Compositional Entailment Learning for Hyperbolic Vision-Language Models

Avik Pal[1*] **Max van Spengler**[1] **Guido Maria D'Amely di Melendugno**[2]
Alessandro Flaborea[2] **Fabio Galasso**[2] **Pascal Mettes**[1]

[1]University of Amsterdam [2]Sapienza University of Rome

2025.1.13 Minghao Liu

目录

CONTENTS

北京大學
PEKING UNIVERSITY

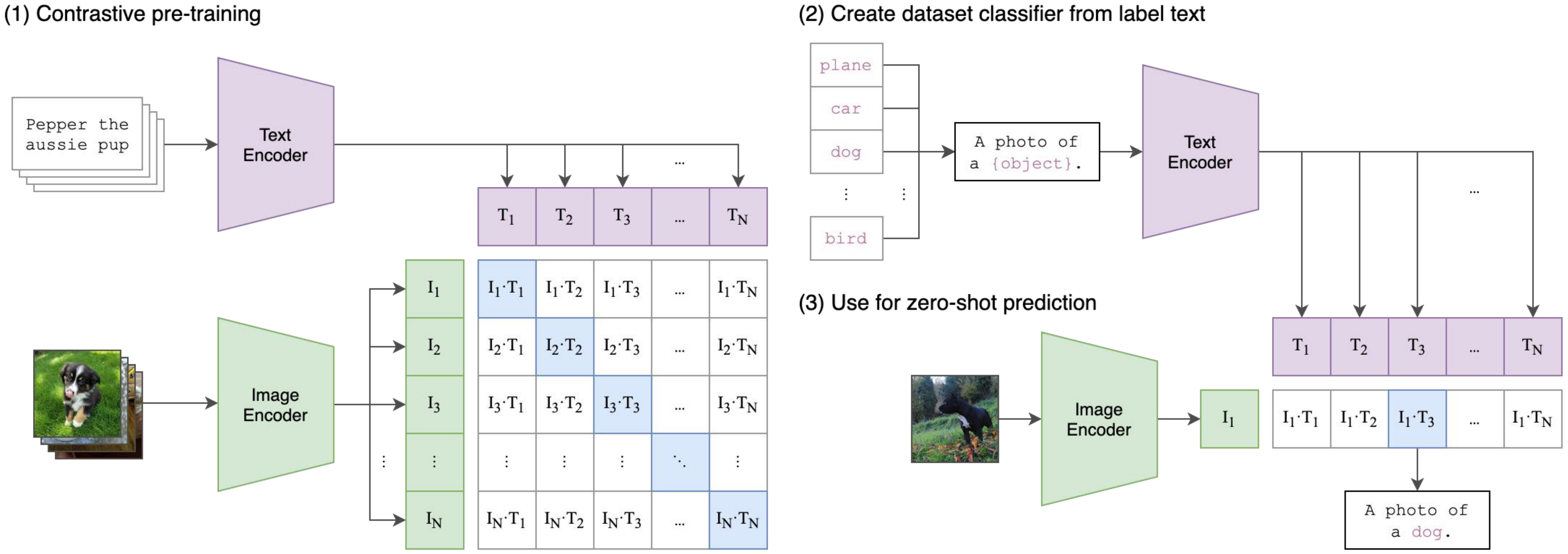
PART 02

Background

Background

PEKING UNIVERSITY

CLIP (Contrastive Language-Image Pre-Training)



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

Background

MERU (Hyperbolic Image-Text Representations)

Desai, Karan, et al. "Hyperbolic image-text representations." International Conference on Machine Learning. PMLR, 2023.

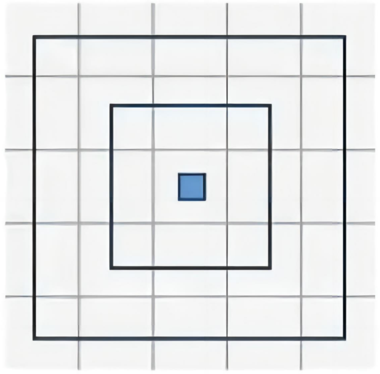
Background

Why Hyperbolic:

Better capture hierarchy informations

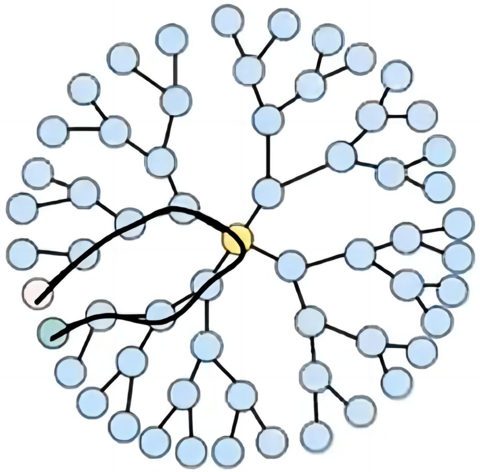
Geometric properties to embed tree-like data

Better representation space



Grid-like

Polynomial level



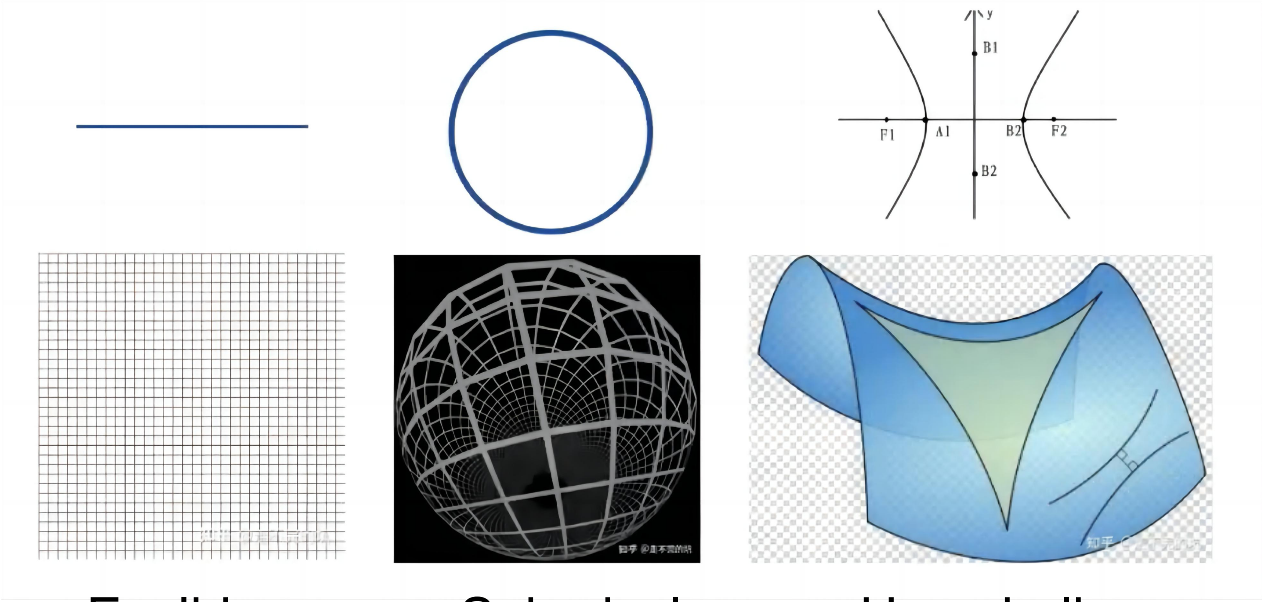
Tree-like

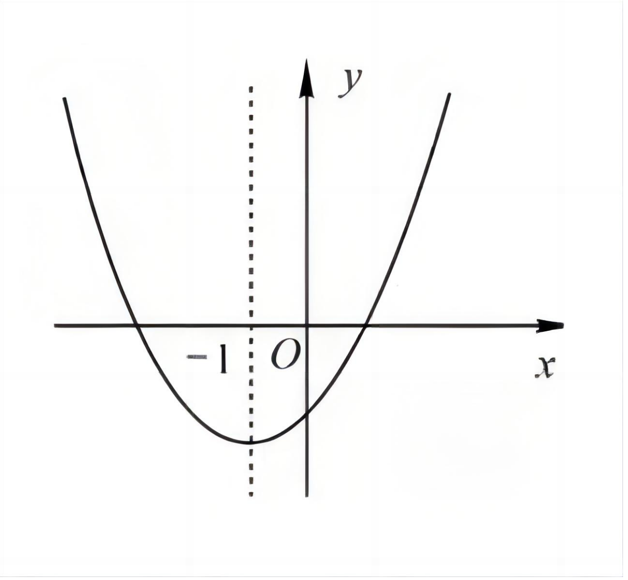
Exponential level

Background

What is Hyperbolic:

