NEURAL INFORMATION PROCESSING SYSTEMS

Visual Diversity and Region-aware Prompt Learning for Zero-shot HOI detection

Chanhyeong Yang[1], Taehoon Song[2], Jihwan Park[1,]
Hyunwoo J. Kim[2]

[1]Korea University [2]KAIST

Task definition

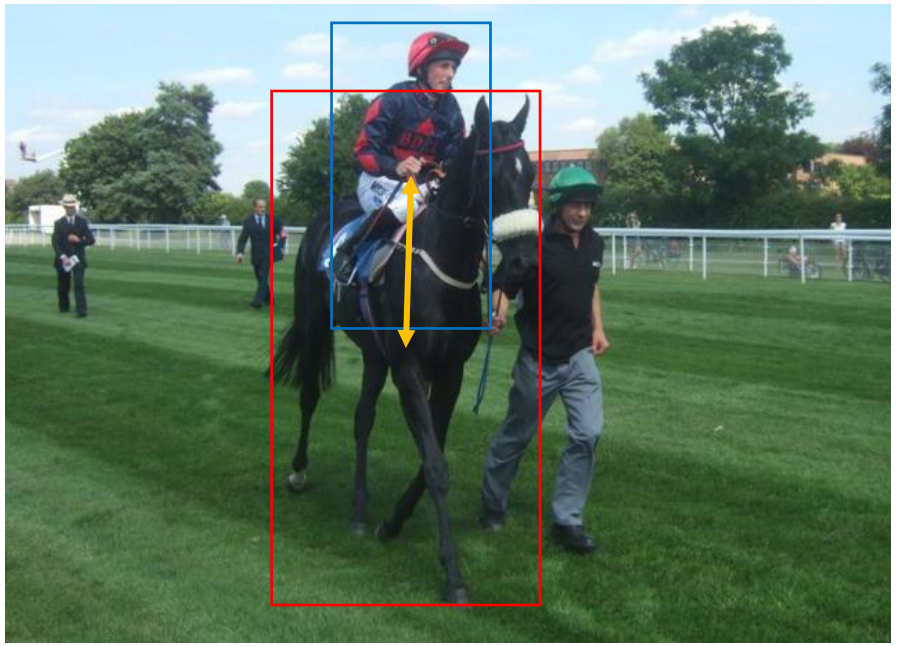
NEURAL INFORMATION PROCESSING SYSTEMS

Human-Object Interaction detection

- Localizing humans and objects in an image
- Recognizing their interactions

- **Zero-shot** HOI detection
  - Addressing **unseen combinations**
  - Ex) unseen combination/object/verb

