JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

上海科技大学
ShanghaiTech University

THU-PM-277

HOICLIP: Efficient Knowledge Transfer for HOI Detection with Vision-Language Models

Shan Ning*, Longtian Qiu*, Yongfei Liu, Xuming He

ShanghaiTech University, ByteDance Inc.

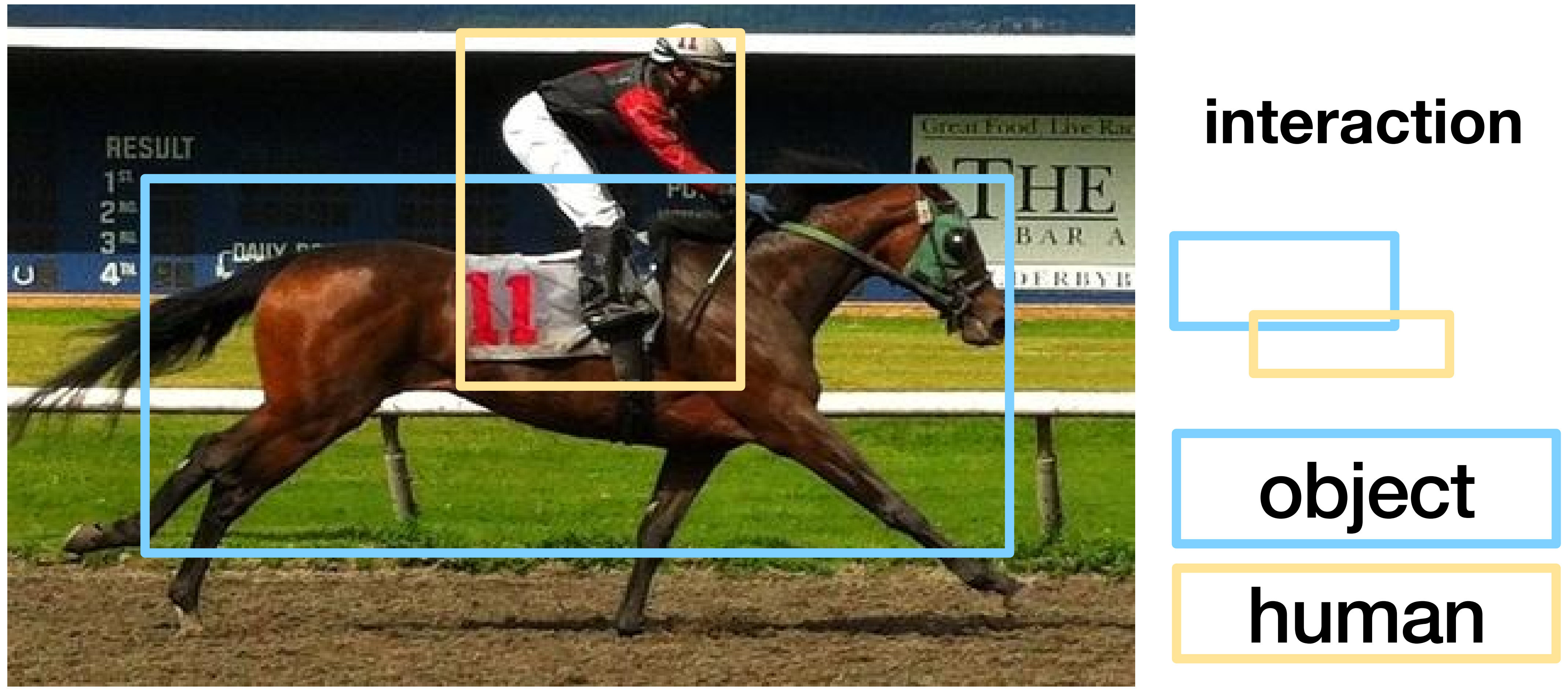
Accepted by CVPR 2023

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

上海科技大学 ShanghaiTech University

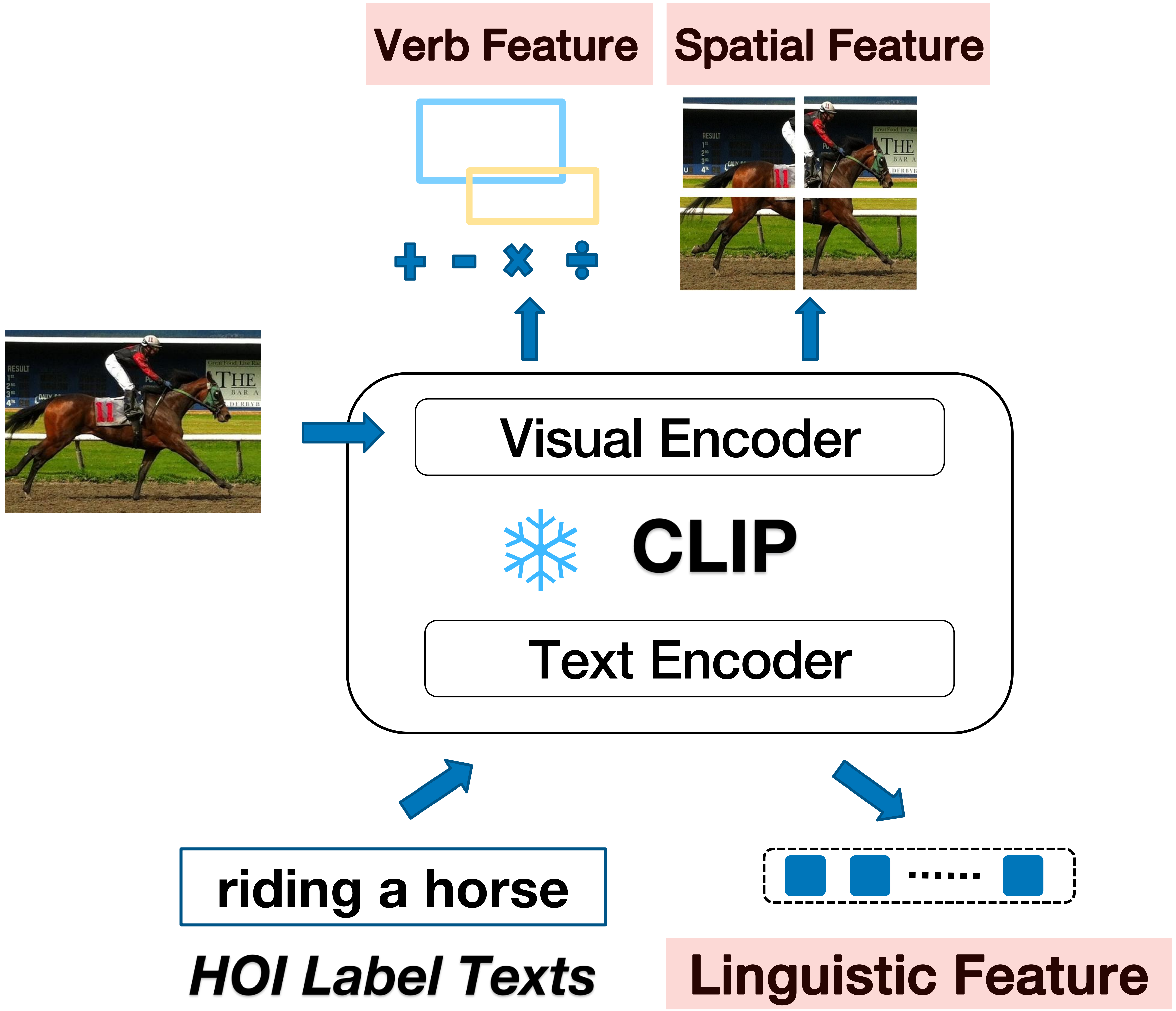
Task: Human Object Interaction Detection