Negative constant curvature space

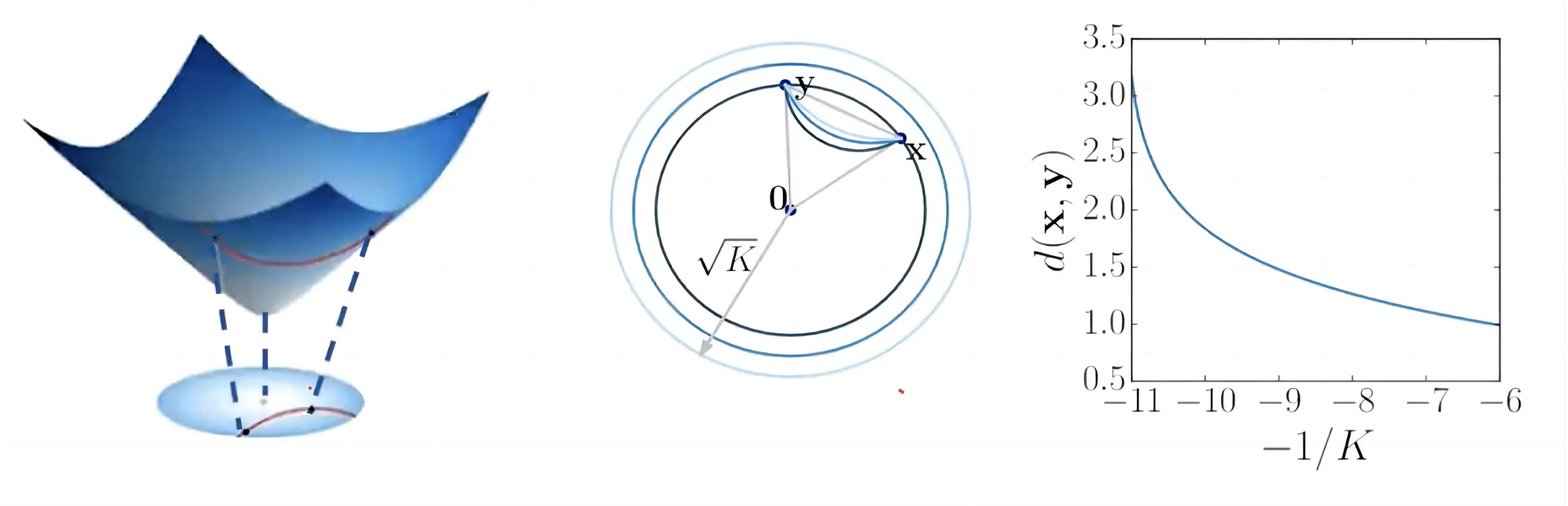




Euclidean $k = 0$	Spherical $k > 0$	Hyperbolic $k < 0$	k is not constant

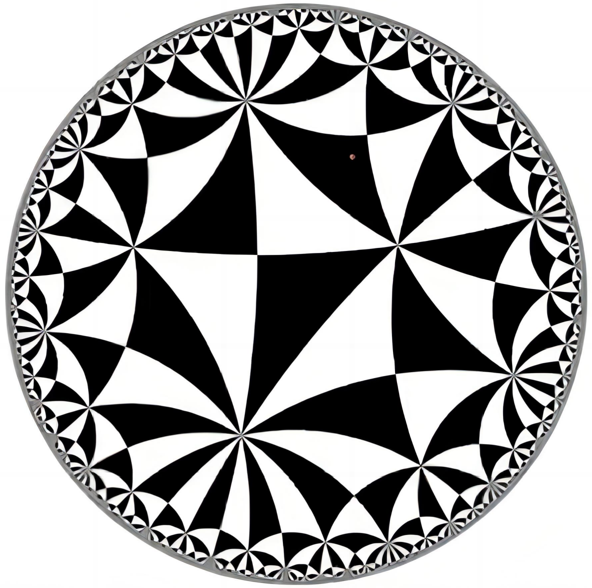
Background

How to Represent Hyperbolic:
e.g. Poincaré disk model



Background

How to Represent Hyperbolic:
e.g. Poincaré disk model



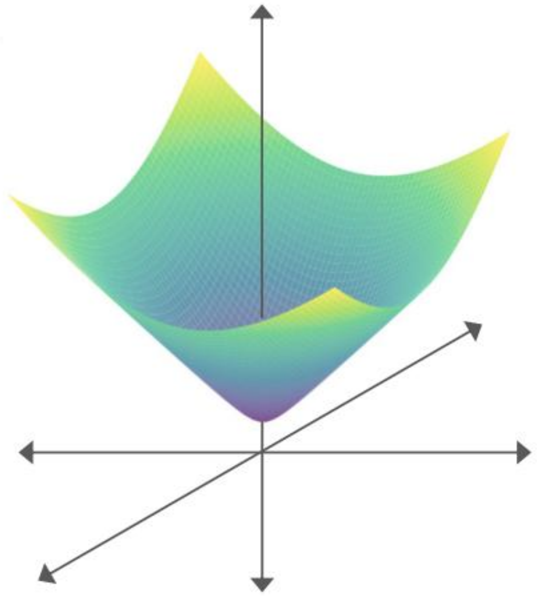
Two-dimensional
Poincare disk model

$$d^1_{\mathbb{D}}(\mathbf{x}, \mathbf{y}) = \operatorname{arcosh}\left(1 + 2\frac{||\mathbf{x}-\mathbf{y}||_2^2}{(1-||\mathbf{x}||_2^2)(1-||\mathbf{y}||_2^2)}\right).$$

Background

MERU uses Lorentz model

use R^{n+1} to represent L^n

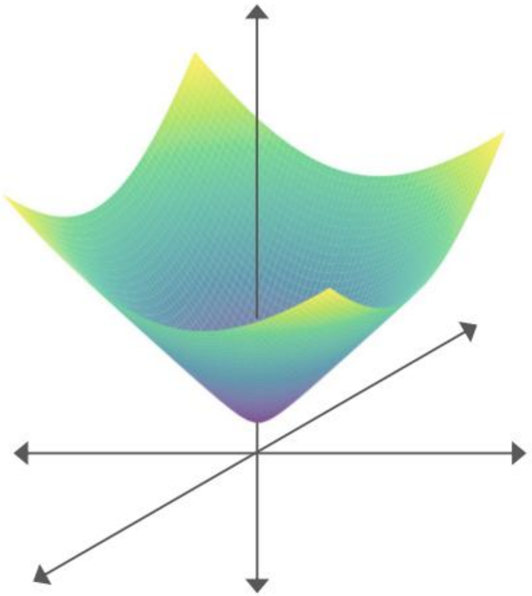


Background

PEKING UNIVERSITY

MERU uses Lorentz model use R^{n+1} to represent L^n

Every vector $\mathbf{x} \in \mathbb{R}^{n+1}$ can be written as $[\mathbf{x}_{space}, x_{time}]$, where $\mathbf{x}_{space} \in \mathbb{R}^n$ and $x_{time} \in \mathbb{R}$.