Example of HOI detection

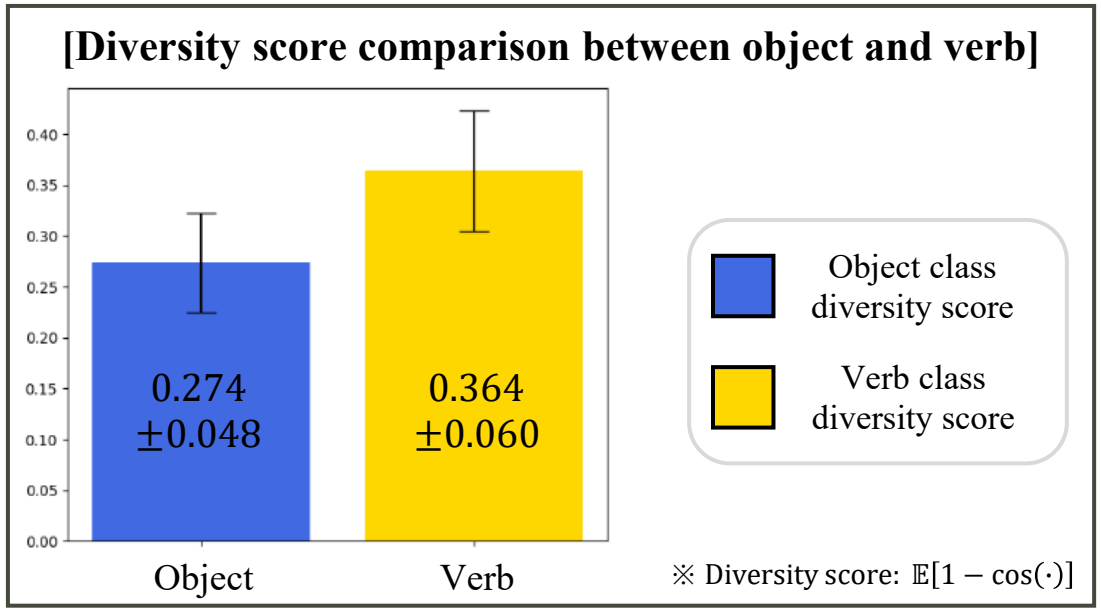


< Person, Riding, Horse >

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

1. Intra-class diversity



Diverse visual patterns

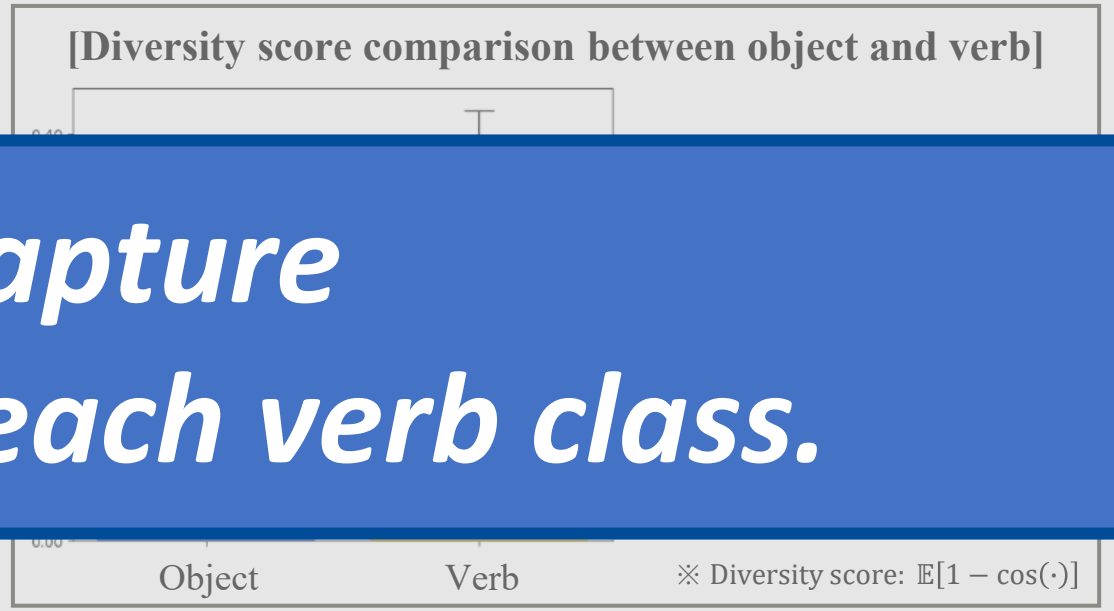
Verbs are more diverse than objects

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

1. Intra-class diversity

"Person **holding** a baseball glove"



Need to capture visual variance of each verb class.

Diverse visual patterns

Verbs are more diverse than objects

Motivation

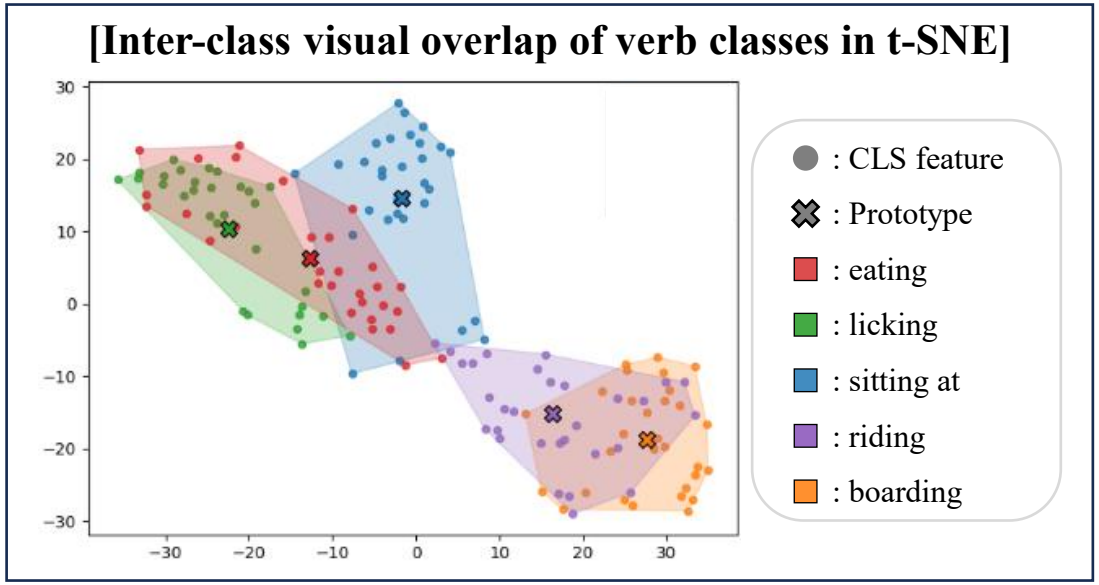
NEURAL INFORMATION PROCESSING SYSTEMS

2. Inter-class visual entanglement

"Person **eating** an object."

