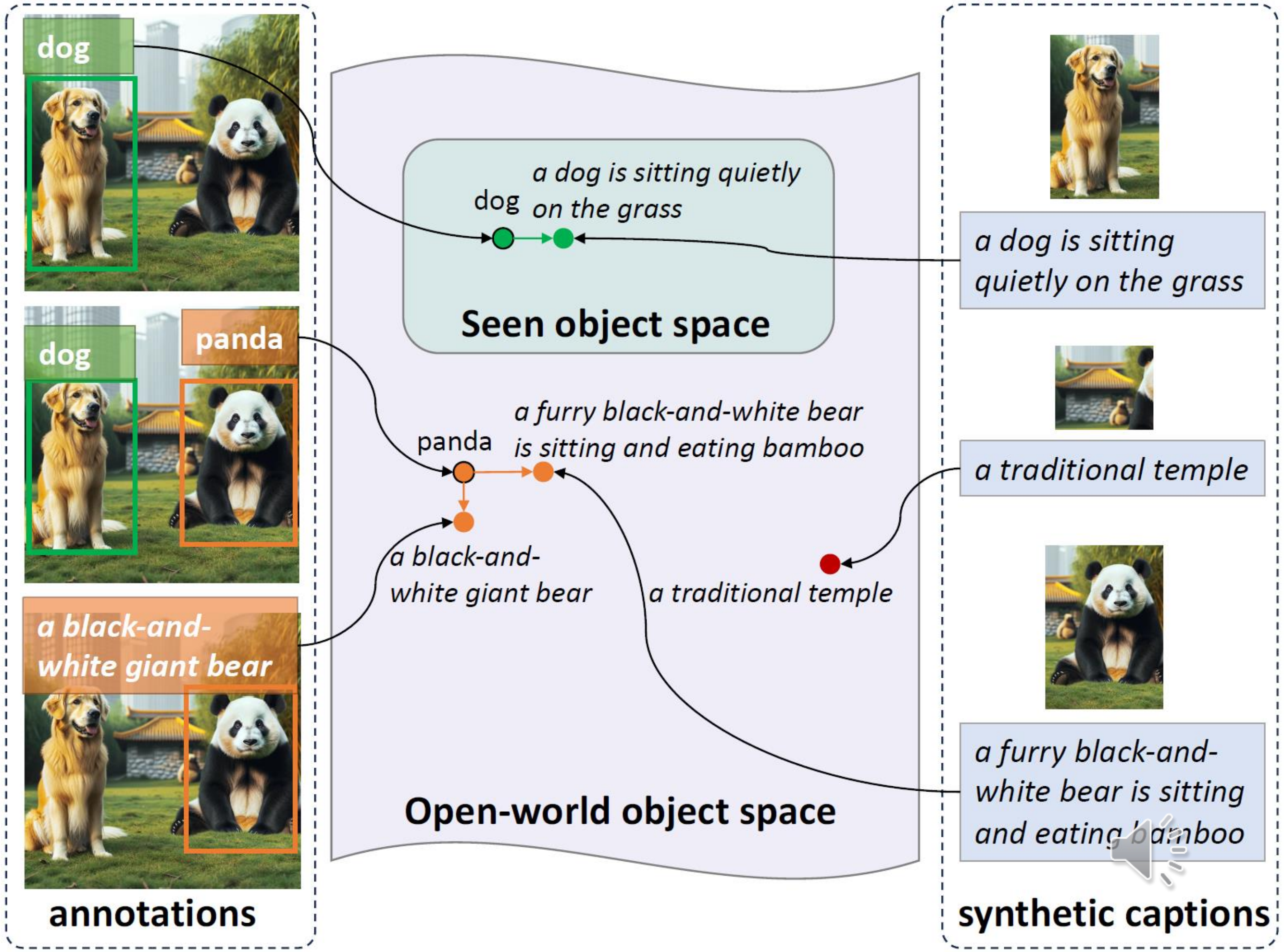
Hyperbolic Learning with Synthetic Captions for Open-World Detection

Fanjie Kong[1], Yanbei Chen[2], Jiarui Cai[2], Davide Modolo[2]