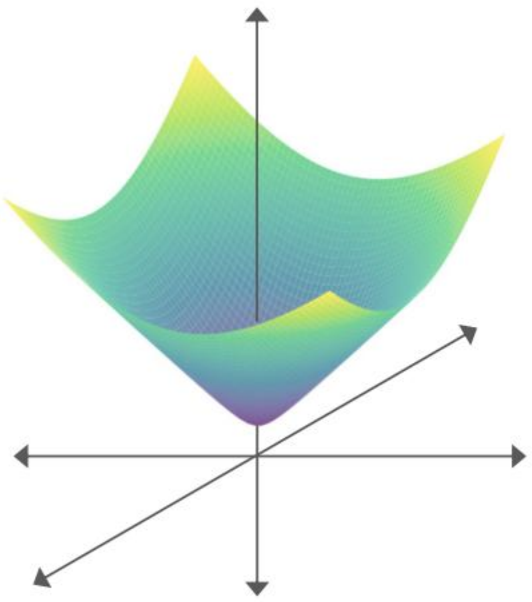


Inner Product: $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = \langle \mathbf{x}_{space}, \mathbf{y}_{space} \rangle - x_{time}\, y_{time}$

Background

MERU uses Lorentz model use R^{n+1} to represent L^n

Every vector $\mathbf{x} \in \mathbb{R}^{n+1}$ can be written as $[\mathbf{x}_{space}, x_{time}]$, where $\mathbf{x}_{space} \in \mathbb{R}^n$ and $x_{time} \in \mathbb{R}$.



Inner Product: $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}} = \langle \mathbf{x}_{space}, \mathbf{y}_{space} \rangle - x_{time}\, y_{time}$

Norm: $\|\mathbf{x}\|_{\mathcal{L}} = \sqrt{|\langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}}|}$

Lorentz model with constantcurvature −c:

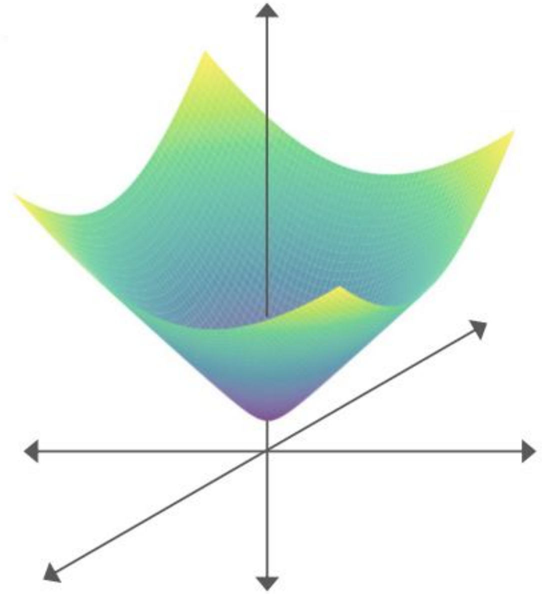
$$\mathcal{L}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} : \langle \mathbf{x}, \mathbf{x} \rangle_{\mathcal{L}} = -{}^{1}/{c}\ ,\ c > 0\}$$

Satisfy:

$$x_{time} = \sqrt{{}^{1}/{c} + \|\mathbf{x}_{space}\|^2}$$

Background

MERU uses Lorentz model use R^{n+1} to represent L^n



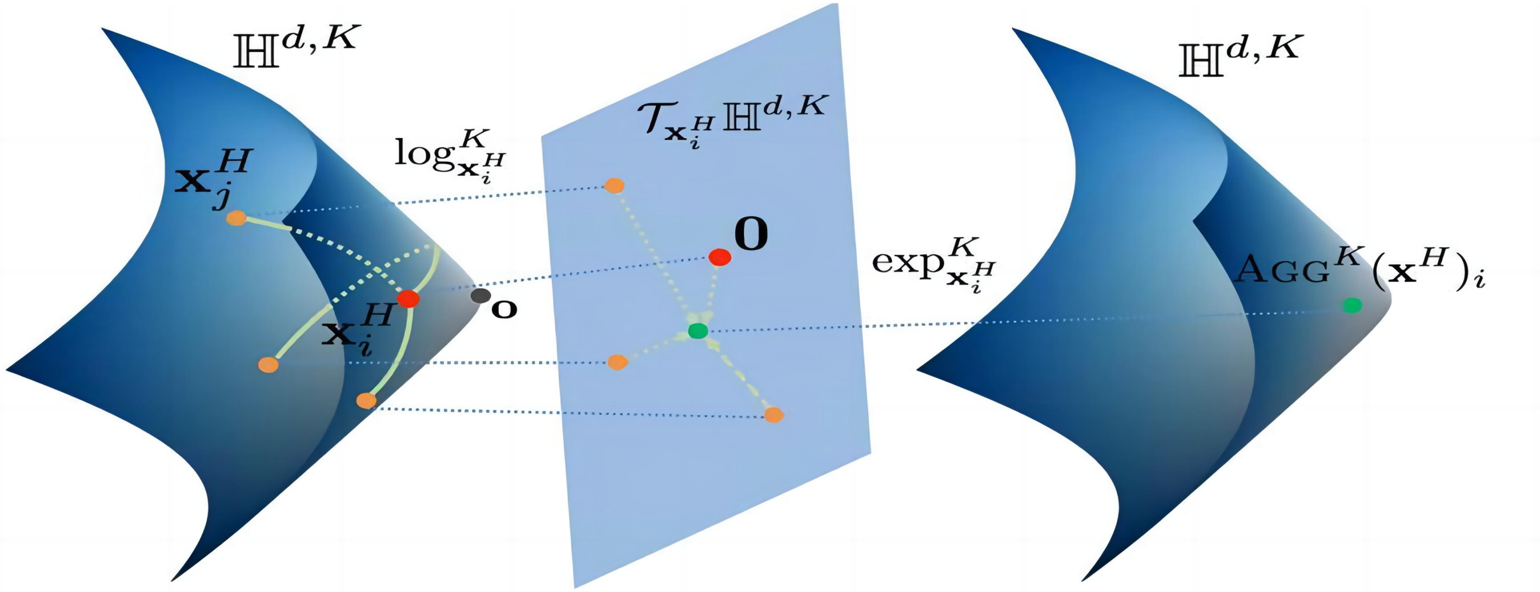
Geodesics: $d_{\mathcal{L}}(\mathbf{x}, \mathbf{y}) = \sqrt{1/c} \cdot \cosh^{-1}(-c \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{L}})$

Tangent space for z: $\mathcal{T}_{\mathbf{z}}\mathcal{L}^n = \{\mathbf{v} \in \mathbb{R}^{n+1} : \langle \mathbf{z}, \mathbf{v} \rangle_{\mathcal{L}} = 0\}$

Project to Tangent space:

$$\mathbf{v} = \mathrm{proj}_{\mathbf{z}}(\mathbf{u}) = \mathbf{u} + c\, \mathbf{z}\, \langle \mathbf{z}, \mathbf{u} \rangle_{\mathcal{L}}$$

