CVPR
JUNE 17-21, 2024
SEATTLE, WA

Discovering Syntactic Interaction Clues for Human-Object Interaction Detection

Jinguo Luo[1], Weihong Ren[1,2], Weibo Jiang[1], Xi’ai Chen[2], Qiang Wang[3], Honghai Liu[1]

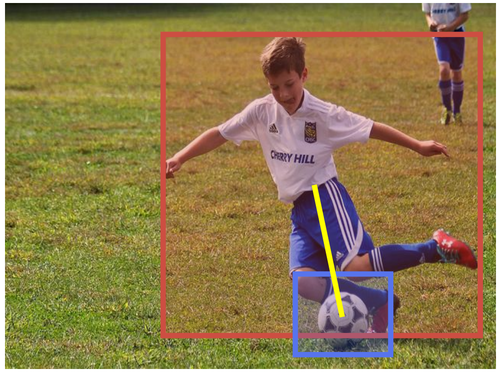
[1]Harbin Institute of Technology, [2]Chinese Academy of Science,
[3]Shenyang University

Speaker: ***Jinguo Luo***

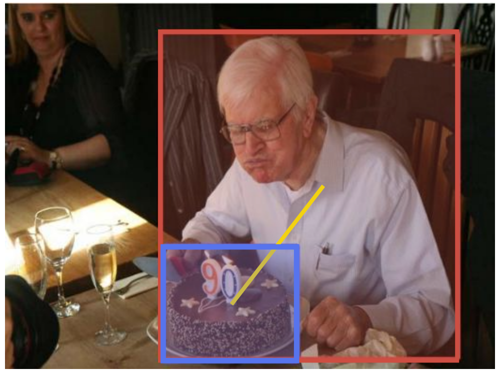
Preview

Definition

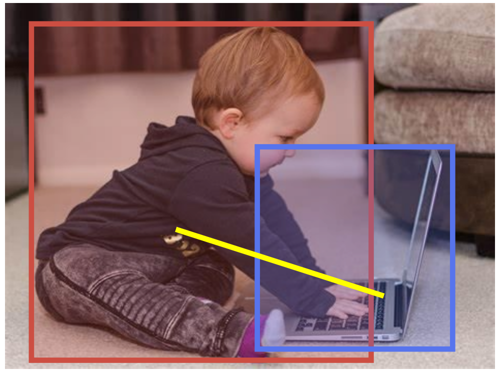
- Human-Object Interaction (abbreviated as **HOI**) aims to localize the **human** and **object** instance and inference the **action** between them;
- HOI should be expressed as a *<human, action, object>* triplet.