"Person **licking** an object."

"Person **sitting at** an object."



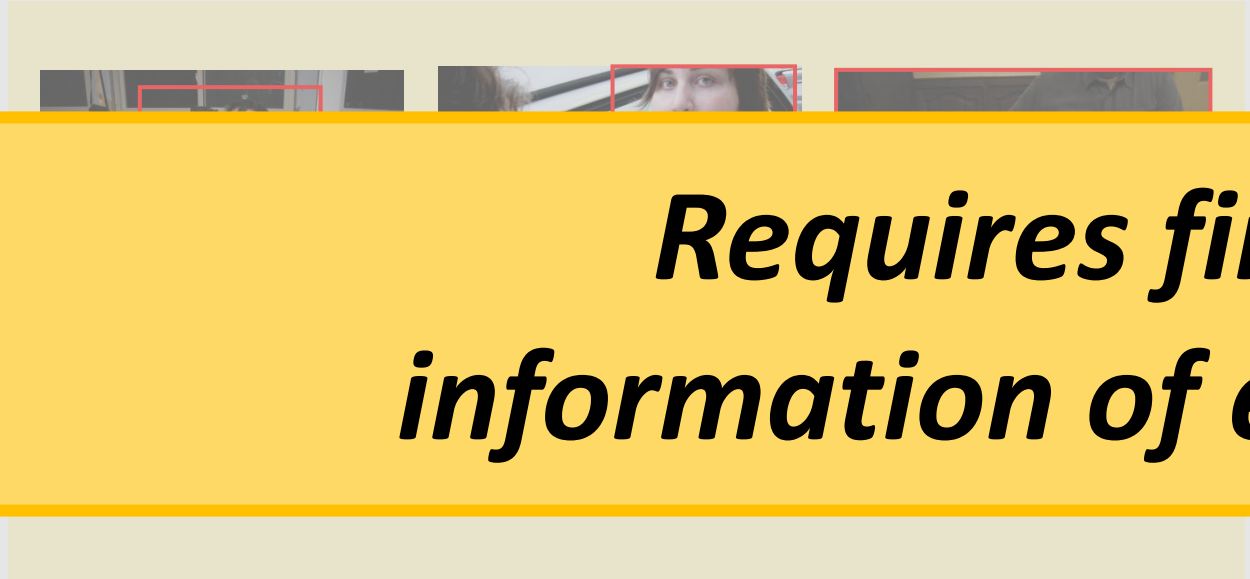
Similar, but different classes

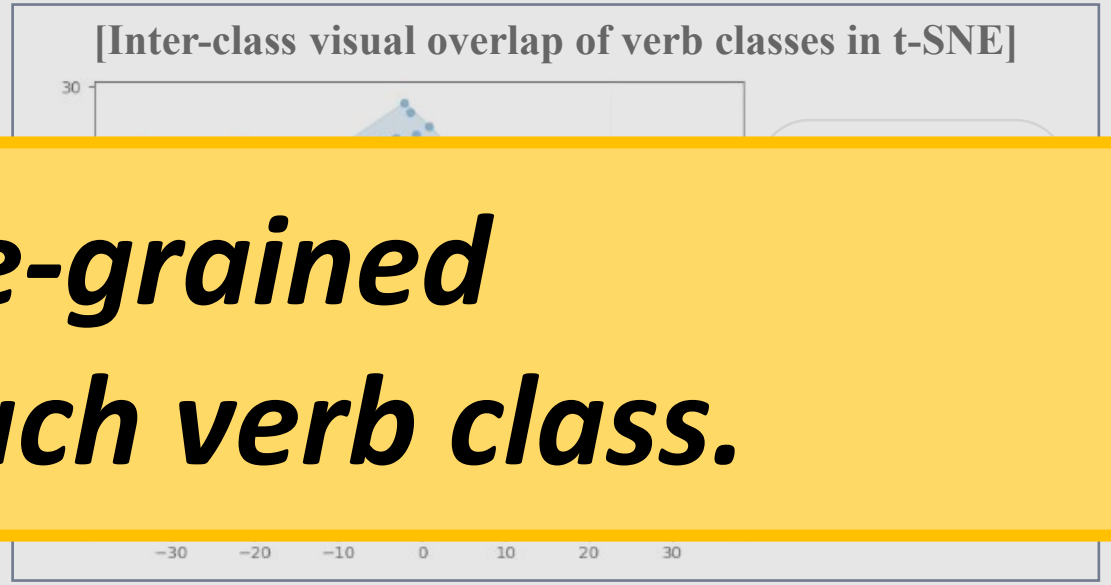
Overlap in visual embedding space

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

2. Inter-class visual entanglement