Background

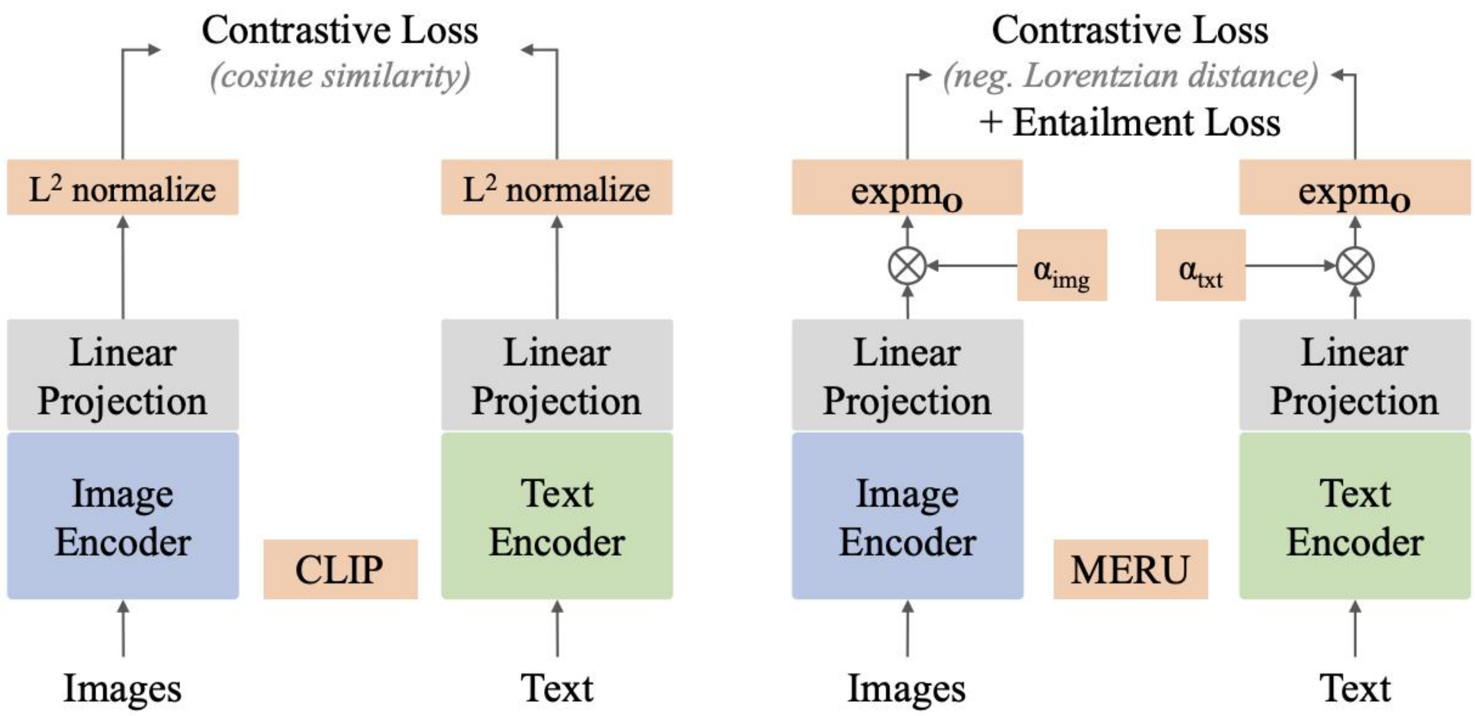


Exponential map: $$\mathbf{x} = \mathrm{expm}_{\mathbf{z}}(\mathbf{v}) = \cosh(\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}})\,\mathbf{z} + \frac{\sinh(\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}})}{\sqrt{c}\,\|\mathbf{v}\|_{\mathcal{L}}}\,\mathbf{v}$$

Logarithmic map: $$\mathbf{v} = \mathrm{logm}_{\mathbf{z}}(\mathbf{x}) = \frac{\cosh^{-1}(-c\,\langle \mathbf{z}, \mathbf{x}\rangle_{\mathcal{L}})}{\sqrt{(c\,\langle \mathbf{z}, \mathbf{x}\rangle_{\mathcal{L}})^2 - 1}}\,\mathrm{proj}_{\mathbf{z}}(\mathbf{x})$$

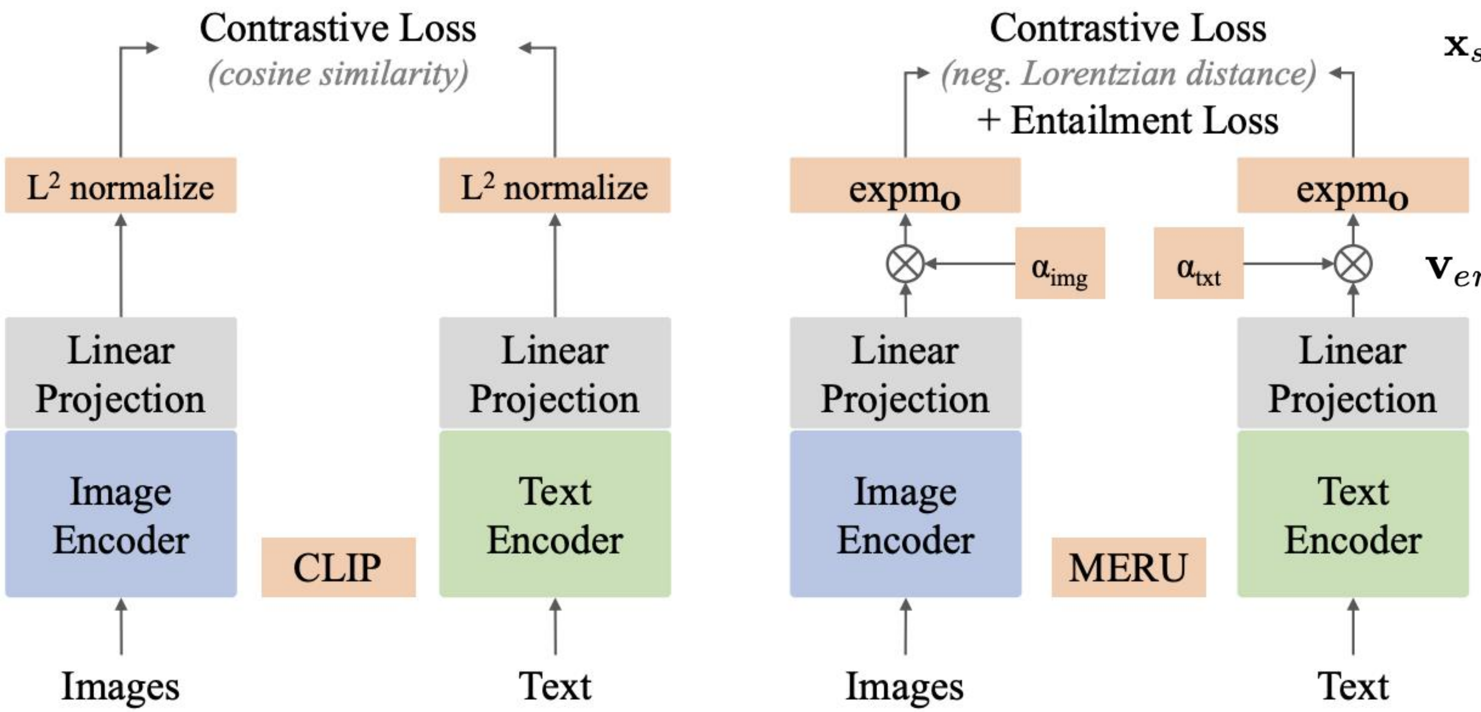
Background

- **MERU Model Design**



Background

- **MERU Model Design**



$$\mathbf{x}_{space} = \frac{sinh(\sqrt{c}\,\|\mathbf{v}_{space}\|)}{\sqrt{c}\,\|\mathbf{v}_{space}\|}\mathbf{v}_{space}$$

$$\mathbf{v}_{enc} \in \mathbb{R}^n \qquad \mathbf{v} = [\mathbf{v}_{enc}, 0] \in \mathbb{R}^{n+1}$$