<human, kick, sportsball>



<human, cut, cake>

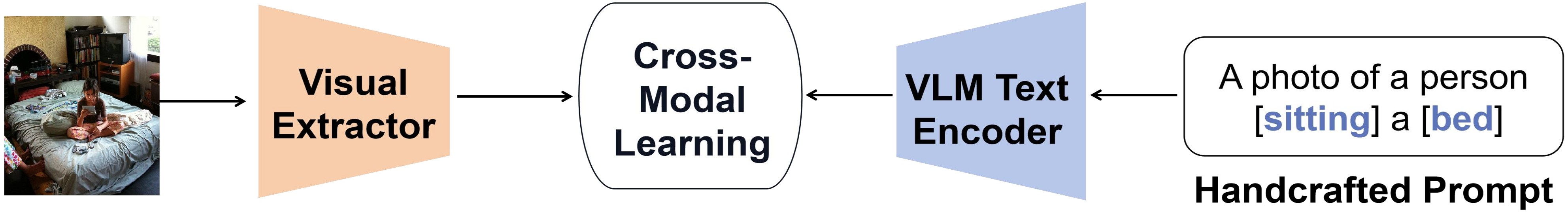


<human, work on, computer>

Preview

Limitation

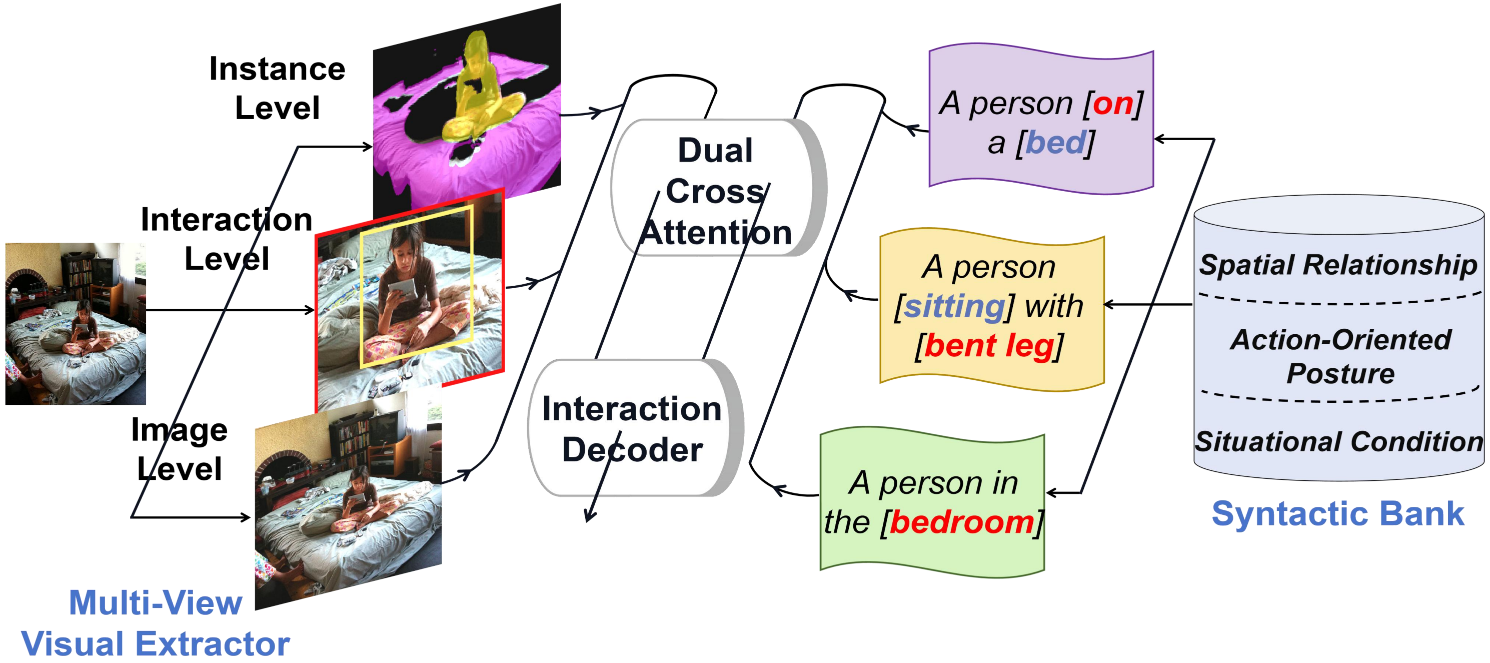
- Existing HOI detectors utilize text prompts to acquire textual prior knowledge.
- Tranditional methods: adopt a handcrafted template to acquire **action-specific** knowledge in vocabulary level. (E.g., “A photo of a person <action> a/an <object>”)