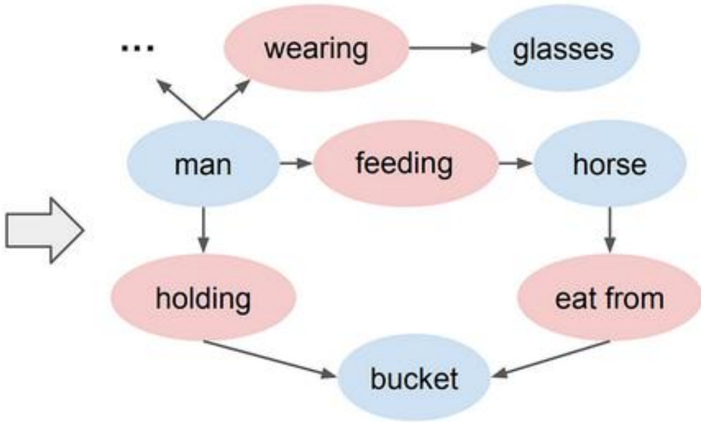
[1]Duke University, [2]AWS AI Labs

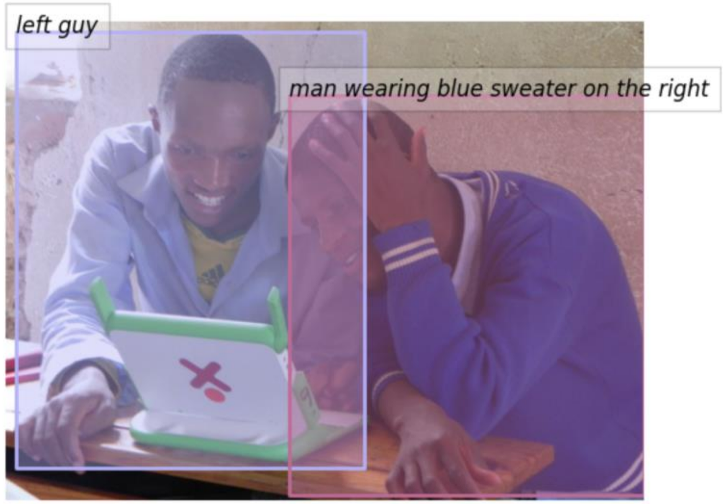
Open-world Object Detection

Goal: localize seen or unseen objects with pre-defined object vocabulary or contextual free-form text queries.



Open-world Object Detection





Free-form text annotations from Visual Genome (Krishna et al., 2016) and RefCOCO (Yu et al., 2016).

Challenges:

- High cost of manual annotations and human-crafted data acquisition pipeline.
- Localize objects described by both class labels and free-form texts

Open-world Object Detection

Previous work:

- Combine grounding data: GLIP (Li et al., 2021), GLIPv2 (Zhang et al., 2022)
- Innovate model design - Grounding DINO (Liu et al., 2023)