Background

- **MERU Model Design**

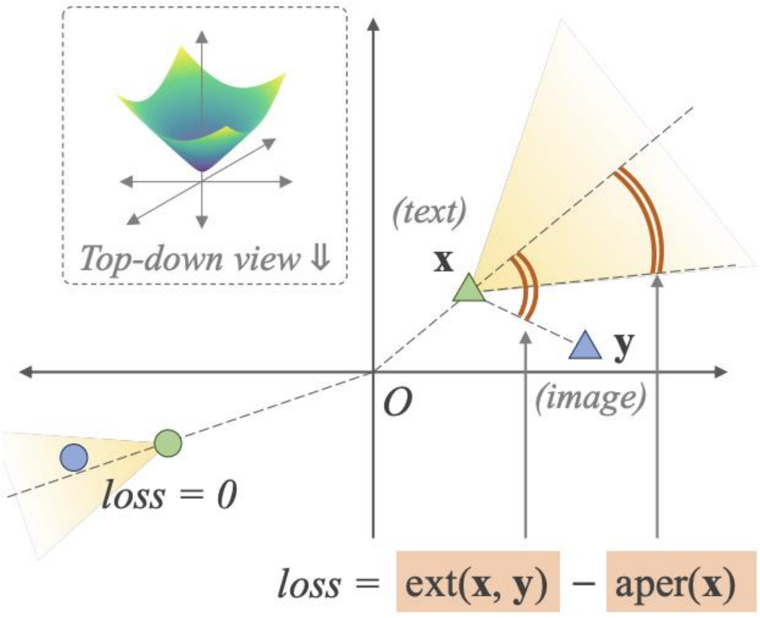
Contrastive Loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathbf{z}_i^I, \mathbf{z}_i^T)/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathbf{z}_i^I, \mathbf{z}_j^T)/\tau)}$$

Not cos similarity, but geodesics

Entailment loss

$$\mathcal{L}_{entail}(\mathbf{x}, \mathbf{y}) = \max(0,\ \text{ext}(\mathbf{x}, \mathbf{y}) - \text{aper}(\mathbf{x}))$$



$$\text{ext}(\mathbf{x}, \mathbf{y}) = \cos^{-1}\left(\frac{y_{time} + x_{time}\, c\, \langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}}}{\|\mathbf{x}_{space}\|\sqrt{(c\,\langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{L}})^2 - 1}}\right)$$

$$\text{aper}(\mathbf{x}) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\,\|\mathbf{x}_{space}\|}\right)$$

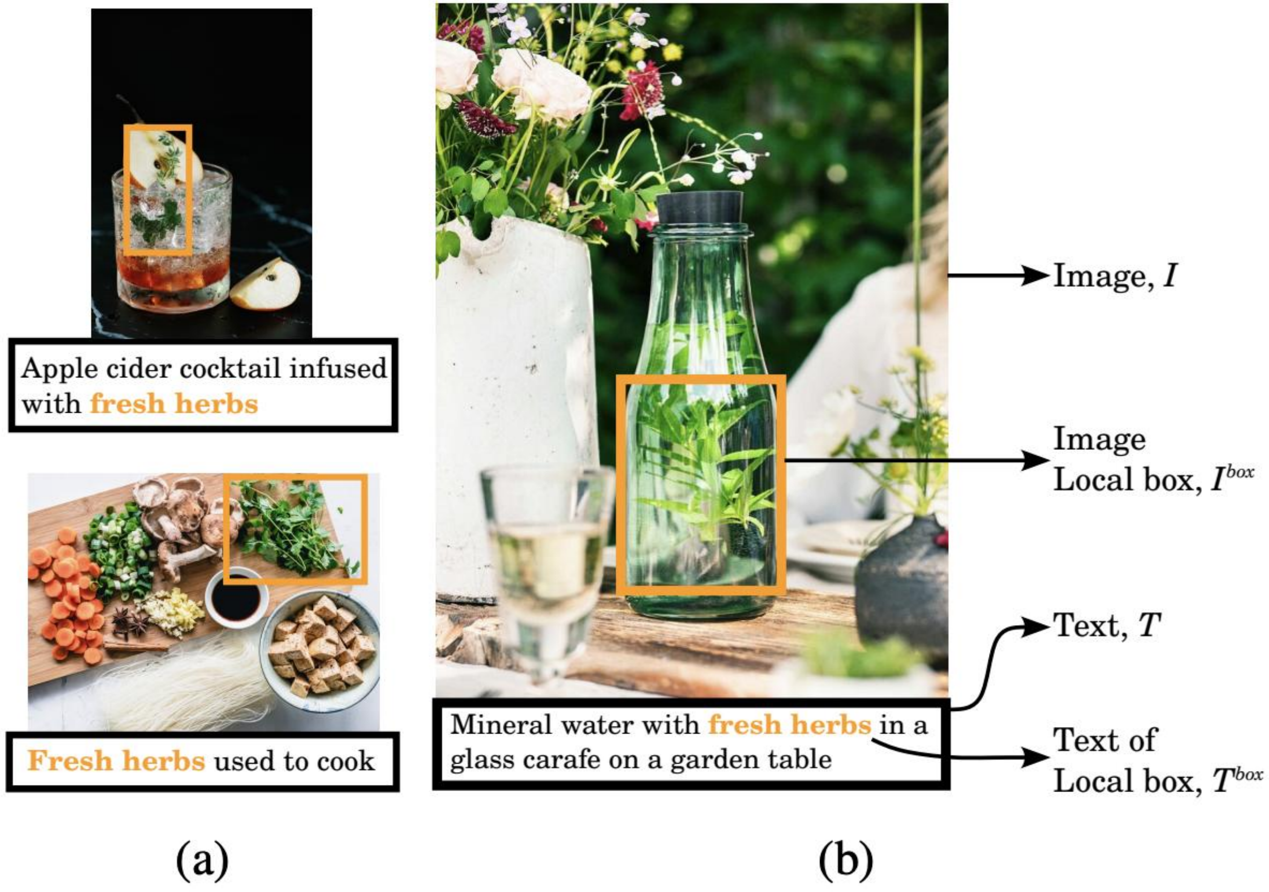
Le, M., Roller, S., Papaxanthos, L., Kiela, D., and Nickel, M. Inferring Concept Hierarchies from Text Corpora via Hyperbolic Embeddings. In Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL), 2019. 3, 4, 9

北京大學
PEKING UNIVERSITY

PART 03

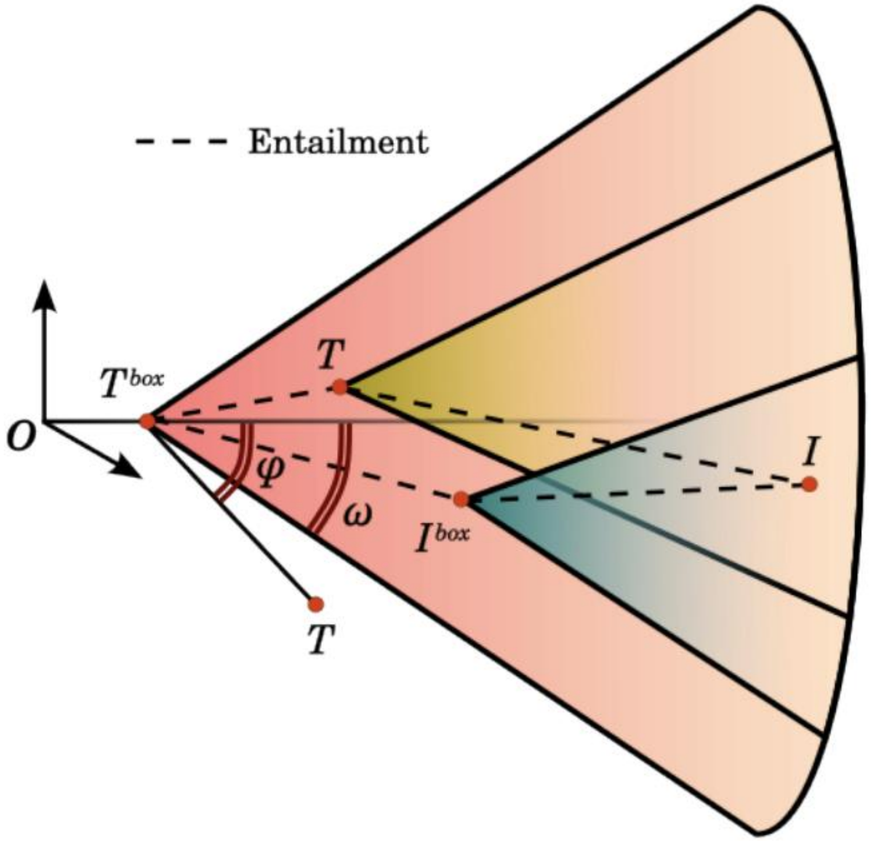
Method

Method



Object-scene hierarchy

Method



(c)