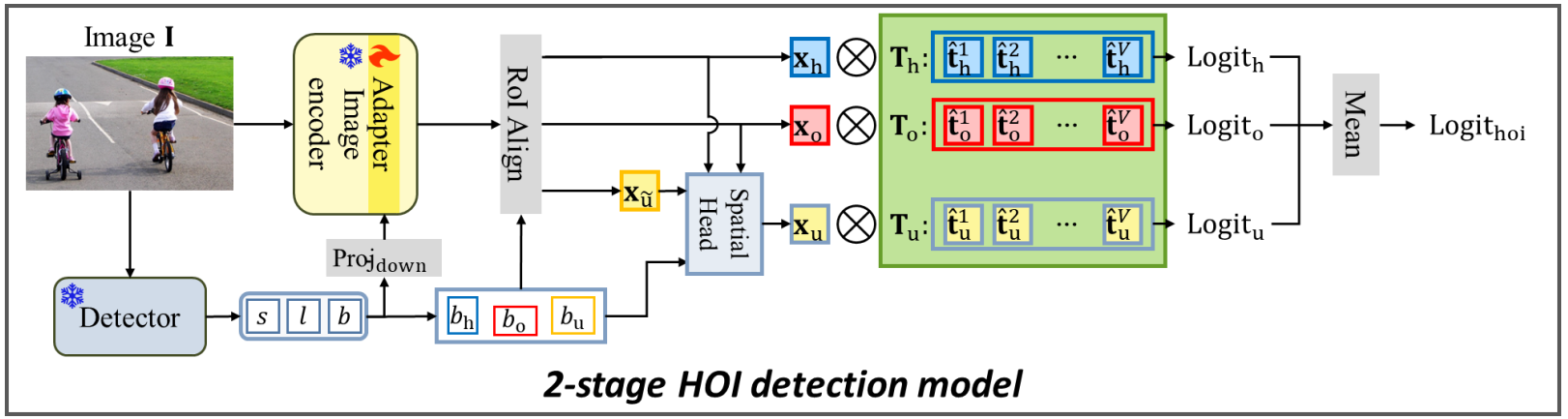
Requires fine-grained information of each verb class.

Similar, but different classes

Overlap in visual embedding space

Methods

Then, how can we handle both challenges?



Capture visual variance of each verb

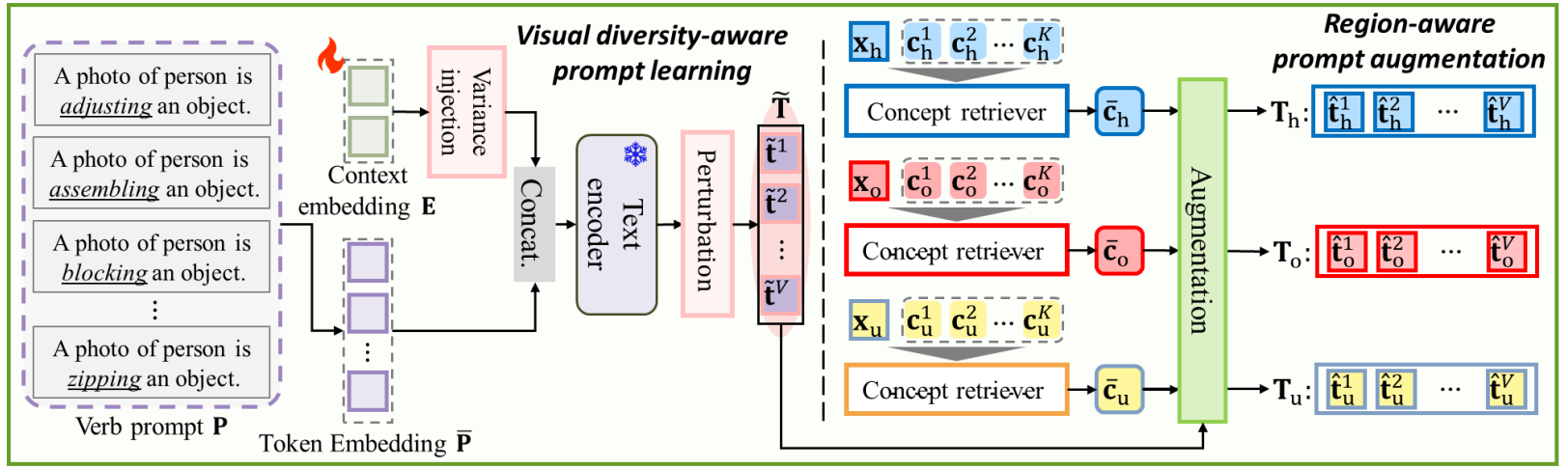
Leverage fine-grained concepts about verbs