<Human, Ride, Horse>



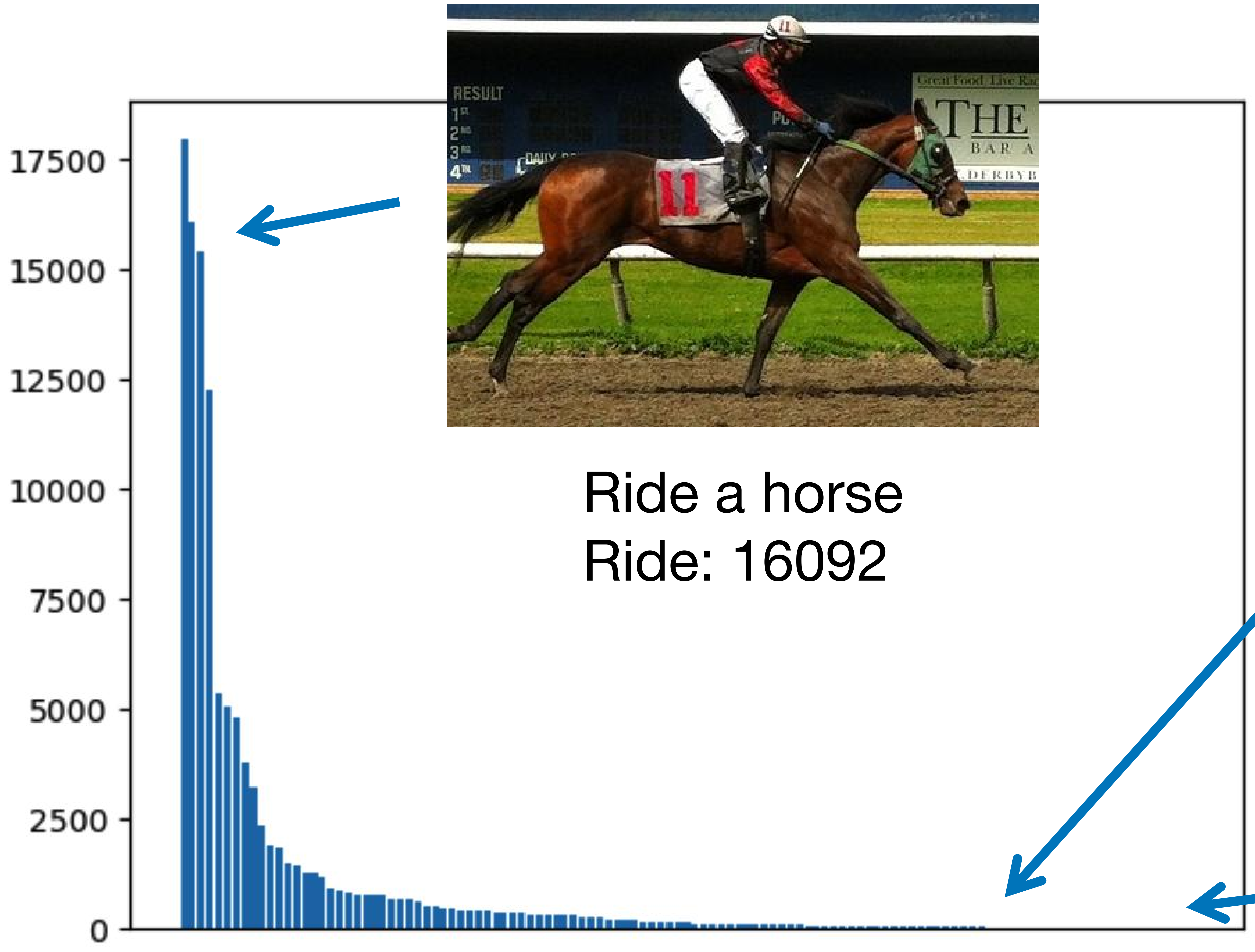
Problem: effective CLIP knowledge transfer for long-tail problem in HOI Detection.

Goal: improve the data efficiency in HOI representation learning and achieve better generalization as well as robustness.



