	77.72
CLIP	✓	61.9	59.3	60.8	63.7	60.8	58.1	89.69	80.33	69.49	**61.63**	**66.47**	80.62	73.55	77.97
MERU		61.3	59.0	59.0	64.0	57.7	56.1	89.10	<u>80.50</u>	69.44	<u>60.82</u>	65.32	80.47	<u>74.90</u>	77.81
MERU	✓	61.0	58.5	58.7	62.6	58.7	56.5	89.39	79.95	**69.65**	60.41	<u>66.07</u>	80.41	**75.34**	77.93
HyCoCLIP	✓	<u>70.4</u>	<u>69.5</u>	<u>67.8</u>	<u>72.6</u>	<u>66.1</u>	<u>67.2</u>	**91.38**	79.74	67.24	54.69	63.66	<u>82.57</u>	74.23	<u>77.99</u>
PHyCLIP	✓	**71.2**	**70.3**	**70.4**	**73.7**	**68.1**	**67.8**	<u>91.06</u>	**81.05**	66.36	57.41	65.87	**83.24**	73.80	**78.32**

A more factorization leads to a better result, but a mixed curvature does not work.

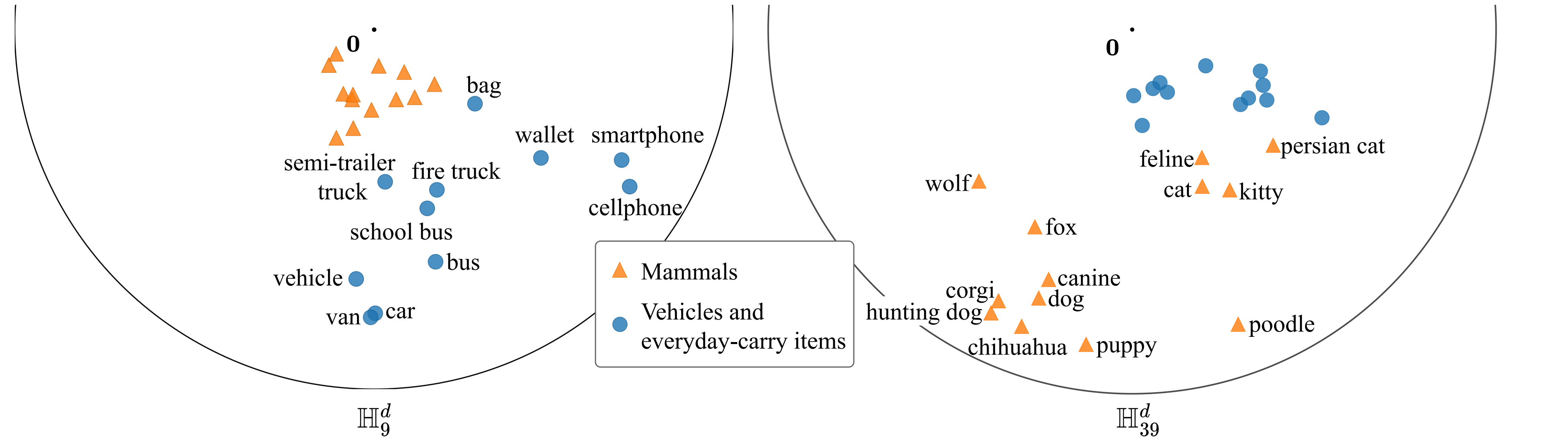
				classification		retrieval COCO, R@5		hierarchical	
# of factors, k	# of dims., d	product metric	curvature	ImageNet	Food-101	Image	Text	TIE	J
1	512	–	hyp.	43.80	52.64	57.11	69.51	3.319	0.8043
8	64	ℓ_1	hyp.	<u>44.38</u>	54.61	57.80	70.80	3.273	0.8072
16	32	ℓ_1	hyp.	44.09	<u>55.29</u>	57.26	69.22	<u>3.287</u>	**0.8066**
32	16	ℓ_1	hyp.	43.90	54.48	56.70	66.92	3.324	0.8035
64	8	ℓ_1	hyp.	**44.31**	**57.26**	**58.03**	<u>70.94</u>	3.294	0.8059
128	4	ℓ_1	hyp.	44.16	53.96	<u>57.79</u>	**71.18**	**3.284**	<u>0.8064</u>
64	8	ℓ_2	hyp.	43.32	53.39	57.09	70.53	3.367	0.8011
64	8	ℓ_∞	hyp.	6.55	10.33	8.77	14.51	9.697	0.4247
-	-	ℓ_2	mixed	39.34	49.05	56.72	70.81	3.712	0.7797

References:

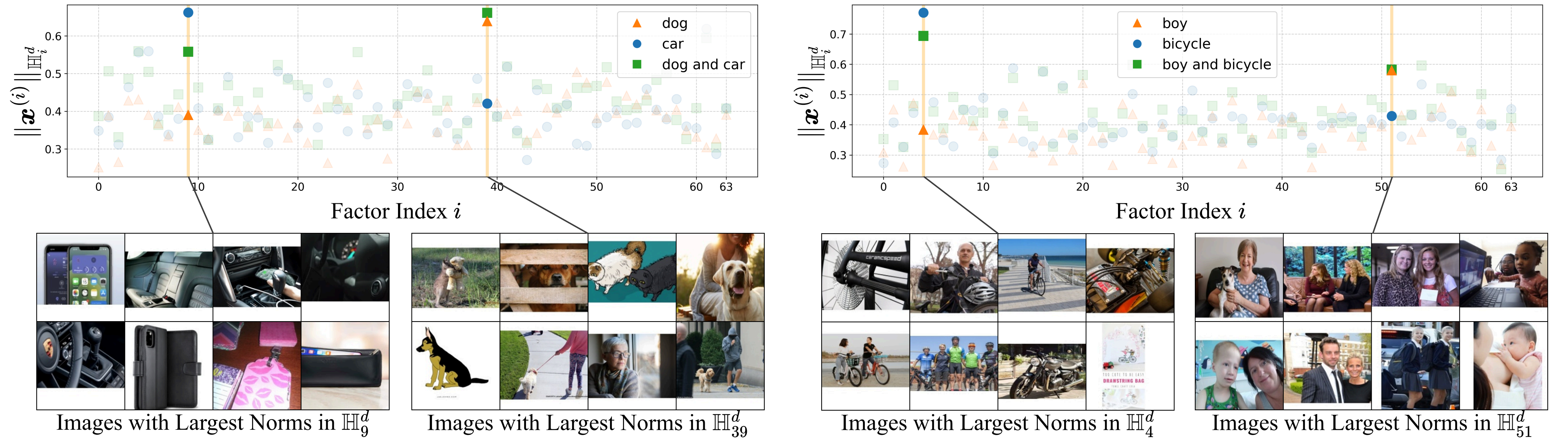
M. R. Bridson & A. Haefliger, *Metric Spaces of Non-Positive Curvature*, Springer, 1999
B. Ganter & R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, 1999.
R. Sarkar, "Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane," *GD*, 2011.
I. Vendrov et al., "Order-Embeddings of Images and Language," *ICLR*, 2016.
M. Nickel & D. Kiela, "Poincaré Embeddings for Learning Hierarchical Representations," *NeurIPS*, 2017
Z. Peng et al., "Kosmos-2: Grounding Multimodal Large Language Models to the World," *arXiv*, 2023.
K. Desai et al., "Hyperbolic Image-text Representations," *ICML*, 2023
A. Pal et al., "Compositional Entailment Learning for Hyperbolic Vision-Language Models," *ICLR*, 2025.

Visualizations:

A taxonomic tree of a concept family emerges selectively in a hyperbolic factor,
even without explicit supervision for factor assignments or hierarchy between atomic concepts



Boolean-like behavior of factor-wise embeddings
-A conjunctive prompt activates all factors that single-concept prompts activate.



-Factor-wise *max* of single-concept prompts behaves like their conjunctive prompt in image retrieval.

"a dog and a car" | Factor-wise *max* of "a dog" and "a car" | "a boy and a bicycle" | Factor-wise *max* of "a boy" and "a bicycle"

Related Work:

	Generalization (hypernymy)	Specialization (hyponymy)	Space	Entailment ($\boldsymbol{x}$ or S entails $\boldsymbol{y}$ or T)
Tree of *is-a* Relations (*is-a* Taxonomy)	join ⊔	(meet ⊓)	T	$\boldsymbol{x} \preceq \boldsymbol{y}$
Order Embedding (as points)	$\min$	$\max$	$\mathbb{R}^n$	$x_i \geq y_i$ for all i
Order Embedding (as orthants)			orthants in $\mathbb{R}^n$	$U(\boldsymbol{x}) \subseteq U(\boldsymbol{y})$
Order Embedding (for entailment)			orthants in $\mathbb{R}^n$	$\boldsymbol{x} \in U(\boldsymbol{y})$
Hyperbolic Entailment Cone	(union ∪)	intersection ∩	cones in $\mathbb{H}^n$	$\boldsymbol{x} \in C(\boldsymbol{y})$
Boolean Lattice (as a power set)	intersection ∩	union ∪	$2^{\mathcal{C}}$	$S \supseteq T$
Boolean Lattice (as a lattice)	meet ⊓	join ⊔		$S \succeq T$
Boolean Lattice (with indicator)	AND	OR	$\{0,1\}^{\lvert\mathcal{C}\rvert}$	$\chi(S)_i \geq \chi(T)_i$ for all i
Dual Lattice (as a set)	union ∪	intersection ∩		$S' \subseteq T'$
Dual Lattice (as a lattice)	join ⊔	meet ⊓		$S' \preceq T'$
Product of Trees	join ⊔	(meet ⊓)	$\prod_{i=1}^k T_i$	$\boldsymbol{x}^{(i)} \preceq \boldsymbol{y}^{(i)}$ for all i
PHyCLIP	(union ∪)	intersection ∩	cones in $(\mathbb{H}^d_i)^k$	$\boldsymbol{x}^{(i)} \in C_i(\boldsymbol{y}^{(i)})$ for all i