Method

Contrastive:

$$L^*_{cont}(I,T) = -\sum_{i\in B} \log \frac{\exp\left(d_{\mathbb{L}}(g_I(I_i), g_T(T_i))/\tau\right)}{\sum_{k=1,k\neq i}^{B} \exp\left(d_{\mathbb{L}}(g_I(I_i), g_T(T_k))/\tau\right)}$$

$$hCC(I,T,I^{box},T^{box}) = \frac{1}{4}\Big(\underbrace{L^*_{cont}(I,T) + L^*_{cont}(T,I)}_{\text{specific-info contrast}} + \underbrace{L^*_{cont}(I^{box},T) + L^*_{cont}(T^{box},I)}_{\text{general-info contrast}}\Big)$$

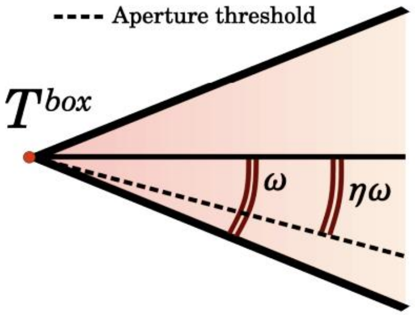
Entailment:

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

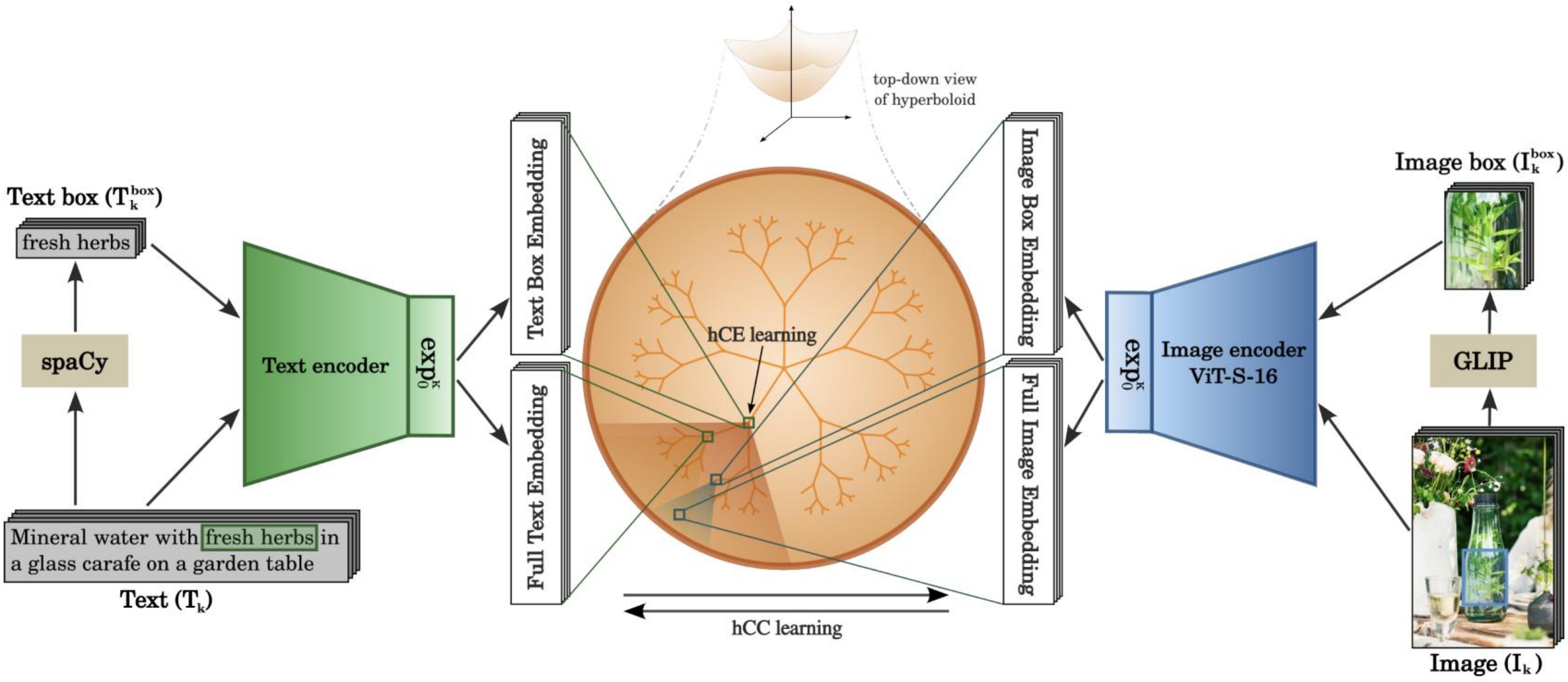
$$hCE(I,T,I^{box},T^{box}) = \underbrace{L^*_{ent}(I^{box},T^{box}) + L^*_{ent}(I,T)}_{\text{inter-modality entailment}} + \underbrace{L^*_{ent}(I,I^{box}) + L^*_{ent}(T,T^{box})}_{\text{intra-modality entailment}}$$

Total Loss:

$$hC = hCC + \gamma hCE$$



Method

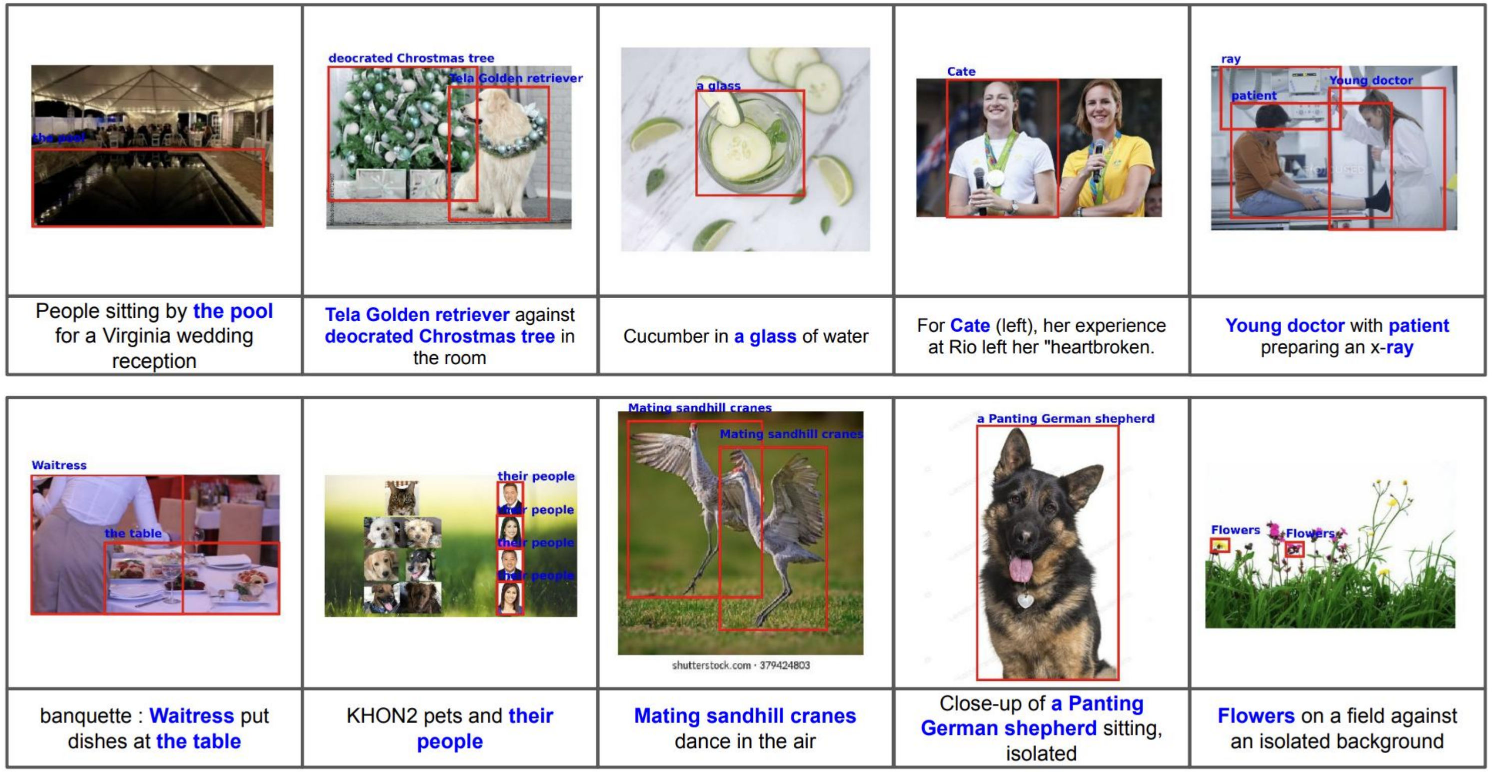

top-down view
of hyperboloid
Text box (T_k^{box})
fresh herbs
spaCy
Text encoder
$\exp_0^\kappa$
Mineral water with fresh herbs in
a glass carafe on a garden table
Text (T_k)
Text Box Embedding
Full Text Embedding
hCE learning
hCC learning
Image Box Embedding
Full Image Embedding
$\exp_0^\kappa$
Image encoder
ViT-S-16
Image box (I_k^{box})
GLIP
Image (I_k)

PART 04

Experiments

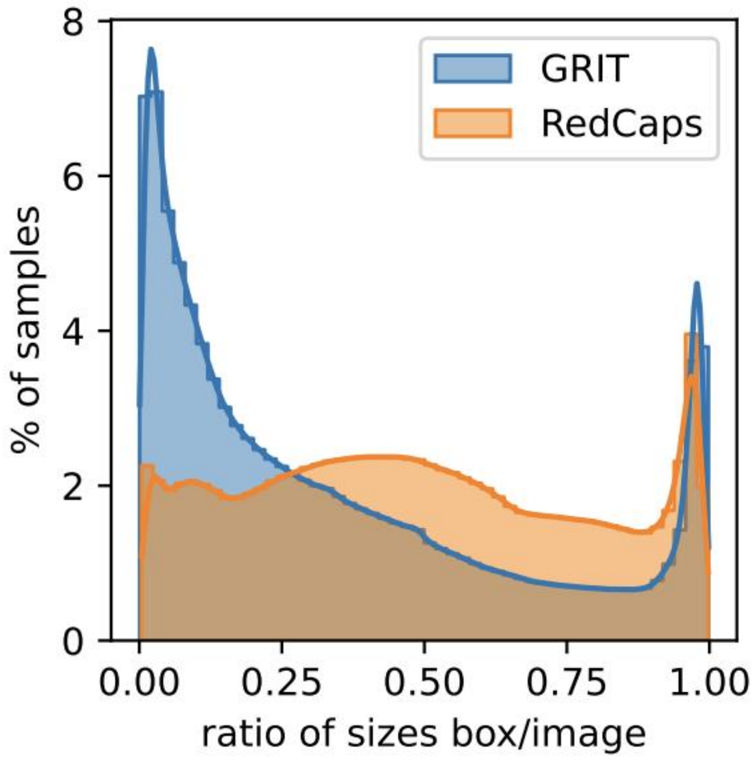
Experiments

- Datasets
- Grounded vision-language pairs
  - Grounded Image-Text Pairs (GRIT) dataset: 20.5 million pairs
  - RedCaps
  - Conceptual Captions 3M（CC3M)



Experiments

PEKING UNIVERSITY



Experiments

		w/ boxes	samples (M)	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
RedCaps																			
ViT S/16	CLIP†	✗	11.4	32.5	66.7	35.8	26.7	60.8	89.8	72.5	29.8	11.1	1.3	72.5	44.9	16.4	30.1	27.7	5.0
	CLIP	✓	11.4 [6.3]	30.2	76.5	42.4	25.8	62.3	89.5	69.6	25.7	8.5	2.2	65.3	38.6	13.6	36.6	28.5	4.6
	MERU†	✗	11.4	31.4	65.9	35.2	26.8	58.1	89.3	71.4	29.0	8.3	1.6	71.0	40.9	17.0	29.9	29.3	4.7
	MERU	✓	11.4 [6.3]	29.9	76.4	39.9	26.6	62.3	89.5	68.4	25.4	8.9	1.2	67.2	37.6	13.0	30.5	27.6	4.3
	HyCoCLIP	✓	5.8 [6.3]	31.9	77.4	37.7	27.6	64.5	90.9	71.1	28.8	9.7	1.1	70.5	41.4	13.4	22.7	30.7	4.4
GRIT																			
ViT S/16	CLIP	✗	20.5	36.7	70.2	42.6	49.5	73.6	89.7	44.7	9.8	6.9	2.0	44.6	14.8	22.3	**40.7**	40.1	5.1
	CLIP	✓	20.5 [35.9]	36.2	84.2	54.8	46.1	74.1	91.6	43.2	11.9	6.0	2.5	45.9	18.1	24.0	32.4	35.5	4.7
	MERU	✗	20.5	35.4	71.2	42.0	48.6	73.0	89.8	48.8	10.9	6.5	2.3	42.7	17.3	18.6	39.1	38.9	5.3
	MERU	✓	20.5 [35.9]	35.0	85.0	54.0	44.6	73.9	91.6	41.1	10.1	5.6	2.2	43.9	15.9	24.5	39.3	33.5	4.8
	HyCoCLIP	✓	20.5 [35.9]	41.7	85.0	53.6	52.5	75.7	92.5	50.2	14.7	8.1	**4.2**	52.0	20.5	22.3	33.8	**45.7**	5.2
ViT B/16	CLIP	✗	20.5	40.6	78.9	48.3	53.0	76.7	92.4	48.6	10.0	9.0	3.4	45.9	21.3	23.4	37.1	42.7	5.7
	MERU	✗	20.5	40.1	78.6	49.3	53.0	72.8	93.2	51.5	11.9	8.6	3.7	48.5	21.2	22.2	31.7	44.2	5.6
	HyCoCLIP	✓	20.5 [35.9]	**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**