Preview

Improvement

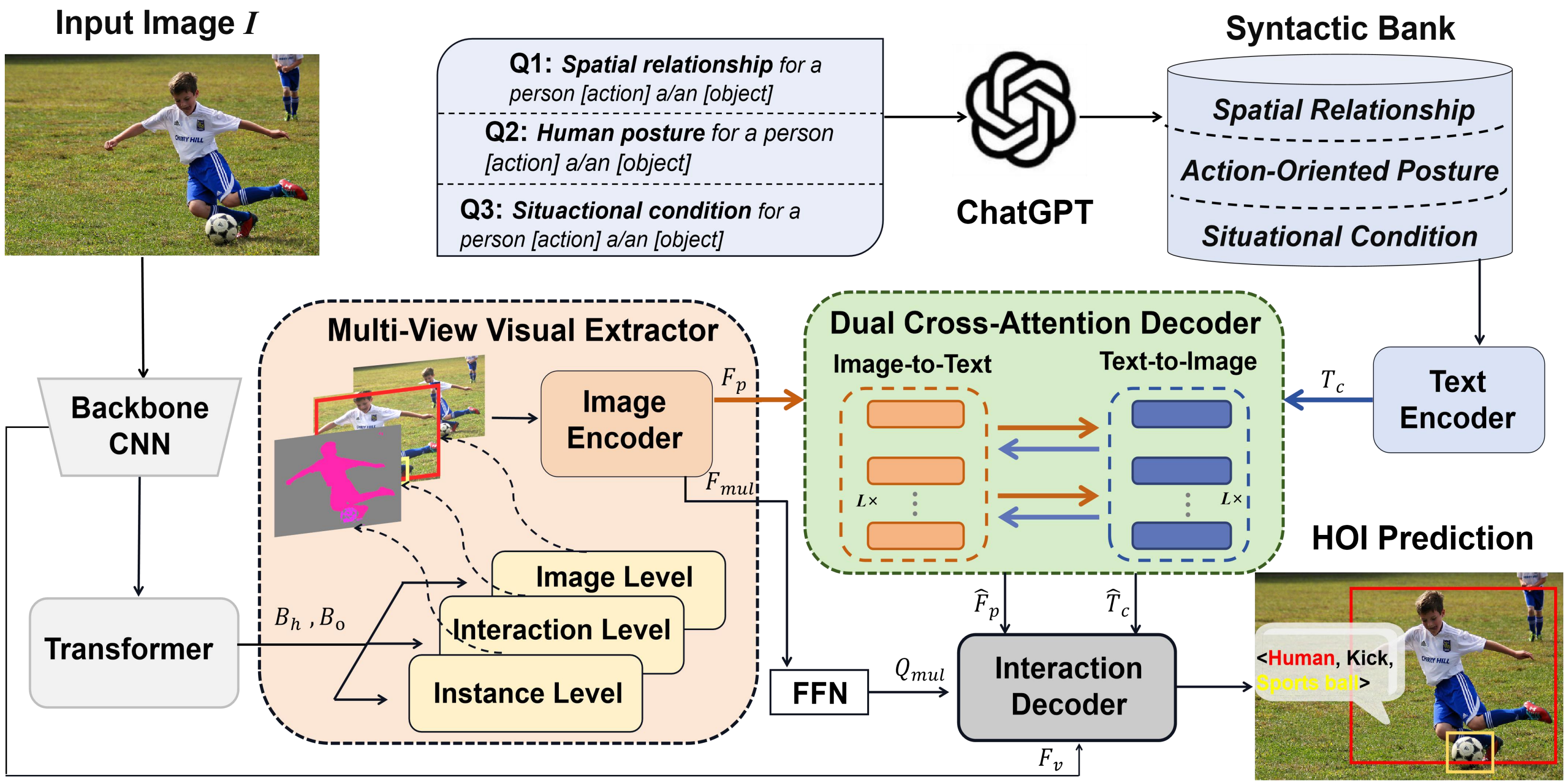
- Our SICHOI model: establishes a syntactic bank to acquire textual knowledge from three levels: **spatial relationship**, **action-oriented posture** and **situational condition**.