HOI detection suffer from long tail problem in interaction understanding

Verb Sample Statistic Distribution in HICO-DET



Ride a horse
Ride: 16092

Spin a frisbee
Spin: 18

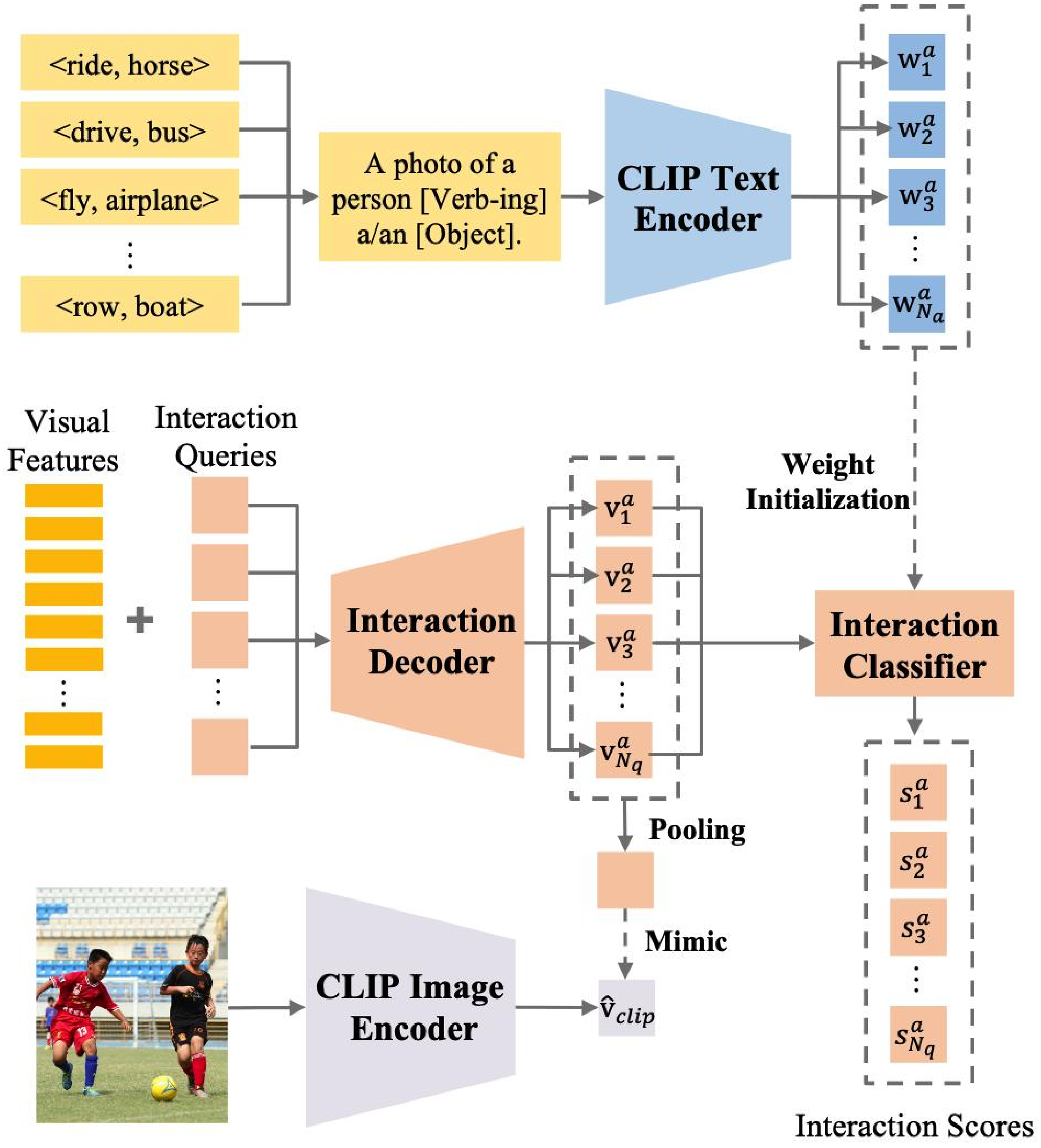
Dry a dog
Dry: 3

Related Works

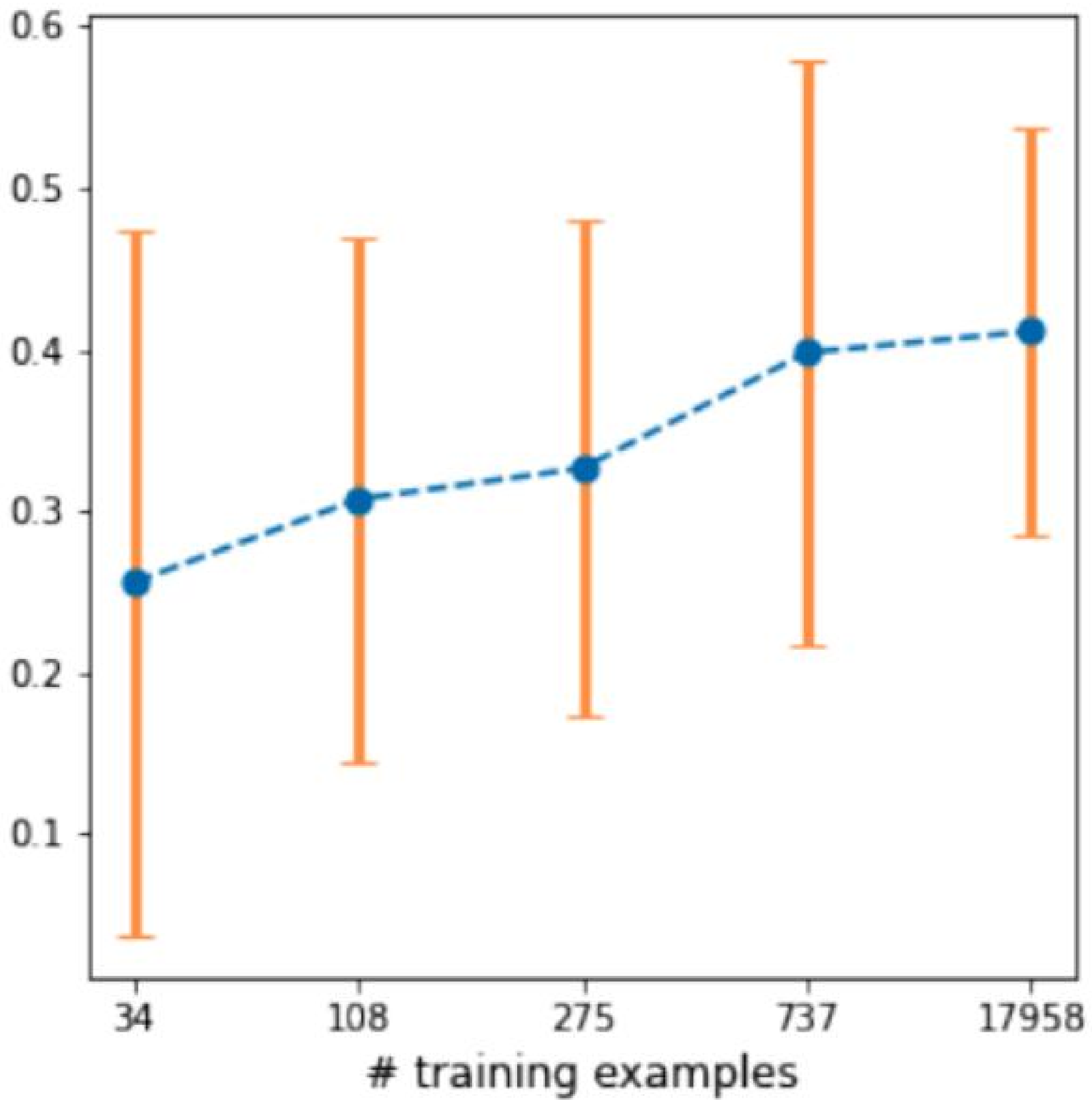
Previous methods leverage vision language model with knowledge distillation:

- Label Text embeddings constructed classifier[1,2].
- Image Feature[1] or logits[2] level knowledge distillation.

Vision Language Knowledge Transfer in GEN-VLKT [1]



Verb performance of GEN-VLKT [1]