Methods

NEURAL INFORMATION PROCESSING SYSTEMS

So, We need dual-module design: VDRP

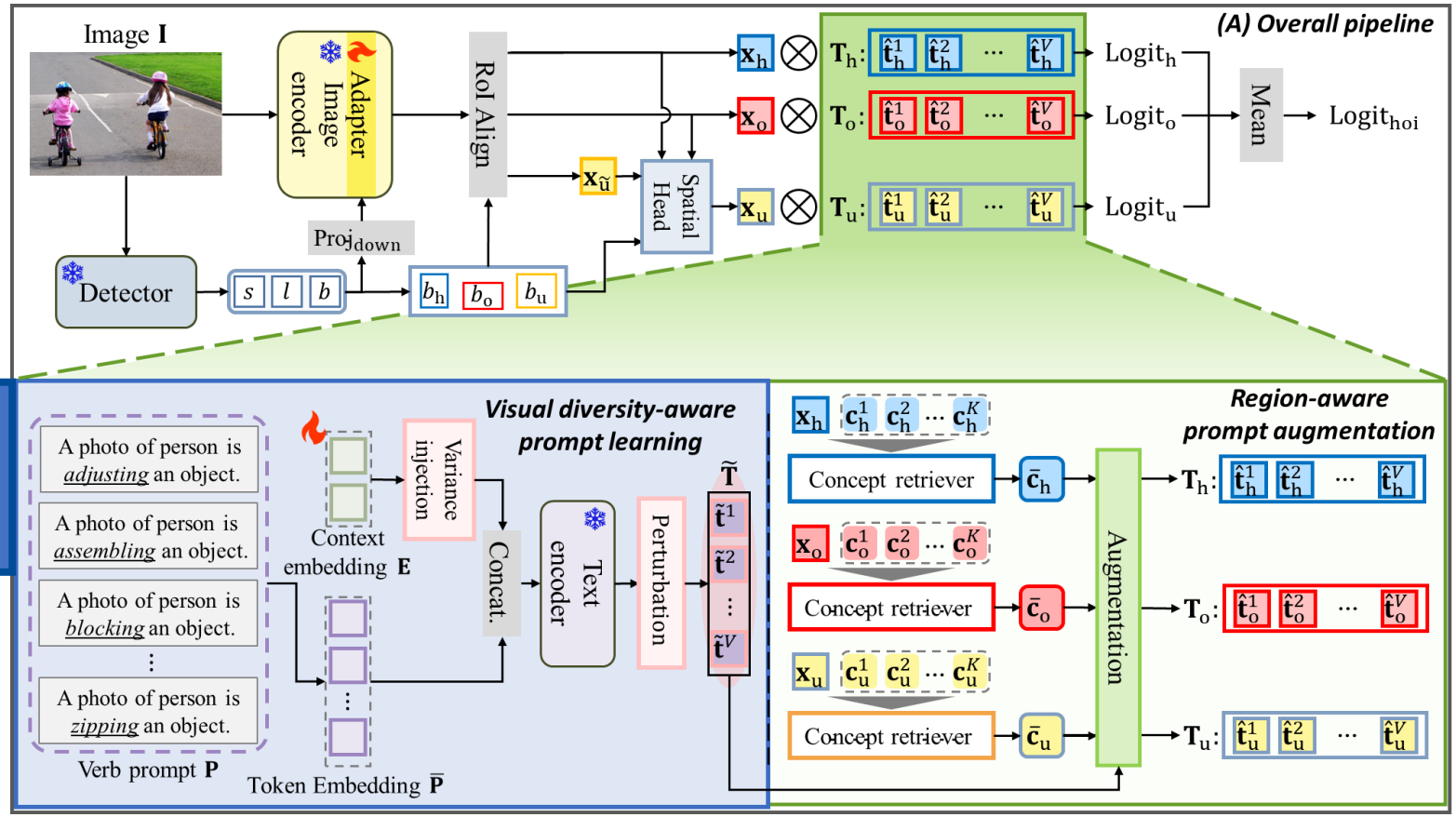
Visual Diversity and Region-aware Prompt Learning