Method

Overall Architecture

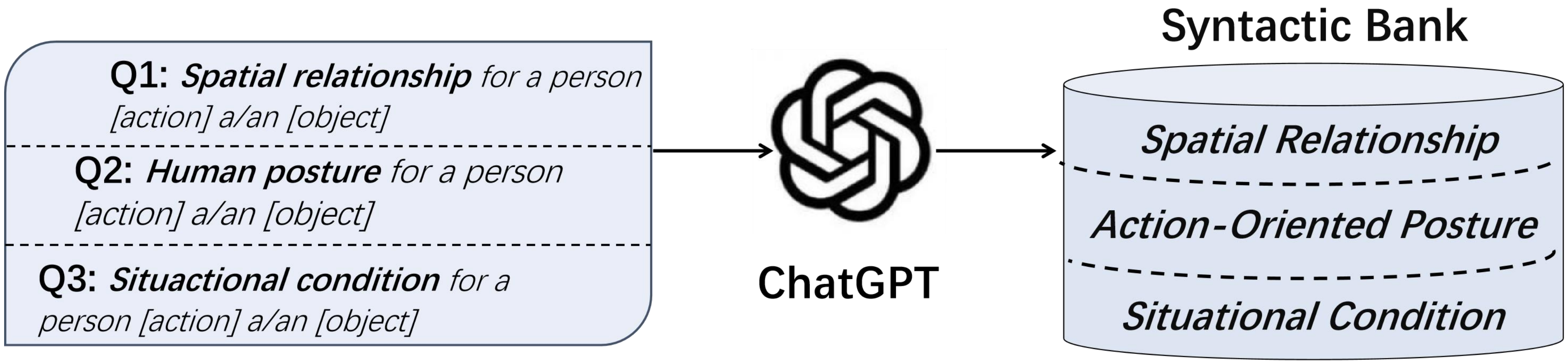
- Syntactic Bank
- Multi-View Visual Extractor
- Dual Cross-Attention Decoder



Method

Syntactic Interaction Bank (SIB)

- Q1: When a person [action] a [object], what is the spatial relationship between the person and the object?
- Q2: How to judgement whether a person [action] a [object] from human posture perspective?
- Q3: What is the situational condition for a person [action] a [object]?