Open-world Object Detection

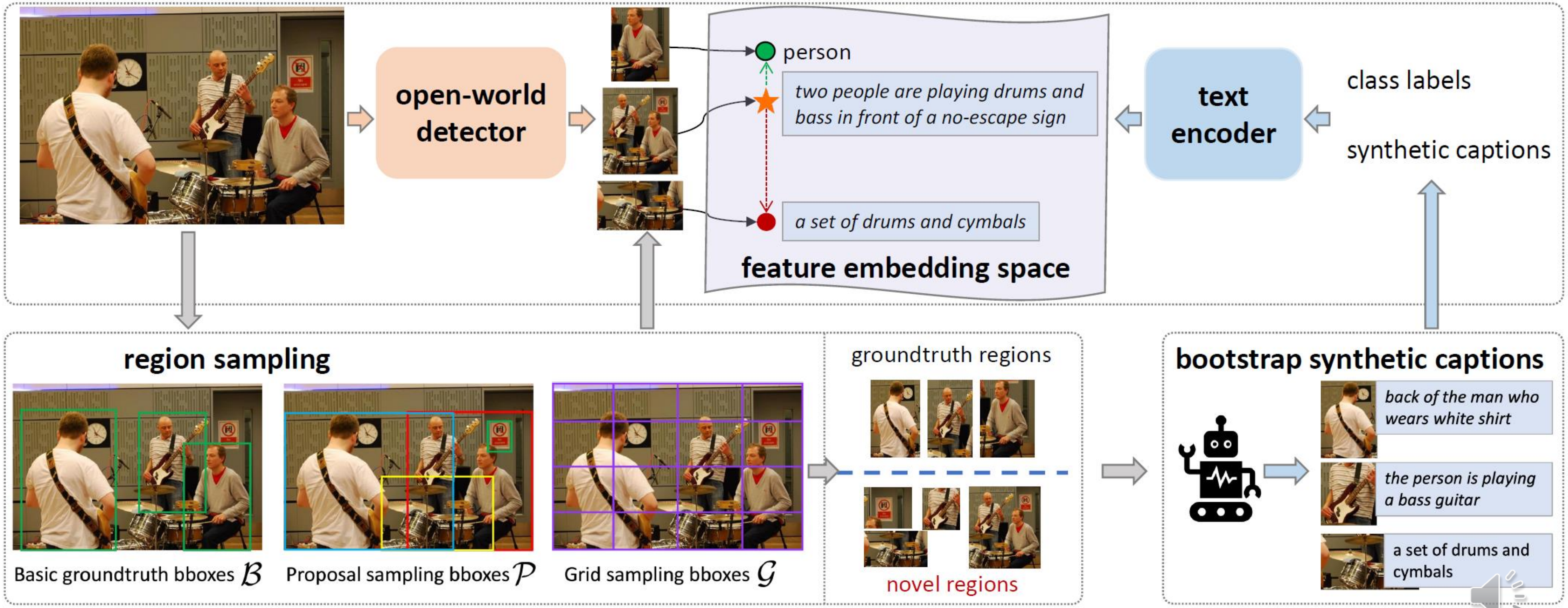
Previous work:

- Combine grounding data: GLIP (Li et al., 2021), GLIPv2 (Zhang et al., 2022)
- Innovate model design - Grounding DINO (Liu et al., 2023)

Our contribution:

- Leverage synthetic captions generated by VLMs to provide rich descriptions across different image regions;
- Introduce a novel hyperbolic vision-language learning method that aligns visual features with textual embeddings in a hierarchical structure.
- Achieve the state-of-the-art performance on a variety of detection and localization datasets in the open-world setting,

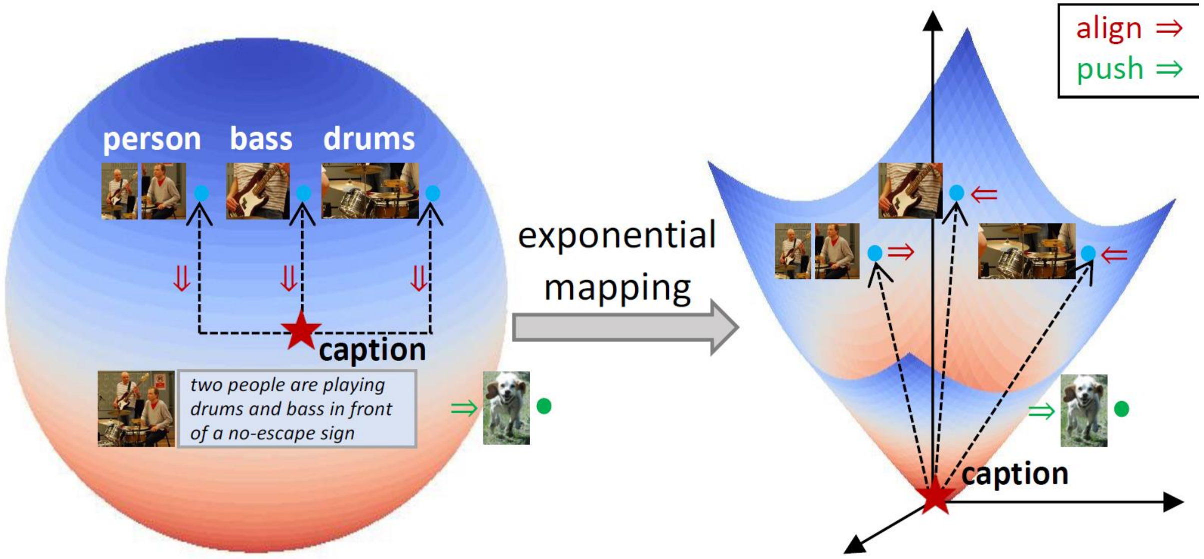
Our approach - *Hyperlearner*



Our hyperbolic vision-language learning approach exploits rich semantics from synthetic captions to boost open-world generalization.