Methods

1. We need VDP: Visual Diversity-aware Prompts

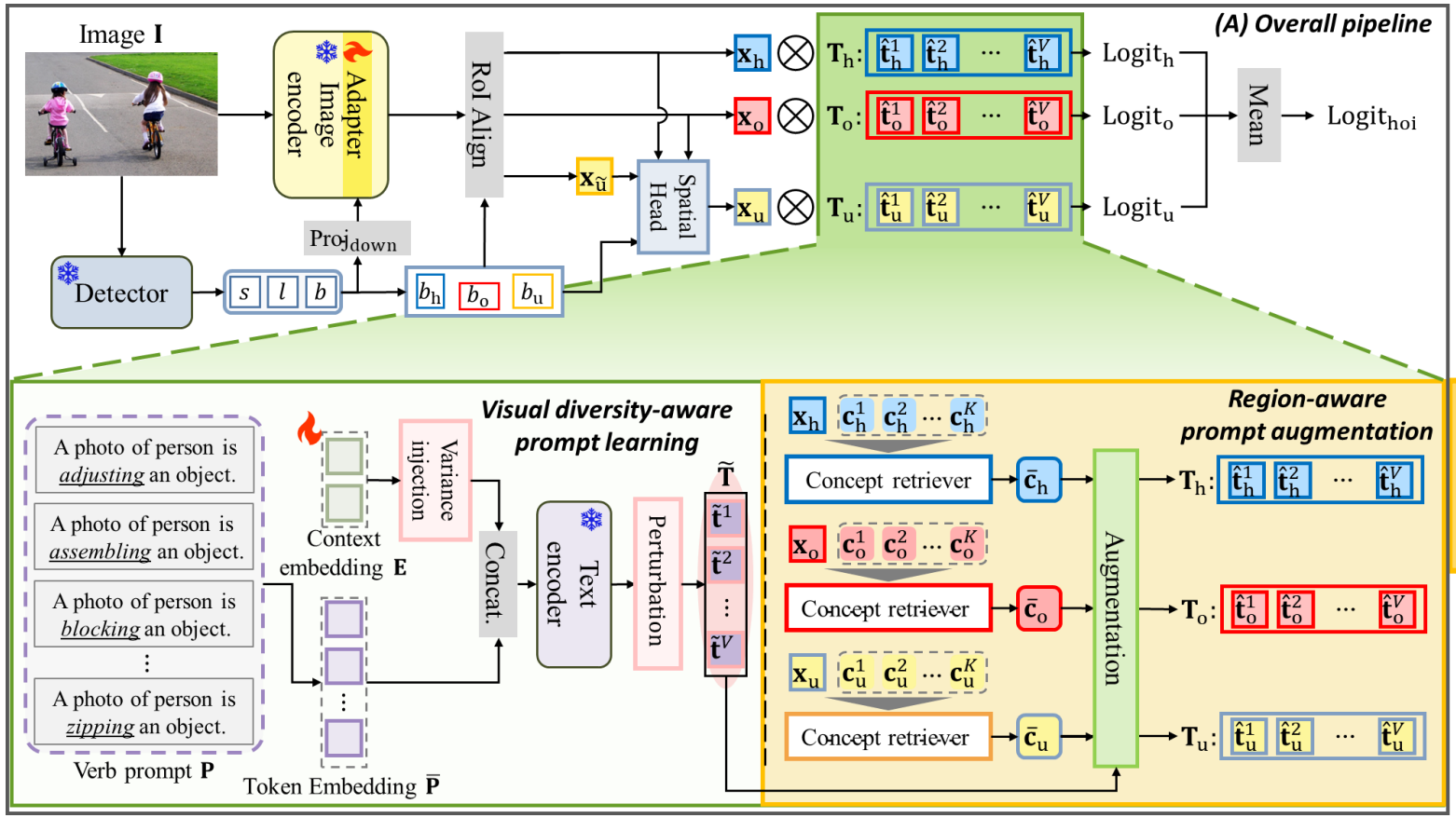
1. VDP: injection and reflect visual variance for each verb