Q1: ***Spatial relationship*** *for a person [action] a/an [object]*

Q2: ***Human posture*** *for a person [action] a/an [object]*

Q3: ***Situactional condition*** *for a person [action] a/an [object]*

ChatGPT

Syntactic Bank

Spatial Relationship

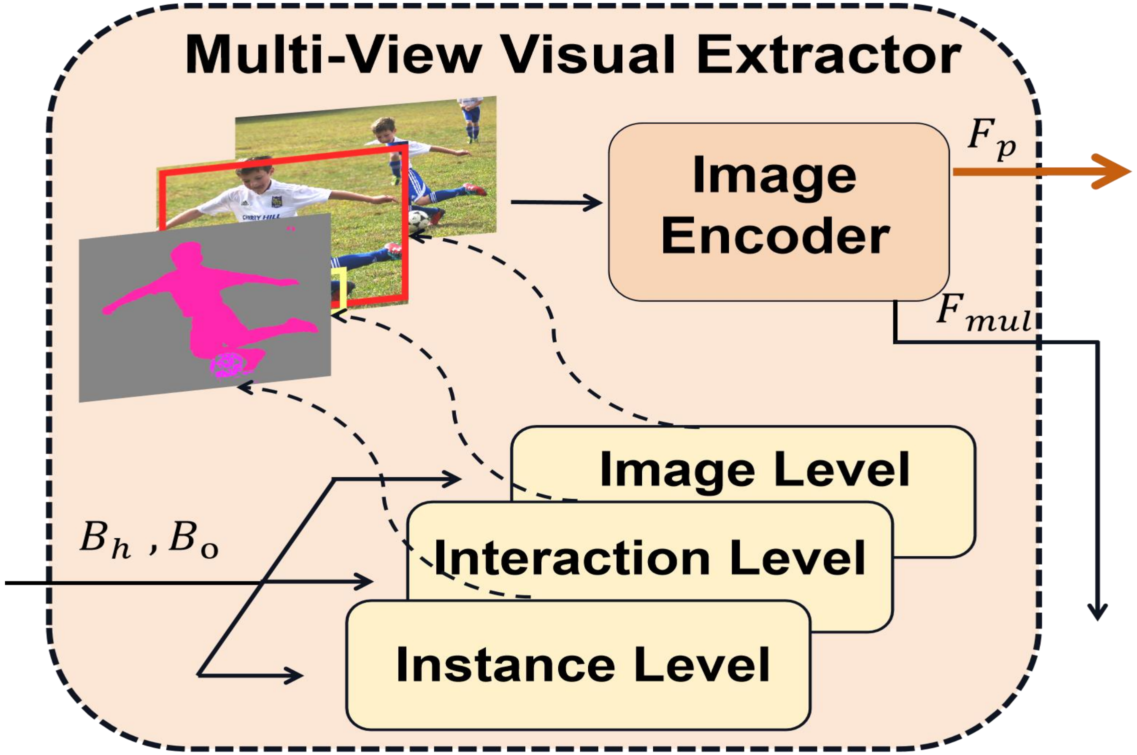
Action-Oriented Posture

Situational Condition

Method

Multi-View Visual Extractor (MVVE)

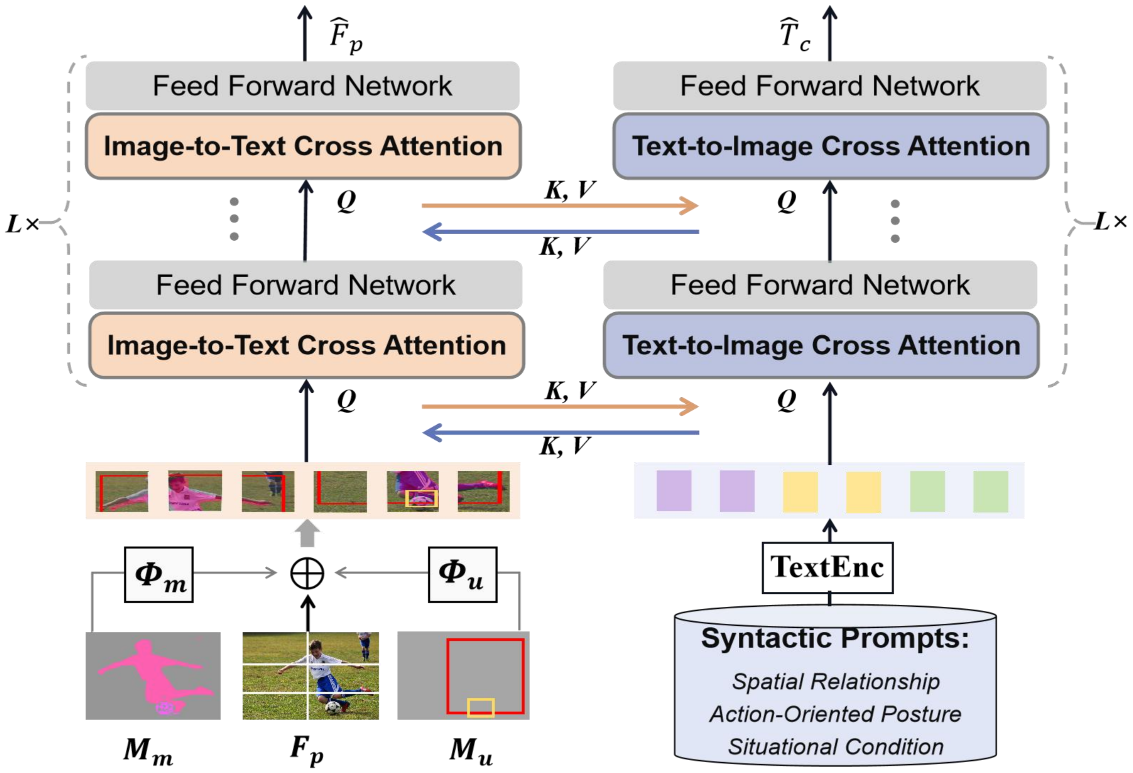
- MVVE aggregates visual features from *instance*, *interaction*, and *image* levels.