Hyperbolic learning loss

Motivation: to mitigate the noise caused by hallucination in synthetic captions, we propose to impose a hierarchical relationship between visual and caption embeddings, where the caption and object adhere to a "caption entails object" hierarchy.



(a) Euclidean space (b) Hyperbolic space

Hyperbolic learning loss

Hyperbolic contrastive loss:

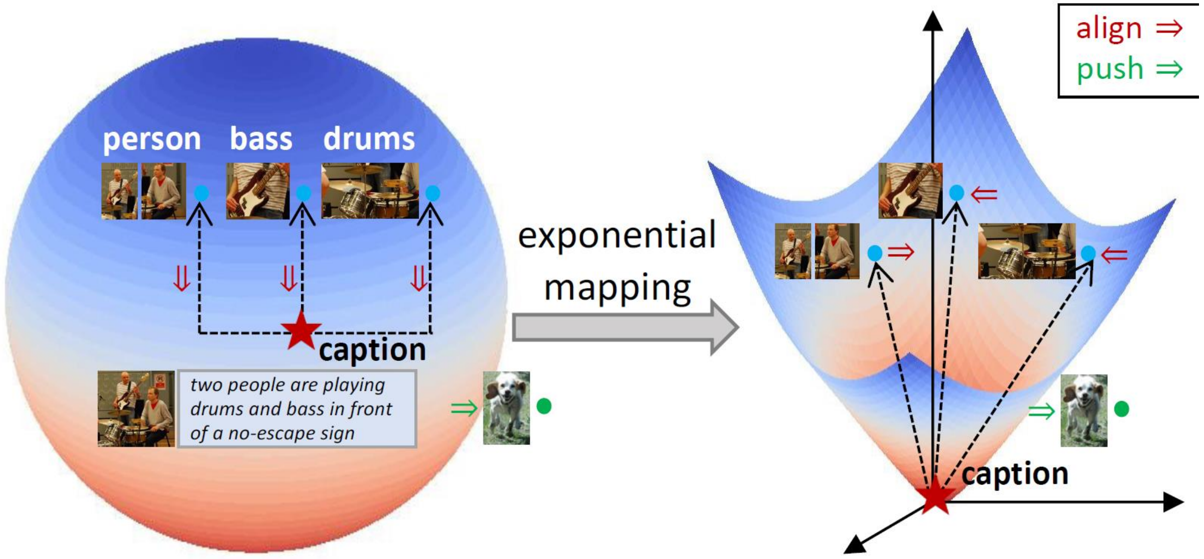
$$\mathrm{expm}_0(x) = \frac{\sinh(\sqrt{C}\|x\|)}{\sqrt{C}\|x\|},$$

$$\mathcal{L}_{cap}^{\mathcal{H}} = -\log \frac{\exp(-d_{\mathcal{H}}(v_i^{\mathcal{H}}, c_i^{\mathcal{H}})/\tau)}{\sum_{j=1}^{B} \exp(-d_{\mathcal{H}}(v_i^{\mathcal{H}}, c_j^{\mathcal{H}})/\tau)},$$

Hyperbolic entailment loss:

$$E(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) = \max(0, \angle(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) - A(c_i^{\mathcal{H}})),$$