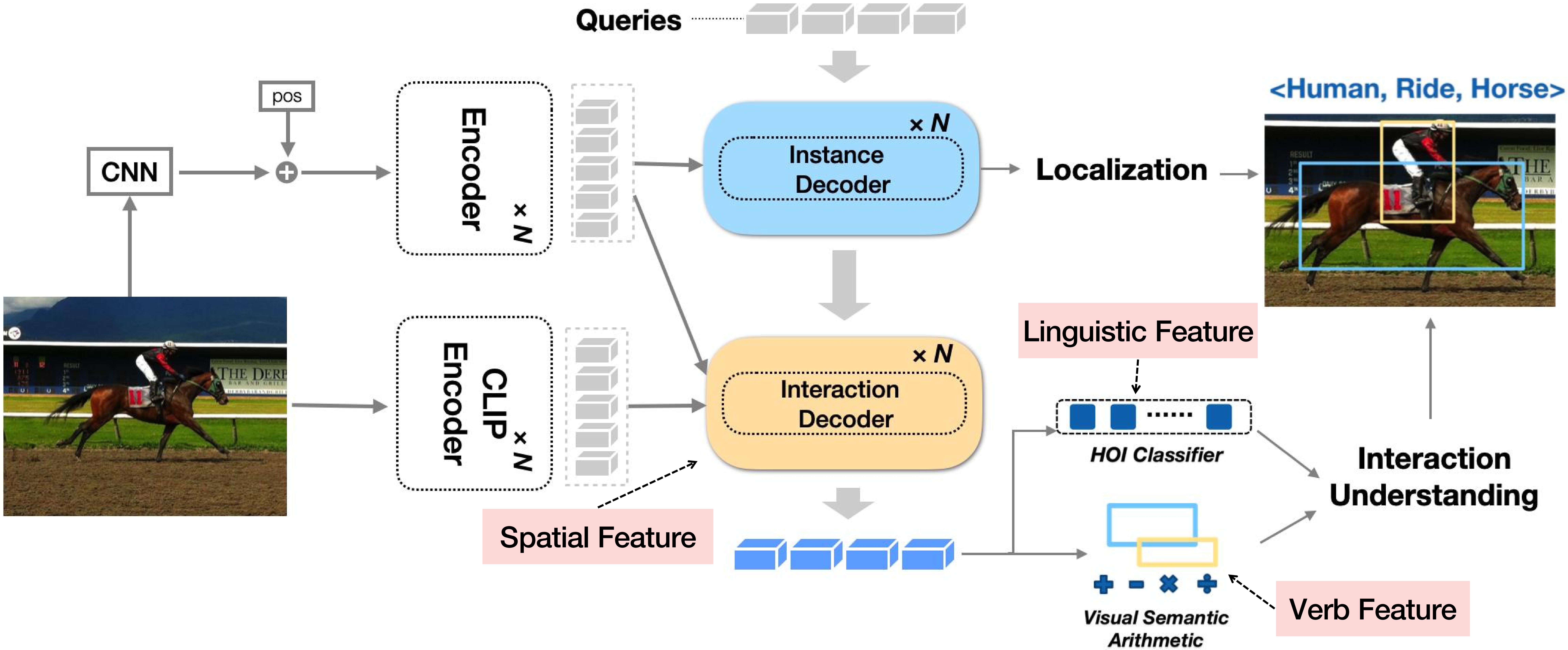


[1] Liao Y, Zhang A, Lu M, et al. CVPR 2022 [2] Gu, Xiuye et al. ICLR2022

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

上海科技大学 ShanghaiTech University

Overview of Our Method



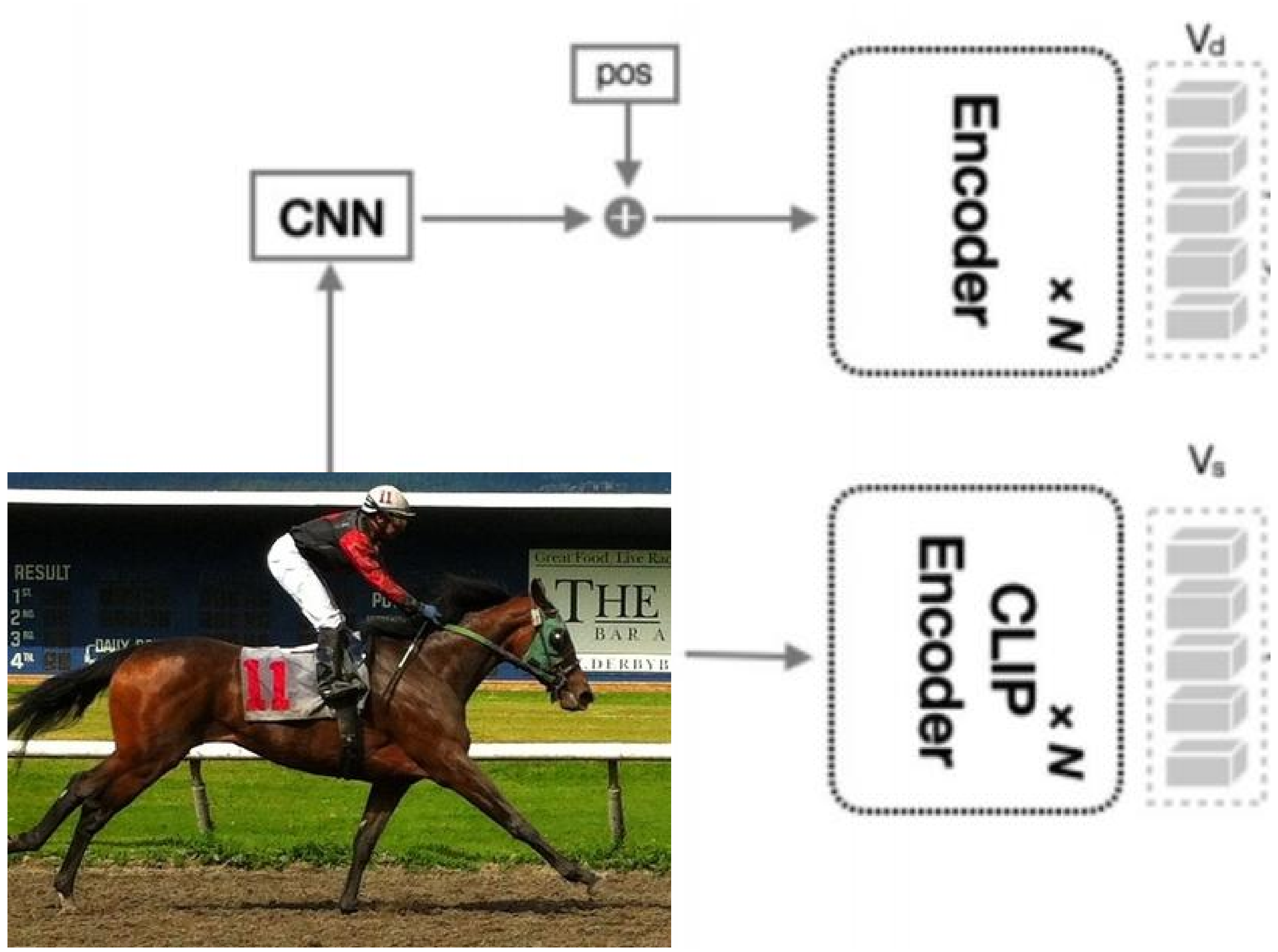
HOICLIP

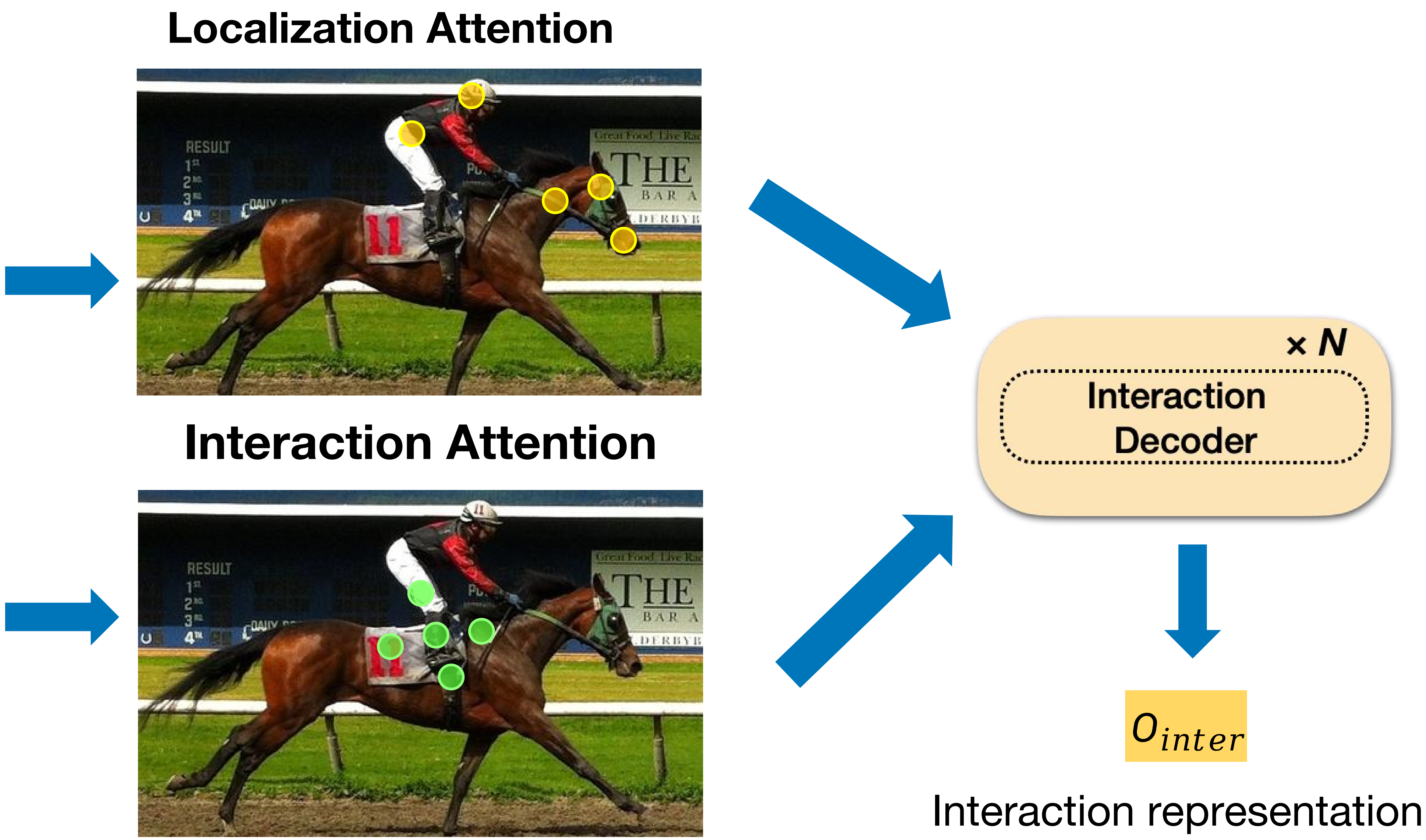
Spatial Feature

Interaction Decoder extracts interaction representations from two visual encoders:

1.projected detection visual feature $V_d^{'}$

2.CLIP spatial visual feature V_s



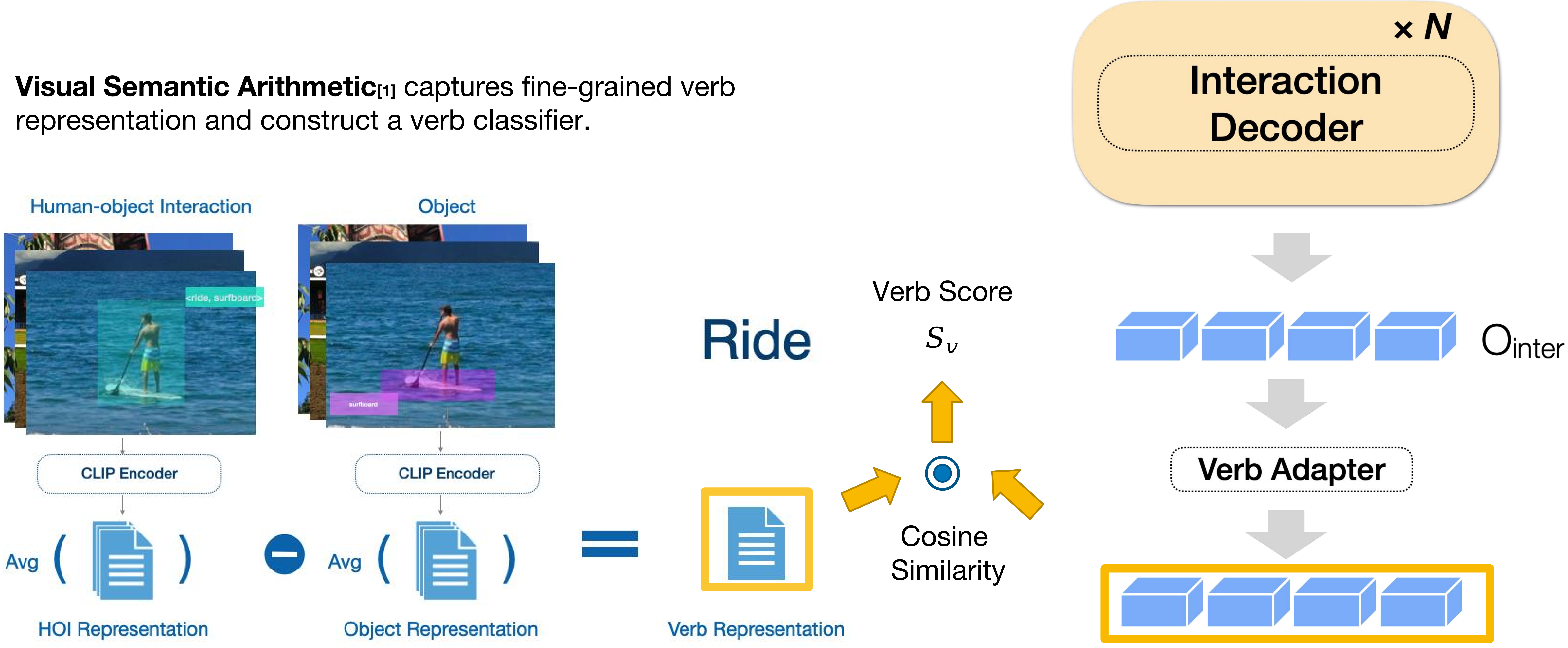

Localization Attention

Interaction Attention

Interaction representation

Verb Feature

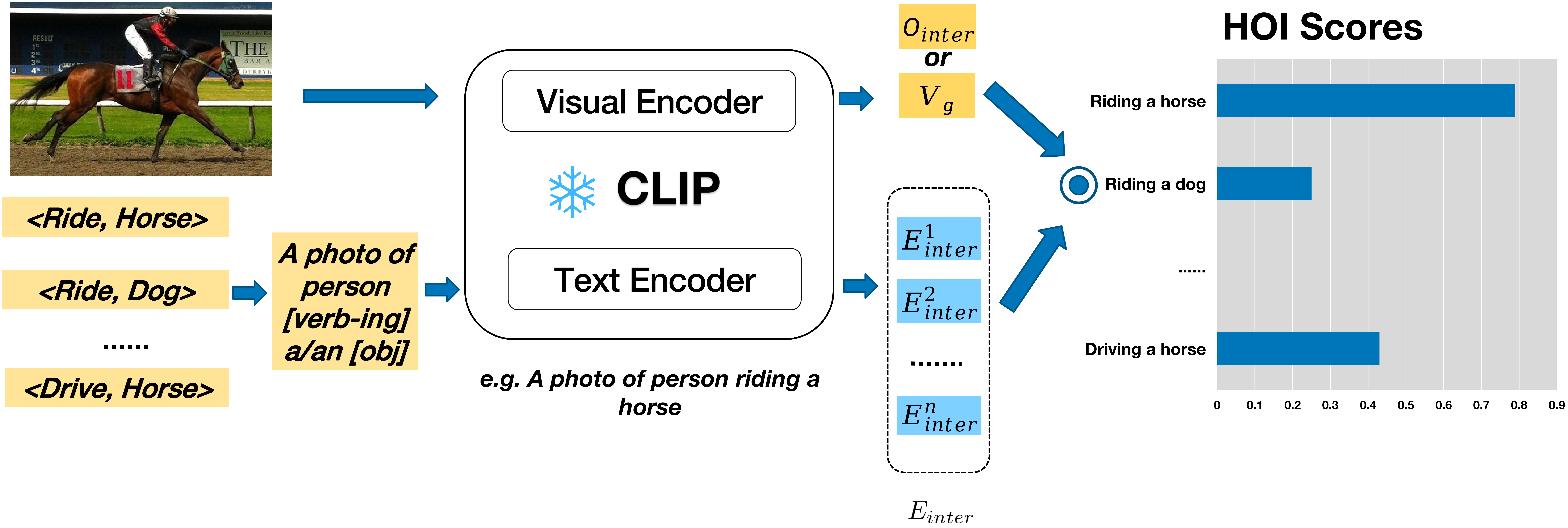
Visual Semantic Arithmetic[1] captures fine-grained verb representation and construct a verb classifier.



[1] Tewel, Yoad et al. CVPR 2022

Linguistic Feature

Linguistic Prior Knowledge in CLIP generate a HOI classifier which provides a training-free Enhancement for HOI classification.



<Ride, Horse>

<Ride, Dog>

......

<Drive, Horse>

A photo of person [verb-ing] a/an [obj]

Visual Encoder

CLIP

Text Encoder

e.g. A photo of person riding a horse

O_{inter} **or** V_g

E^1_{inter}

E^2_{inter}

....