Method

Dual Cross-Attention Decoder (DCAD)

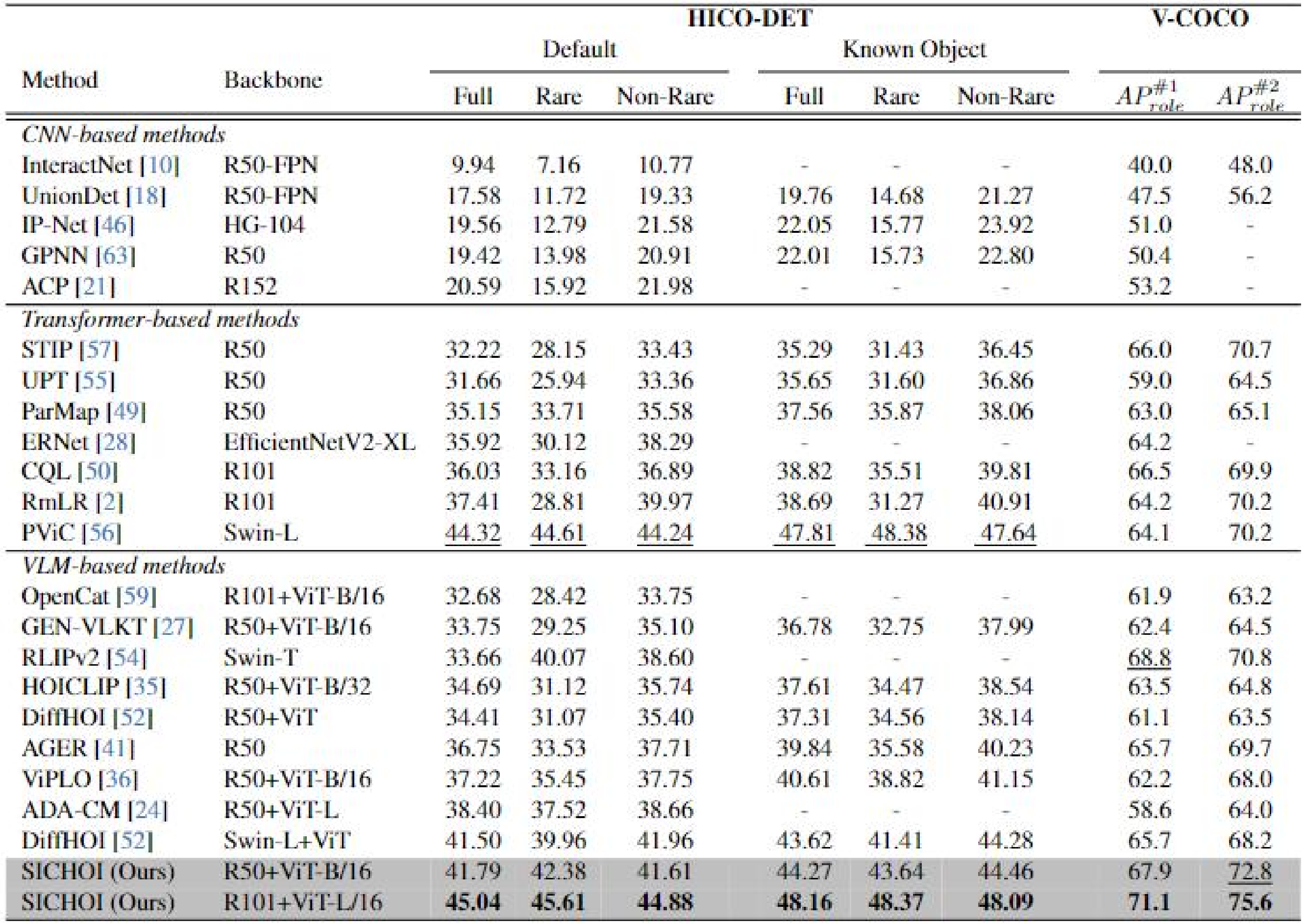
- DCAD performs context propagation between text knowledge and visual features.
- In text-to-image cross attention: text knowledge ->query, visual features-> key&value
- In image-to-text cross attention: visual features ->query, text knowledge-> key&value