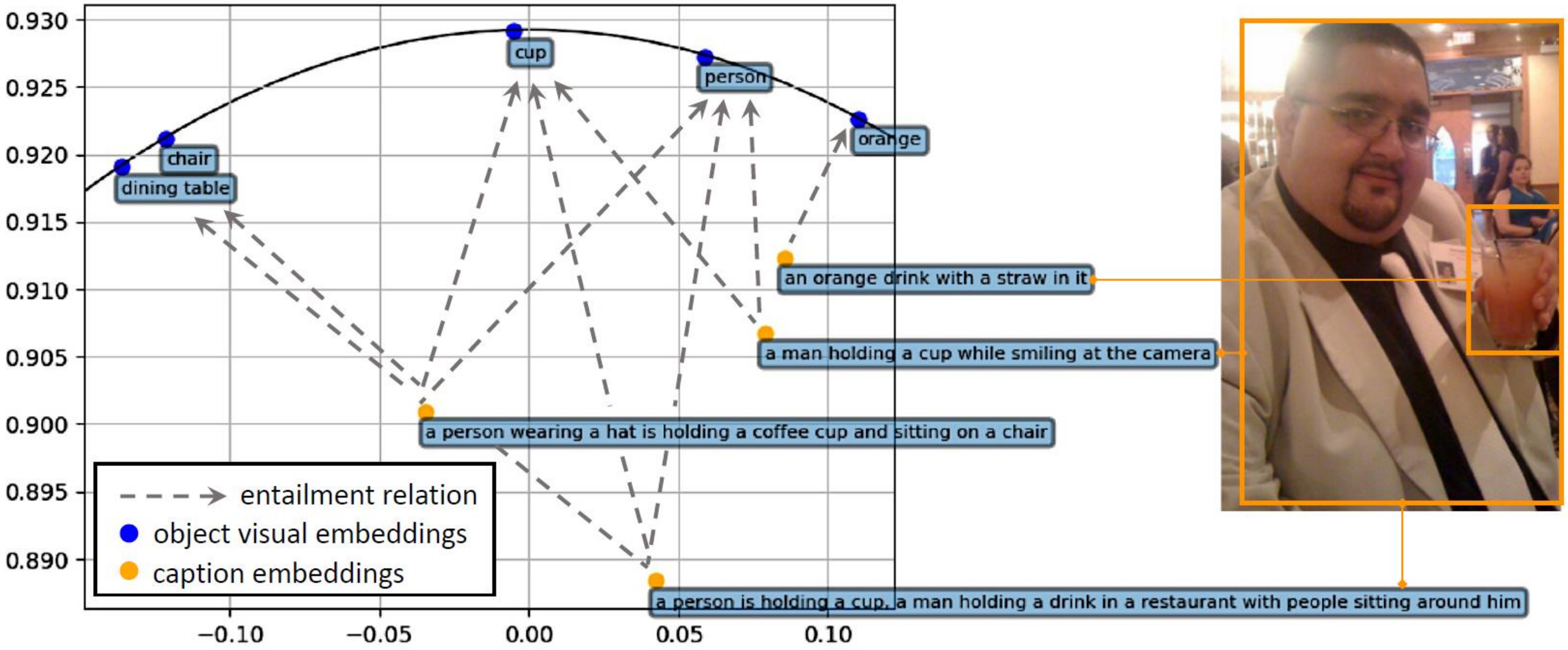
$$\mathcal{L}_{\text{entail}} = E(c_i^{\mathcal{H}}, v_i^{\mathcal{H}}) + \sum_{j\neq i} \max(0, \gamma - E(c_i^{\mathcal{H}}, v_j^{\mathcal{H}})),$$



(a) Euclidean space

(b) Hyperbolic space

Visualization of caption-object hierarchy



Evaluation

Tasks:

- open-world object detection
  - Datasets: COCO, LVIS, ODinW
- free-form text localization
  - RefCOCO/+/g
- Metric
  - mAP (Mean Average Precision) for detection tasks
  - Top-1 accuracy for referential expression localization tasks.

Evaluation – open-world object detection

	Method	Backbone	#Params	FLOPs	Pre-training Data	COCO2017 val	
						Zero-shot	Fine-tuning
1	Faster-RCNN [14]	RN50-FPN	42M	180G	COCO	-	40.2
2	Faster-RCNN [14]	RN101-FPN	54M	313G	COCO	-	42.0
3	Deformable DETR(DC5) [64]	RN50	41M	187G	COCO	-	41.1
4	CenterNetv2 [62]	RN50	76M	288G	COCO	-	42.9
5	Dyhead-T [6]	Swin-T	232M	361G	O365	43.6	53.3
6	GLIP-T(A) [28]	Swin-T	232M	488G	O365	42.9	52.9
7	GLIP-T(B) [28]	Swin-T	232M	488G	O365	44.9	53.8
8	GLIP-T(C) [28]	Swin-T	232M	488G	O365, GoldG	46.7	55.1
9	DINO-T [58]	Swin-T	-	-	O365	46.2	56.9
10	Grounding-DINO-T[1] [32]	Swin-T	172M	464G	O365	46.7	56.9
11	Grounding-DINO-T[2] [32]	Swin-T	172M	464G	O365, GoldG	48.1	57.1
12	Grounding-DINO-T[3] [32]	Swin-T	172M	464G	O365, GoldG, Cap4M	48.4	57.2
13	HyperLearner (Ours)	Swin-T	**90M**	**324G**	O365	**47.6**	**56.8**
14	HyperLearner (Ours)	Swin-T	**90M**	**324G**	O365, GoldG	**48.4**	**57.4**