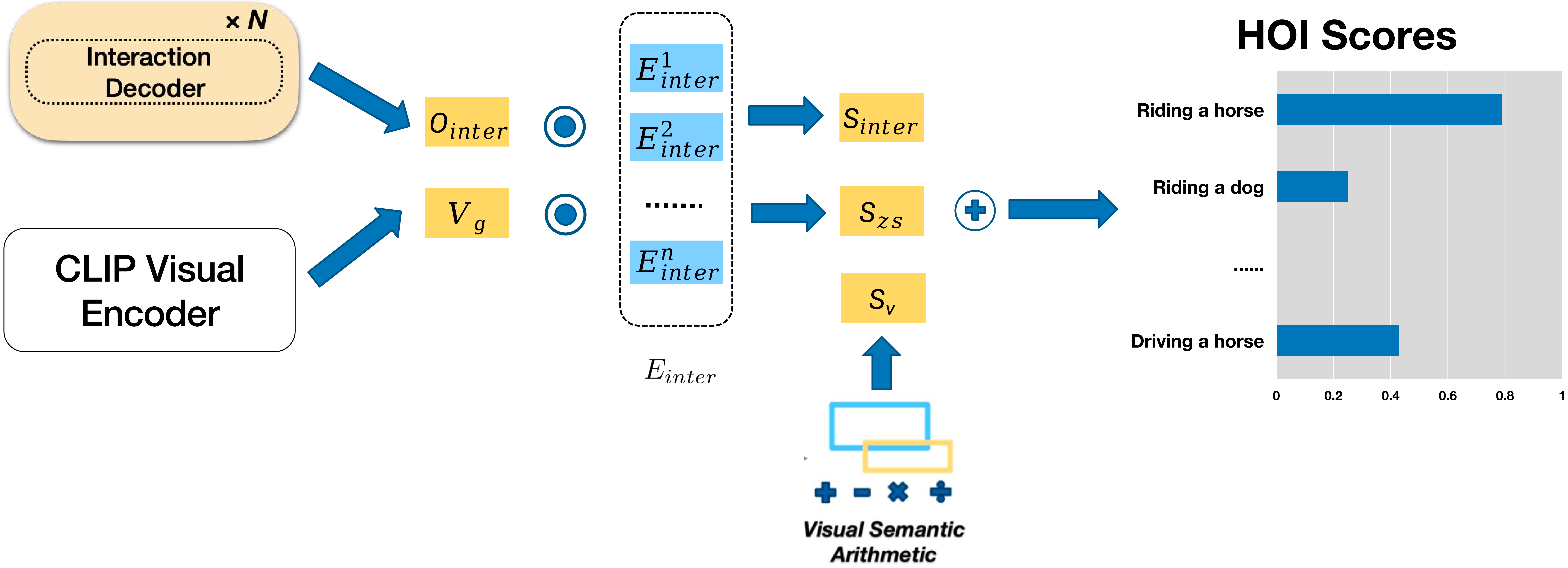
E^n_{inter}

E_{inter}

HOI Scores

Interaction Inference

Final interaction prediction is a weight sum of verb score, interaction predcition score and zero-shot interaction prediction score.



Experiments

Benchmark:

- HICO-DET contains ~**48k** images, **600** HOI categories
- V-COCO contains ~**10k** images, **29** verb categories

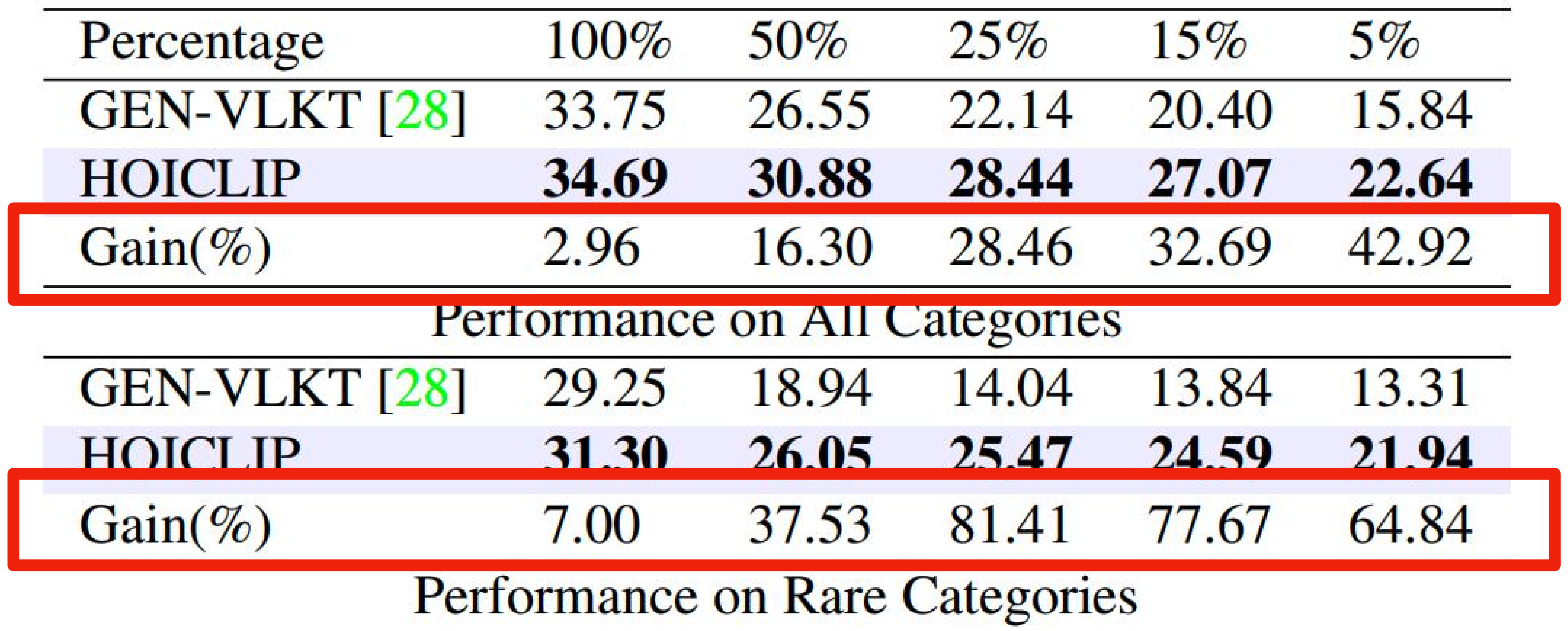
Experiment settings

- **Low-data** HOI Detection
- **Zero-shot** HOI Detection
- **Standard** HOI Detection