Qualitative Results

Comparsion on HICO-DET and V-COCO dataset

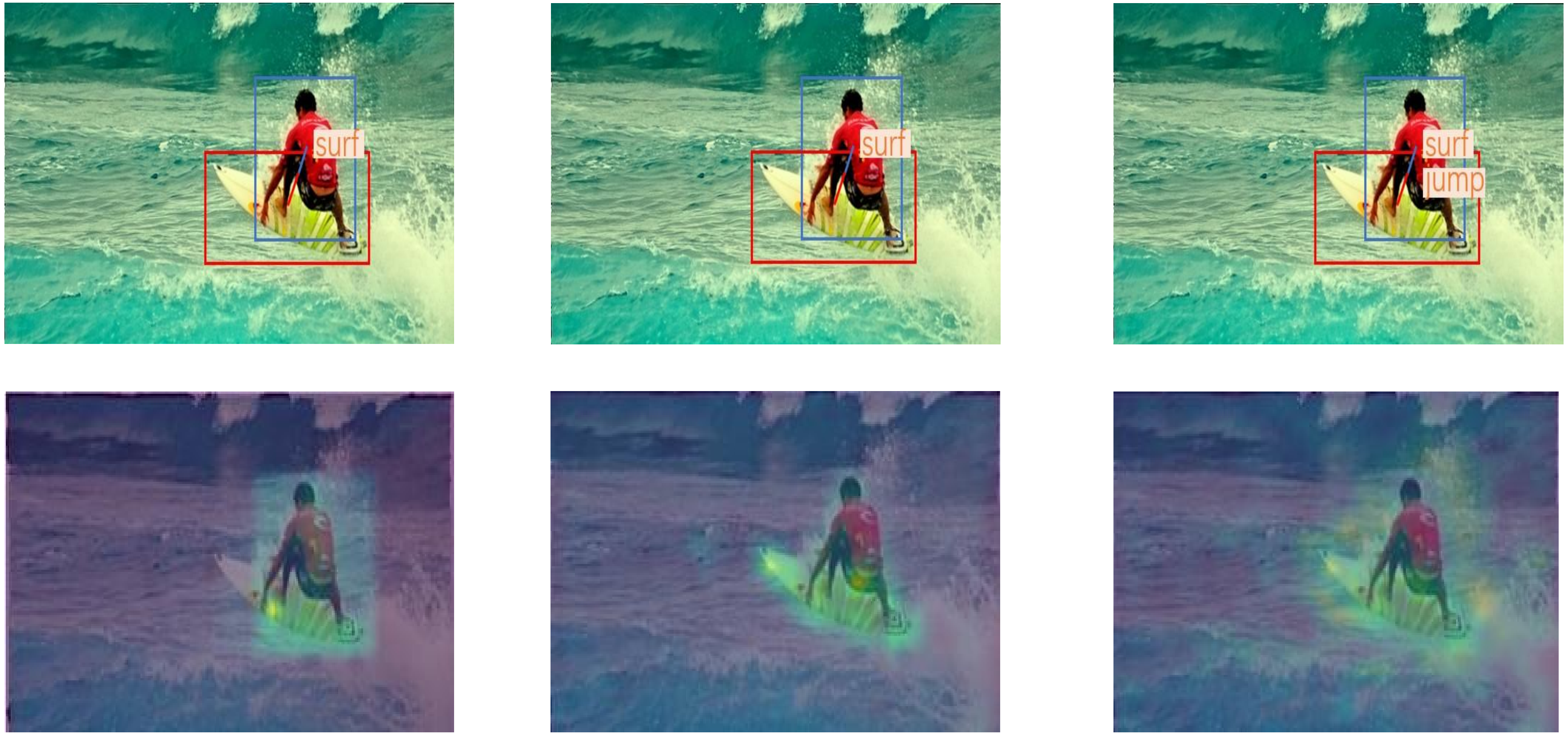
HICO-DET: ↑0.**72map** V-COCO: ↑2.3**map**