Table 1. Comparison on COCO benchmark. Results are given on both zero-shot and fine-tuning settings. Metric: mAP.

Evaluation – open-world object detection

Method	Pre-training Data	LVIS minival AP	APr	APc	APf
MDETR [19]	GoldG, RefCOCO	24.2	20.9	24.9	24.3
DETCLIP-T [50]	O365	28.8	26.0	28.0	30.0
GLIP-T (C) [28]	O365, GoldG	24.9	17.7	19.5	31.0
GLIP-T [28]	O365, GoldG, Cap4M	26.0	20.8	21.4	31.0
Grounding-DINO-T [32]	O365, GoldG	25.6	14.4	19.6	32.2
Grounding-DINO-T [32]	O365, GoldG, Cap4M	27.4	20.8	21.4	31.0
HyperLearner (Ours)	O365	25.5	25.9	27.5	23.7
HyperLearner (Ours)	O365, GoldG	**31.3**	**30.7**	**32.6**	**30.3**

Table 2. Comparison on LVIS benchmark. Metric: mAP.

Method	Backbone	Pre-training Data	Test AP_{avg} zero-shot	full-shot
Detic-R [61]	RN50	LVIS, COCO, IN-21K	29.4	64.4
Detic-B [61]	Swin-B	LVIS, COCO, IN-21K	38.7	70.1
GLIP-T(A) [28]	Swin-T	O365	28.7	63.6
GLIP-T(B) [28]	Swin-T	O365	33.2	62.7
GLIP-T(C) [28]	Swin-T	O365, GoldG	44.4	63.9
Grounding-DINO-T [32]	Swin-T	O365, GoldG,Cap4M	44.9	-
HyperLearner (Ours)	Swin-T	O365	37.9	66.7
HyperLearner (Ours)	Swin-T	O365, GoldG	**45.2**	**68.9**