Evaluation Metric: **mAP**

Low-data HOI Detection

Results on HICO-DET

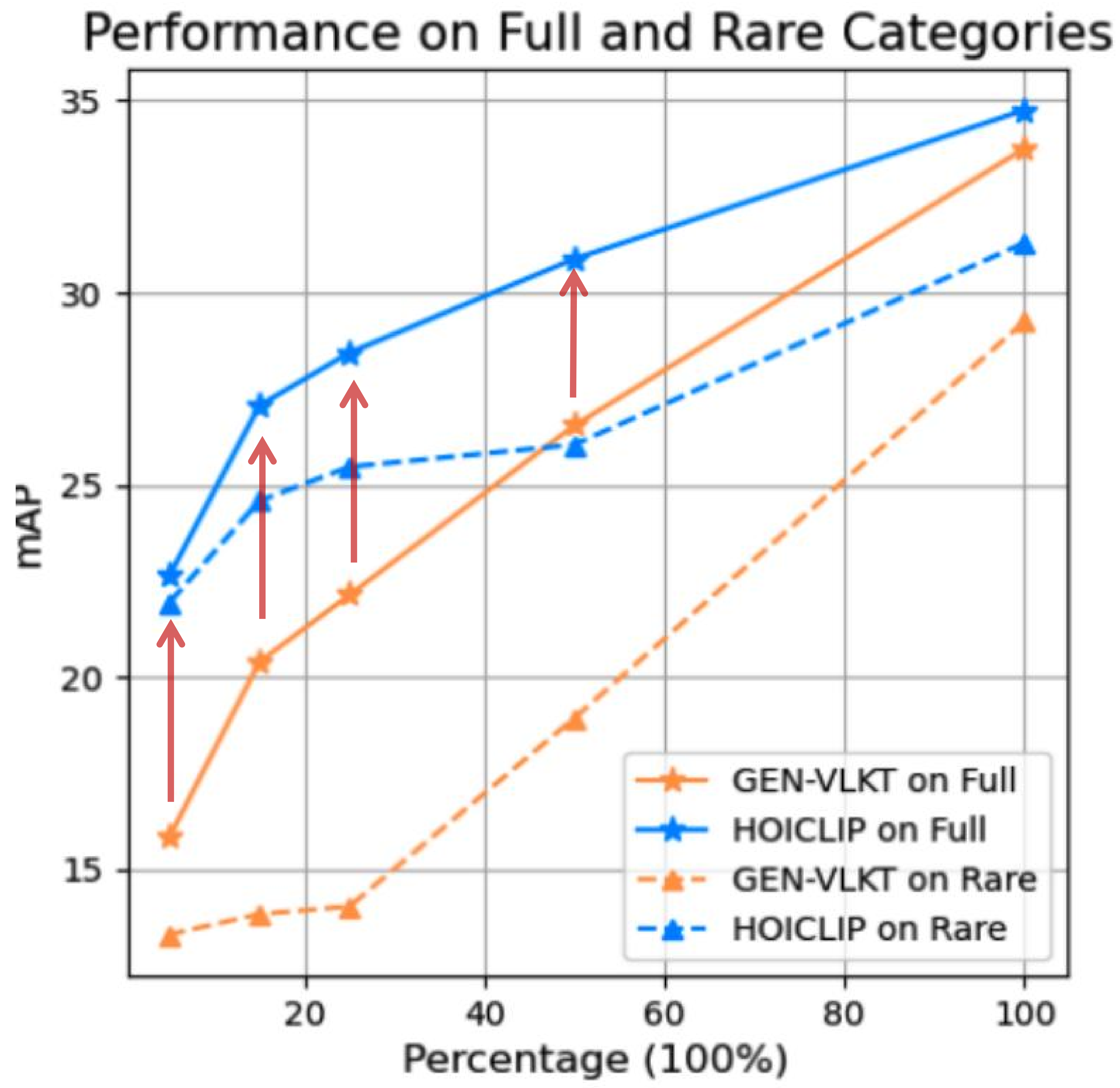


Percentage	100%	50%	25%	15%	5%
GEN-VLKT [28]	33.75	26.55	22.14	20.40	15.84
HOICLIP	**34.69**	**30.88**	**28.44**	**27.07**	**22.64**
Gain(%)	2.96	16.30	28.46	32.69	42.92

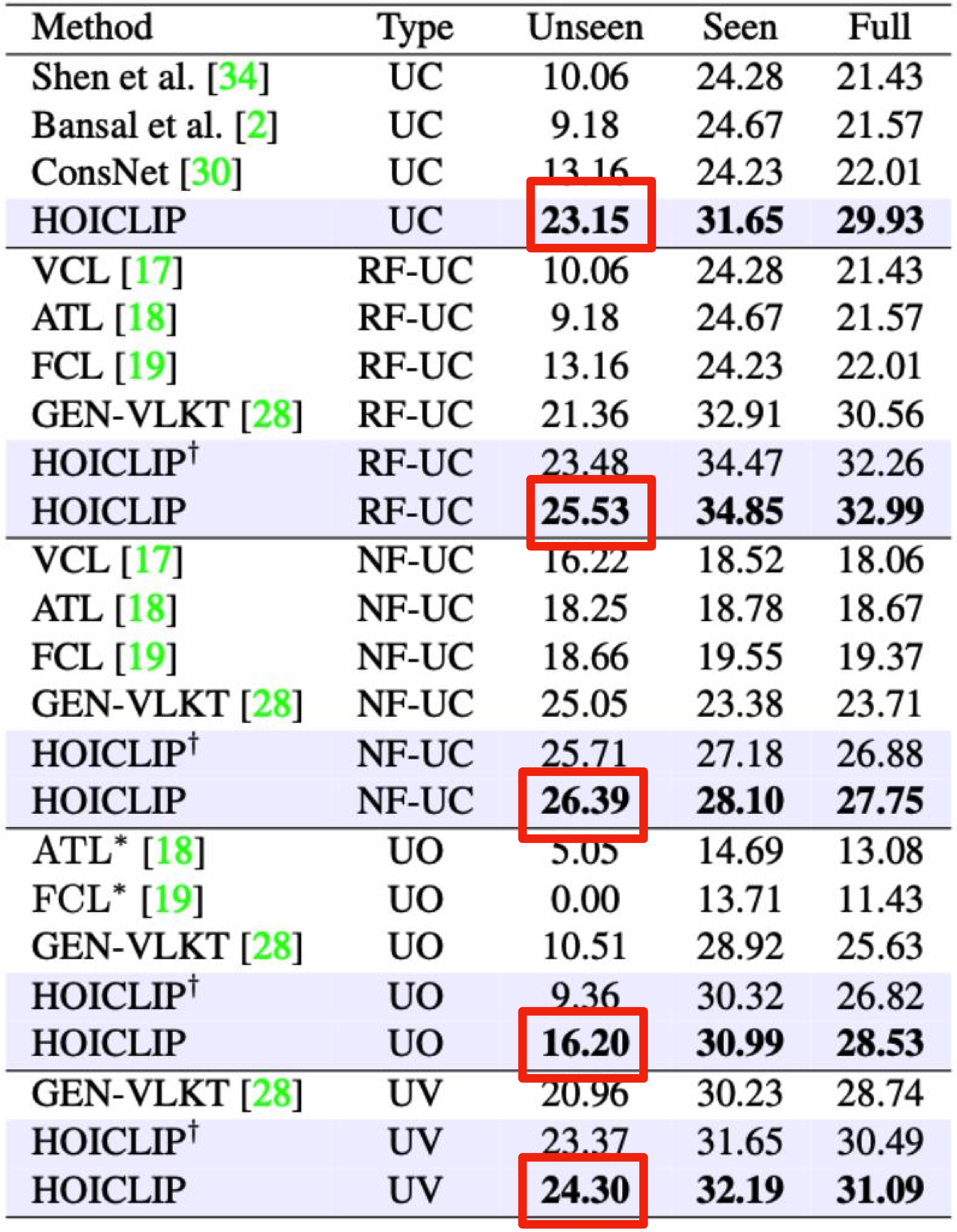
Performance on All Categories

GEN-VLKT [28]	29.25	18.94	14.04	13.84	13.31
HOICLIP	**31.30**	**26.05**	**25.47**	**24.59**	**21.94**
Gain(%)	7.00	37.53	81.41	77.67	64.84

Performance on Rare Categories



Zero-shot HOI Detection



Method	Type	Unseen	Seen	Full
Shen et al. [34]	UC	10.06	24.28	21.43
Bansal et al. [2]	UC	9.18	24.67	21.57
ConsNet [30]	UC	13.16	24.23	22.01
HOICLIP	UC	**23.15**	**31.65**	**29.93**
VCL [17]	RF-UC	10.06	24.28	21.43
ATL [18]	RF-UC	9.18	24.67	21.57
FCL [19]	RF-UC	13.16	24.23	22.01
GEN-VLKT [28]	RF-UC	21.36	32.91	30.56
HOICLIP†	RF-UC	23.48	34.47	32.26
HOICLIP	RF-UC	**25.53**	**34.85**	**32.99**
VCL [17]	NF-UC	16.22	18.52	18.06
ATL [18]	NF-UC	18.25	18.78	18.67
FCL [19]	NF-UC	18.66	19.55	19.37
GEN-VLKT [28]	NF-UC	25.05	23.38	23.71
HOICLIP†	NF-UC	25.71	27.18	26.88
HOICLIP	NF-UC	**26.39**	**28.10**	**27.75**
ATL* [18]	UO	5.05	14.69	13.08
FCL* [19]	UO	0.00	13.71	11.43
GEN-VLKT [28]	UO	10.51	28.92	25.63
HOICLIP†	UO	9.36	30.32	26.82
HOICLIP	UO	**16.20**	**30.99**	**28.53**
GEN-VLKT [28]	UV	20.96	30.23	28.74
HOICLIP†	UV	23.37	31.65	30.49
HOICLIP	UV	**24.30**	**32.19**	**31.09**