Method	Backbone	HICO-DET						V-COCO	
		Default			Known Object				
		Full	Rare	Non-Rare	Full	Rare	Non-Rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
CNN-based methods									
InteractNet [10]	R50-FPN	9.94	7.16	10.77	-	-	-	40.0	48.0
UnionDet [18]	R50-FPN	17.58	11.72	19.33	19.76	14.68	21.27	47.5	56.2
IP-Net [46]	HG-104	19.56	12.79	21.58	22.05	15.77	23.92	51.0	-
GPNN [63]	R50	19.42	13.98	20.91	22.01	15.73	22.80	50.4	-
ACP [21]	R152	20.59	15.92	21.98	-	-	-	53.2	-
Transformer-based methods									
STIP [57]	R50	32.22	28.15	33.43	35.29	31.43	36.45	66.0	70.7
UPT [55]	R50	31.66	25.94	33.36	35.65	31.60	36.86	59.0	64.5
ParMap [49]	R50	35.15	33.71	35.58	37.56	35.87	38.06	63.0	65.1
ERNet [28]	EfficientNetV2-XL	35.92	30.12	38.29	-	-	-	64.2	-
CQL [50]	R101	36.03	33.16	36.89	38.82	35.51	39.81	66.5	69.9
RmLR [2]	R101	37.41	28.81	39.97	38.69	31.27	40.91	64.2	70.2
PViC [56]	Swin-L	44.32	44.61	44.24	47.81	48.38	47.64	64.1	70.2
VLM-based methods									
OpenCat [59]	R101+ViT-B/16	32.68	28.42	33.75	-	-	-	61.9	63.2
GEN-VLKT [27]	R50+ViT-B/16	33.75	29.25	35.10	36.78	32.75	37.99	62.4	64.5
RLIPv2 [54]	Swin-T	33.66	40.07	38.60	-	-	-	68.8	70.8
HOICLIP [35]	R50+ViT-B/32	34.69	31.12	35.74	37.61	34.47	38.54	63.5	64.8
DiffHOI [52]	R50+ViT	34.41	31.07	35.40	37.31	34.56	38.14	61.1	63.5
AGER [41]	R50	36.75	33.53	37.71	39.84	35.58	40.23	65.7	69.7
ViPLO [36]	R50+ViT-B/16	37.22	35.45	37.75	40.61	38.82	41.15	62.2	68.0
ADA-CM [24]	R50+ViT-L	38.40	37.52	38.66	-	-	-	58.6	64.0
DiffHOI [52]	Swin-L+ViT	41.50	39.96	41.96	43.62	41.41	44.28	65.7	68.2
SICHOI (Ours)	R50+ViT-B/16	41.79	42.38	41.61	44.27	43.64	44.46	67.9	72.8
SICHOI (Ours)	R101+ViT-L/16	**45.04**	**45.61**	**44.88**	**48.16**	**48.37**	**48.09**	**71.1**	**75.6**

Quantitative Results

HOI detections and attention maps



(a) Base (b) Base+SP (c) Base+SIB