Methods

NEURAL INFORMATION PROCESSING SYSTEMS

We need RAP: Region-Aware Prompts



2. RAP: augmentation with region-level concepts for each verb

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

HICO-DET: NF-UC / RF-UC

Method	Backbone	NF-UC				RF-UC			
		HM	Full	Unseen	Seen	HM	Full	Unseen	Seen
GEN-VLKT [20]	ResNet-50 + ViT-B	24.17	23.71	25.05	23.38	26.08	30.56	21.36	32.91
EoID [50]	ResNet-50	26.71	26.69	26.76	26.66	26.11	29.52	22.04	31.39
HOICLIP [24]	ResNet-50 + ViT-B	28.70	27.75	29.36	28.10	26.55	32.99	25.83	28.47
ADA-CM [9]	ResNet-50 + ViT-B	31.76	31.39	32.41	31.13	30.48	33.01	27.63	34.35
CLIP4HOI [23]	ResNet-50 + ViT-B	29.54	28.90	31.44	28.26	31.23	**34.08**	27.88	**35.48**
CMMP [21]	ResNet-50 + ViT-B	30.82	30.18	32.09	29.71	31.10	32.18	29.45	32.87
EZ-HOI [22]	ResNet-50 + ViT-B	31.76	31.17	33.66	30.55	31.18	33.13	29.02	34.15
Ours	ResNet-50 + ViT-B	**33.85**	**32.57**	**36.45**	**31.60**	**32.77**	33.78	**31.29**	34.41
UniHOI [10]	ResNet-50 + ViT-L	30.40	31.79	28.45	32.63	30.76	32.27	28.68	33.16
CMMP [21]	ResNet-50 + ViT-L	34.50	35.13	33.52	35.53	36.69	37.13	35.98	37.42
EZ-HOI [22]	ResNet-50 + ViT-L	35.38	34.84	36.33	34.47	35.73	36.73	34.24	37.35
Ours	ResNet-50 + ViT-L	**36.83**	**36.46**	**37.48**	**36.21**	**37.58**	**38.13**	**36.72**	**38.48**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