↑ 75.91%

↑ 19.52%

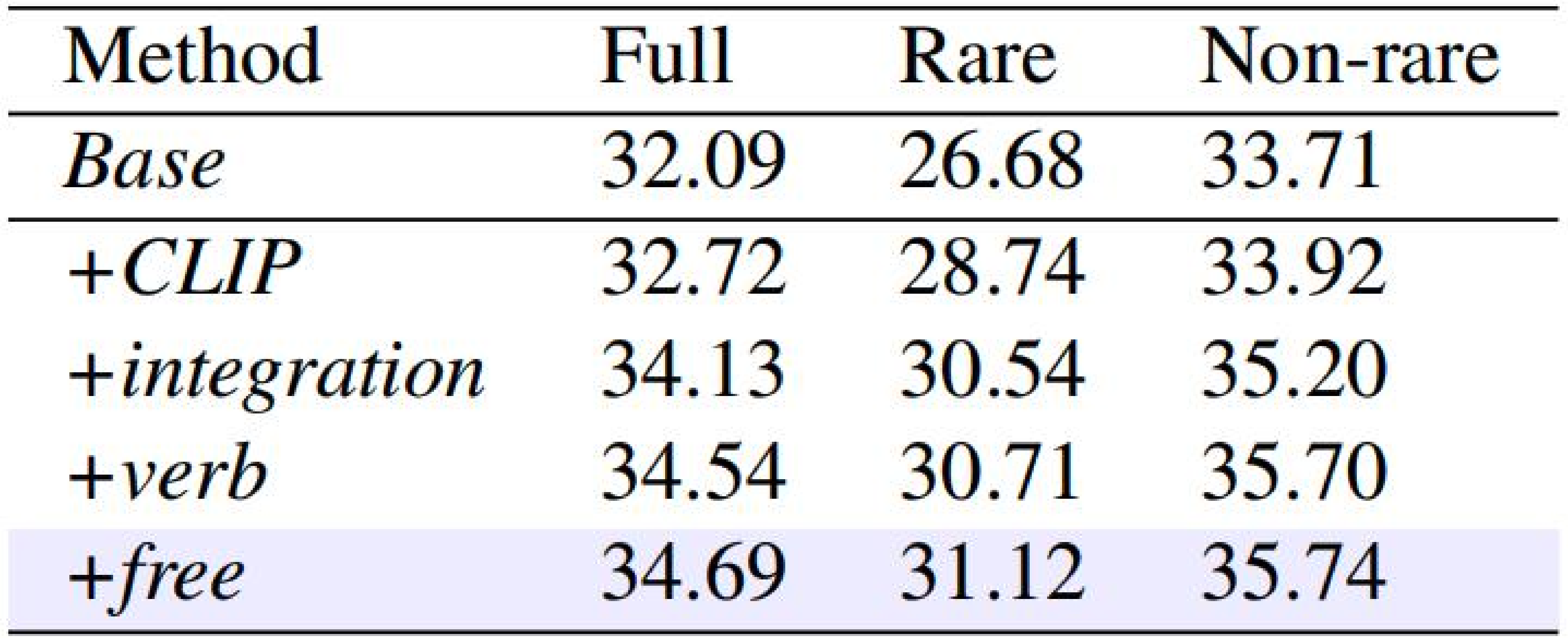
↑ 5.35%

↑ 54.14%

↑ 15.94%

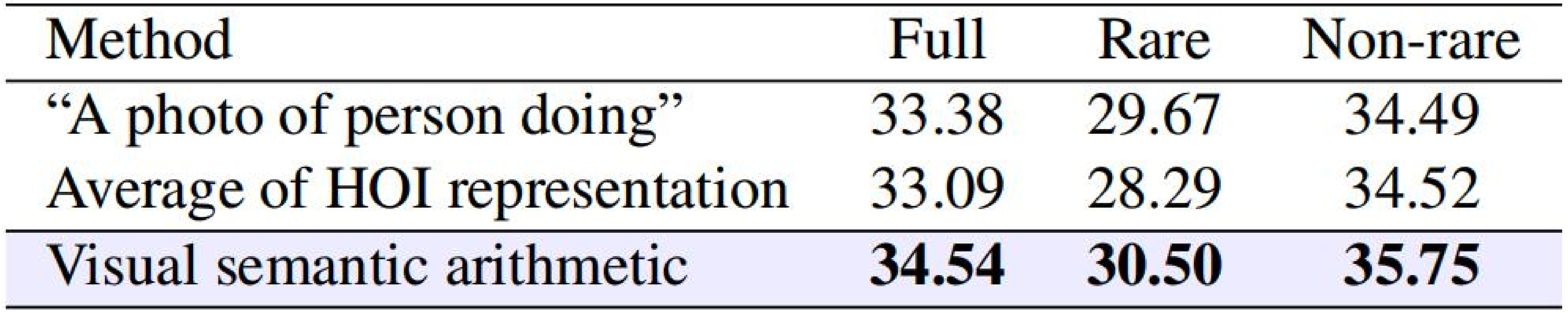
Ablation study

Network Architecture Design Albation



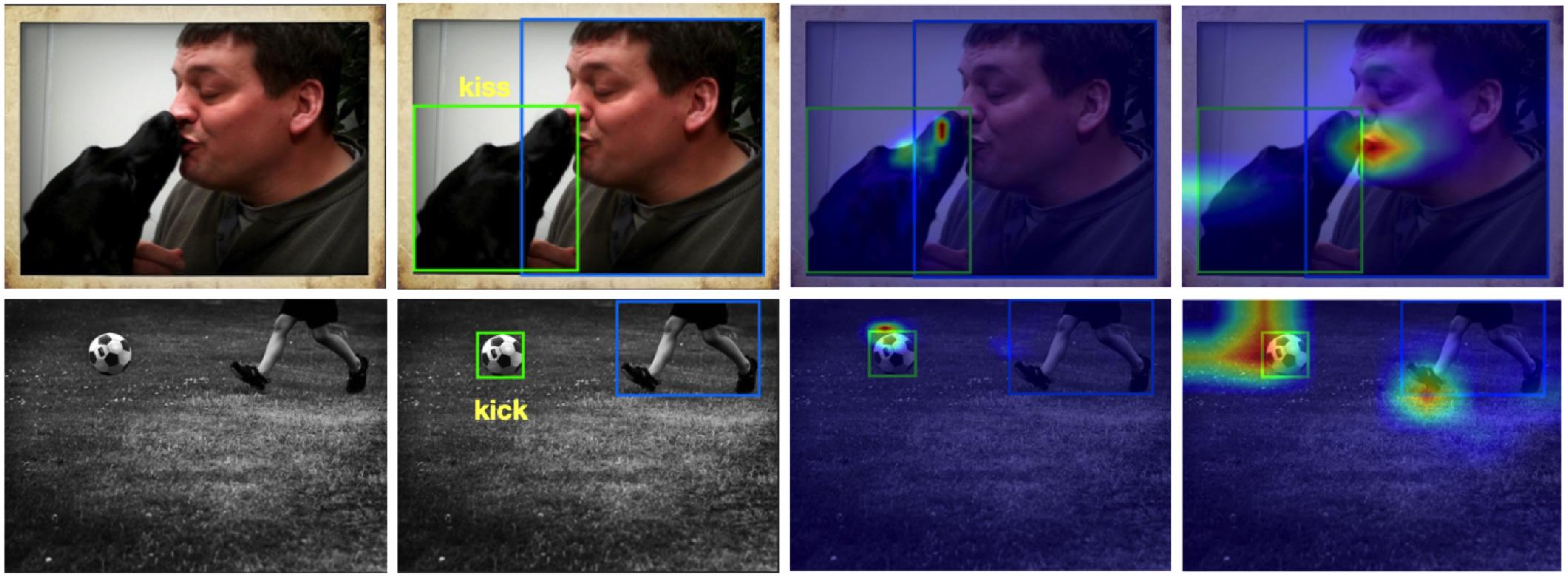
Method	Full	Rare	Non-rare
Base	32.09	26.68	33.71
+*CLIP*	32.72	28.74	33.92
+*integration*	34.13	30.54	35.20
+*verb*	34.54	30.71	35.70
+*free*	34.69	31.12	35.74

Verb Representation Extraction Albation



Method	Full	Rare	Non-rare
“A photo of person doing”	33.38	29.67	34.49
Average of HOI representation	33.09	28.29	34.52
Visual semantic arithmetic	**34.54**	**30.50**	**35.75**

Visualization



input Image

prediction result

localization attention maps

interaction attention maps

Thanks for listening!

For more information please refer to our paper and code

Paper & code