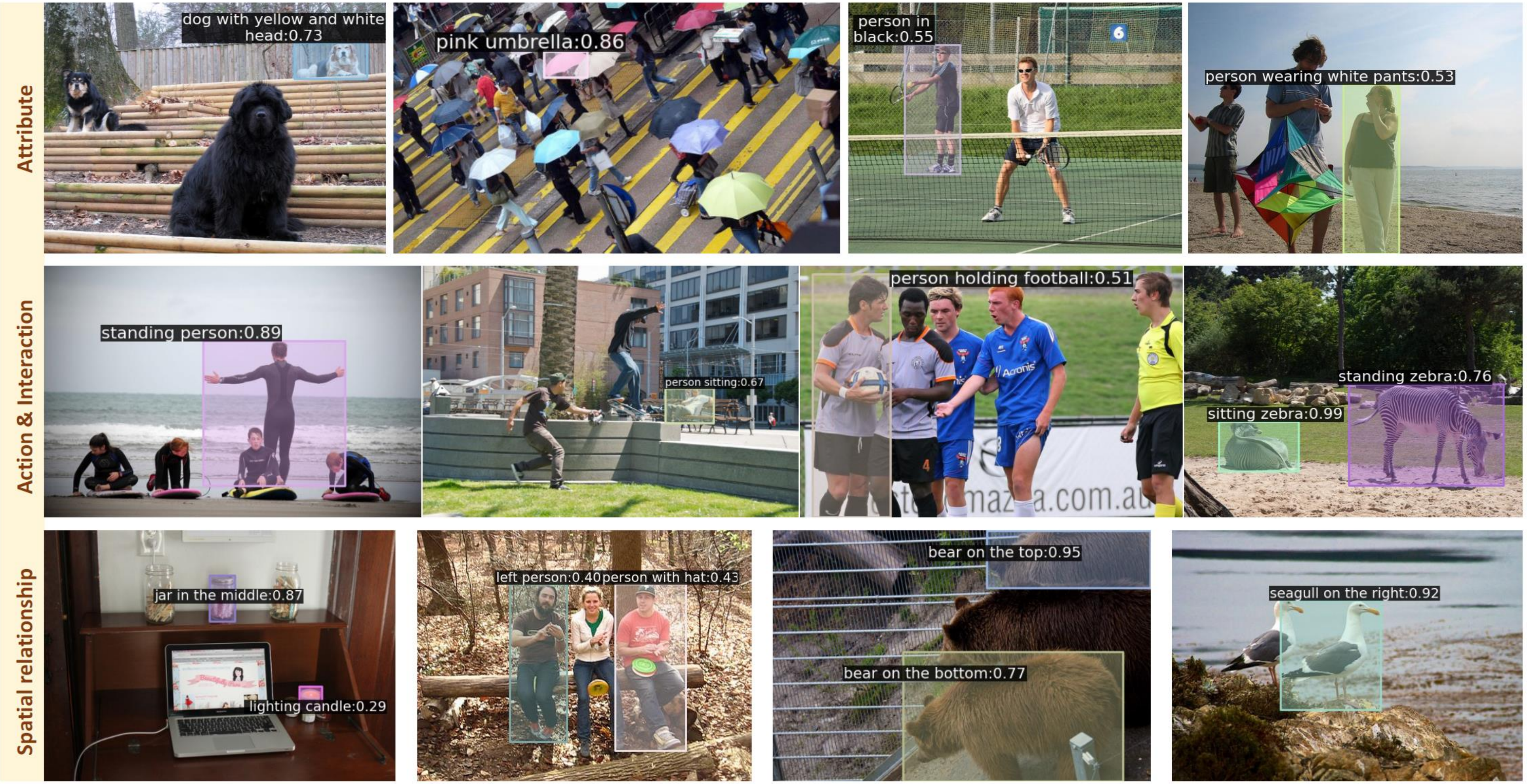
Table 4. Comparison on ODinW benchmark. Metric: mAP.

Evaluation – free-form text localization

	Method	Pre-training Data	Fine-tuning	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
6	GLIP-T(B) [28]	O365,GoldG	×	49.96	54.69	43.06	49.01	53.44	43.42	65.58	66.08
7	GLIP-T(C) [28]	O365,GoldG,Cap4M	×	50.42	54.30	43.83	49.50	52.78	44.59	66.09	66.89
8	Grounding-DINO-T [32]	O365,GoldG	×	50.41	57.24	43.21	51.40	57.59	45.81	67.46	67.13
9	Grounding-DINO-T [32]	O365,GoldG,RefC	×	73.98	74.88	59.29	66.81	69.91	56.09	71.06	72.07
10	Grounding-DINO-T [32]	O365,GoldG,RefC	✓	89.19	91.86	**85.99**	81.09	**87.40**	**74.71**	**84.15**	**84.94**
11	HyperLearner (Ours)	O365,GoldG	×	**50.66**	**60.87**	**44.66**	**59.29**	**62.29**	**45.43**	**67.02**	**67.44**
12	HyperLearner (Ours)	O365,GoldG,RefC	×	**77.89**	**76.92**	**72.99**	**67.54**	**75.55**	**57.54**	**77.00**	**76.79**
13	HyperLearner (Ours)	O365,GoldG,RefC	✓	**90.74**	**92.09**	85.46	**82.35**	84.70	72.64	82.53	82.39

Table 3. Comparison on RefCOCO/+/g benchmark. Metric: Top-1 accuracy.

Visualization



Summary

- We introduce a novel hyperbolic vision-language learning approach that effectively utilizes synthetic captions to enhance open-world object detection.
- Our comprehensive experiments demonstrate competitive performance across multiple benchmark datasets, supported by insightful ablation studies and qualitative analysis.
- This work establishes a foundational framework for extending hyperbolic learning to other vision learning tasks using synthetic data.

Thank you!