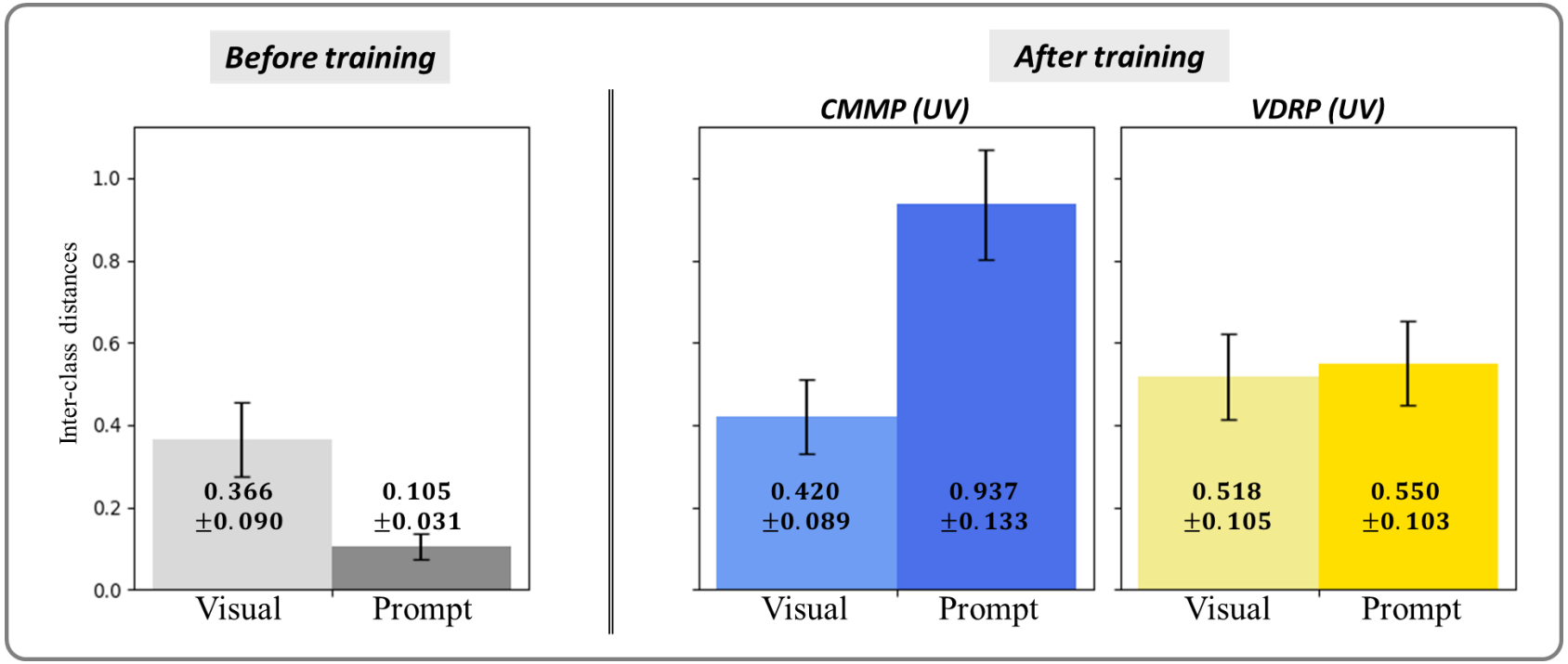
HICO-DET: Unseen Object / Unseen Verb

Method	Backbone	#TP	UO				UV			
			HM	Full	Unseen	Seen	HM	Full	Unseen	Seen
FCL [1]	ResNet-50	–	17.65	19.87	15.54	20.74	–	–	–	–
ATL [2]	ResNet-50	–	17.79	20.47	15.11	21.54	–	–	–	–
GEN-VLKT [20]	ResNet-50	42.05M	20.11	25.63	15.01	28.92	24.35	28.74	20.96	30.23
EoID [50]	ResNet-50	–	–	–	–	–	26.29	29.61	22.71	30.73
HOICLIP [24]	ResNet-50 + ViT-B	66.18M	20.32	28.53	16.30	30.99	27.72	31.09	24.30	32.19
CLIP4HOI [23]	ResNet-50 + ViT-B	56.70M	31.98	32.58	31.79	32.73	28.35	30.42	26.02	31.14
CMMP [21]	ResNet-50 + ViT-B	2.30M	32.44	31.59	33.76	31.15	29.23	31.84	26.23	32.75
EZ-HOI [22]	ResNet-50 + ViT-B	6.85M	32.14	32.27	33.28	32.06	29.09	32.32	25.10	33.49
Ours	ResNet-50 + ViT-B	4.50M	**34.41**	**33.39**	**36.13**	**32.84**	**29.80**	**32.73**	**26.69**	**33.72**
UniHOI [10]	ResNet-50 + ViT-L	52.3M	25.17	31.56	19.72	34.76	30.50	34.68	26.05	36.78
CMMP [21]	ResNet-50 + ViT-L	5.40M	37.83	36.74	**39.67**	36.15	33.75	36.38	30.84	37.28
EZ-HOI [22]	ResNet-50 + ViT-L	14.07M	37.06	36.38	38.17	36.02	32.84	36.84	28.82	38.15
Ours	ResNet-50 + ViT-L	10.29M	**38.41**	**37.81**	39.36	**37.50**	**34.31**	**37.18**	**31.16**	**38.16**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Qualitative results



- *CMMP causes modality mismatch with over-separated prompts*
- *VDRP maintains balanced alignment between the two modalities*

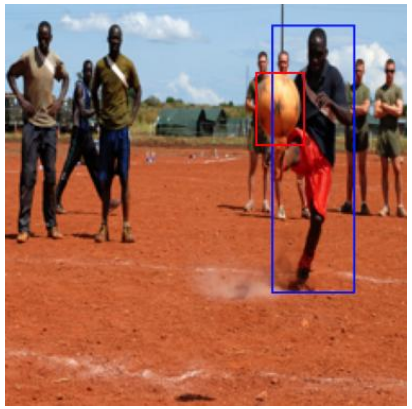
Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

Qualitative results

: Human region concepts : Object region concepts : Union region concepts : Human box : Object box

Retrieved concepts



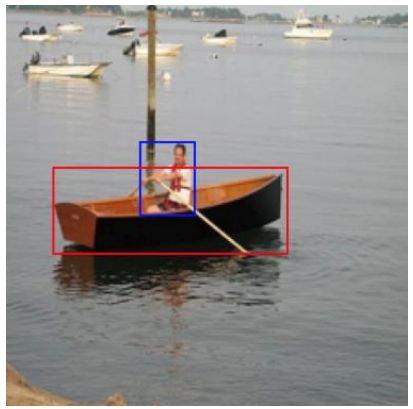
Kicking an object.

(0.27) Buttocks are tense and slightly lifted.
(0.27) Person is extending leg to make contact with object.
(0.25) Lower leg is straight and foot is flat on ground.
(0.21) Back is straight and shoulders are relaxed.

(0.50) Object appears to be in mid-air or suspended.
(0.35) Object's surface appears to be vibrating or trembling.
(0.07) Object's position changes or shifts.

(0.27) The object is propelled through the air.
(0.14) The person's foot is cocked back before making contact.
(0.13) The object is sent flying in a diagonal direction.
(0.13) The object is struck with a firm, direct kick.

Retrieved concepts



Rowing an object.

(0.59) Arms are extended and pulling on oars.
(0.20) Legs are bent and feet are planted firmly on the platform.
(0.14) Back is straight and facing forward.

(0.67) Rowing motion creates ripples in the water.
(0.20) Object is positioned at an angle in the water.
(0.13) Object is partially submerged in water.

(0.82) The person's legs are tucked in to maintain balance.
(0.07) The person's face is focused on the task at hand.
(0.05) The water is calm and reflective of the surrounding scenery.

▶ ***With Sparsemax, irrelevant concepts are zero-out***

Conclusions

- We propose **VDRP**, a dual-module prompt learning framework for zero-shot HOI detection.
- By combining **visual diversity-aware prompts** and **region-aware prompts,**
- **VDRP** addresses both **intra-class diversity** and **inter-class entanglement** in once.

NEURAL INFORMATION PROCESSING SYSTEMS

Paper

Github

KOREA UNIVERSITY

KAIST