Zero-shot classification

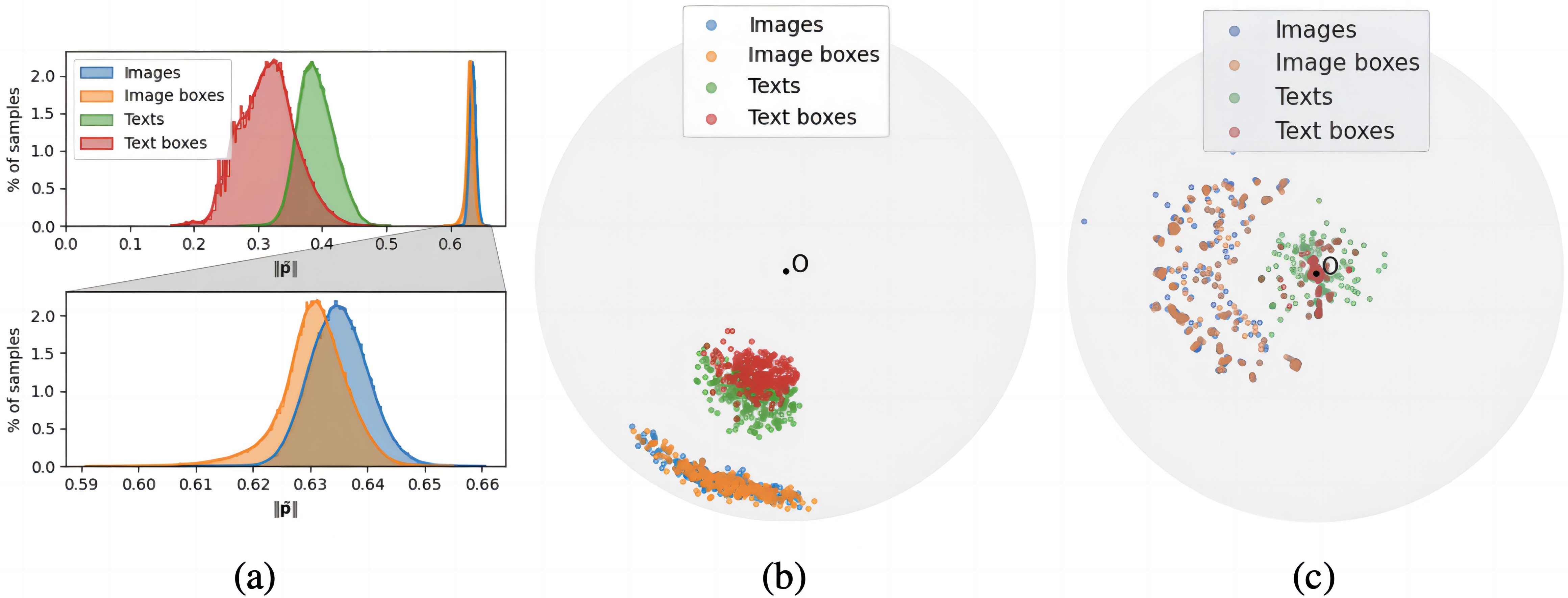
Experiments

Vision encoder	Model	w/ boxes	Text retrieval COCO R@5	Text retrieval COCO R@10	Text retrieval Flickr R@5	Text retrieval Flickr R@10	Image retrieval COCO R@5	Image retrieval COCO R@10	Image retrieval Flickr R@5	Image retrieval Flickr R@10	Hierarchical metrics WordNet TIE(↓)	LCA(↓)	J(↑)	P_H(↑)	R_H(↑)
ViT S/16	CLIP	✗	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
	CLIP	✓	60.7	71.8	84.2	91.3	47.1	58.6	73.1	82.1	4.03	2.38	0.76	0.83	0.83
	MERU	✗	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
	MERU	✓	**72.7**	**81.9**	83.5	90.1	46.6	58.3	60.0	71.7	4.08	2.39	0.75	0.83	0.83
	HyCoCLIP	✓	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**
ViT B/16	CLIP	✗	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	✗	**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
	HyCoCLIP	✓	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**

Zero-shot retrieval, detection, and hierarchical classification

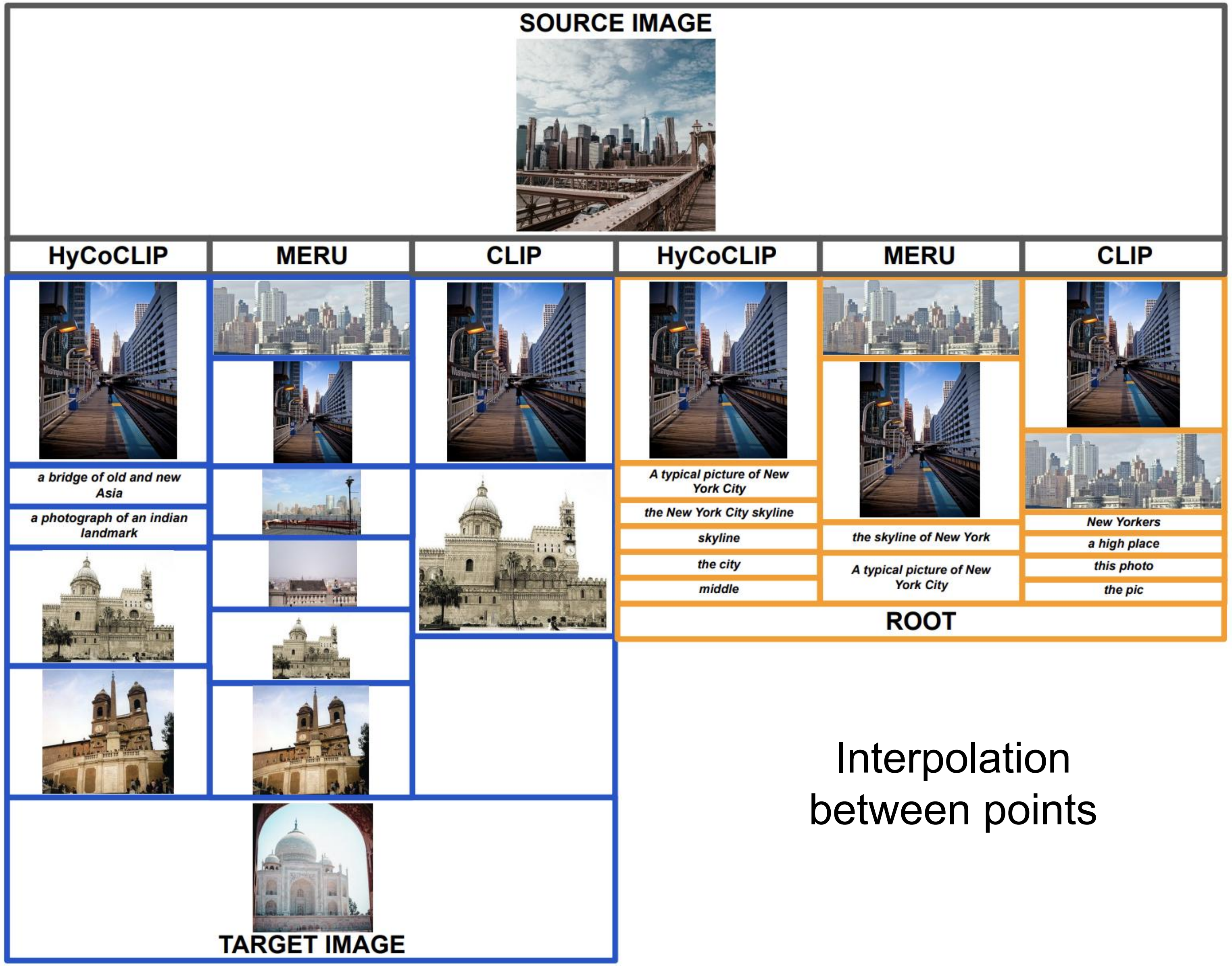
Model	AP
CLIP	51.2
MERU	55.8
RegionCLIP	65.2
HyCoCLIP	**68.5**

Experiments



Visualizing the learned hyperbolic space of HyCoCLIP in lower dimensions

Experiments



Interpolation between points

Thanks for Listening！