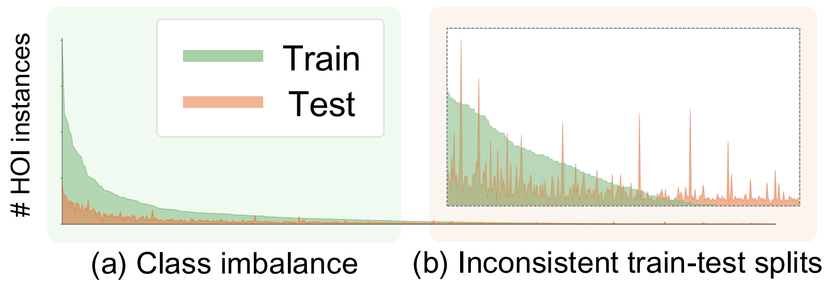
B-RIGHT: Benchmark Re-evaluation for Integrity in Generalized Human-Object Interaction Testing

Yoojin Jang[1*], Junsu Kim[1*], Hayeon Kim[1] , Eun-ki Lee[2], Eun-sol Kim[2], Seungryul Baek[1] and Jaejun Yoo[1†]

[1]Ulsan National Institute of Science and Technology, [2]Hanyang University

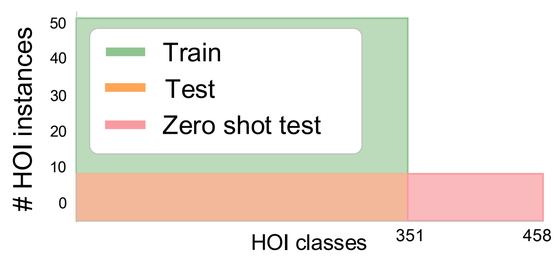
Do standard HOI benchmarks truly reflect HOI detector performance? No!

Limitations in HOI benchmarks (e.g., HICO-DET)

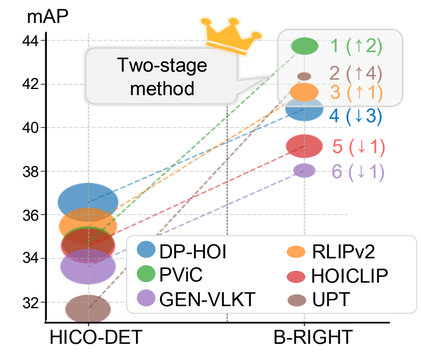


(a) Class imbalance (b) Inconsistent train-test splits

Ours Balanced HOI benchmarks : B-RIGHT



Re-train and re-evaluate HOI detectors



Imbalanced benchmark
- ☞ **High** class-score variance
- ☞ **Unfair** class-wise evaluation
- ☞ **Unreliable** metrics

Balanced benchmark
- ☞ **Low** class-score variance
- ☞ **Fair** class-wise evaluation
- ☞ **Reliable** metrics

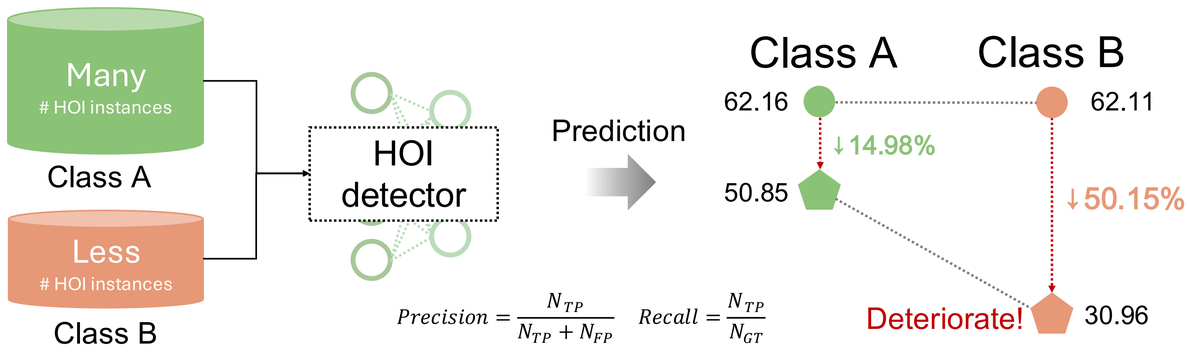
Balanced benchmark reveals what **imbalanced** benchmark hides:
- ☞ **Rankings shift**
- ☞ **Two-stage** models **win**

Motivation

Class-imbalanced train set → Overfitting to frequent classes
Class-imbalanced test set → Unreliable and inconsistent evaluation scores

The impact of **inconsistent train-test splits in certain classes**

When **flipping 1 TP to 1 FP**, the change of class AP is:

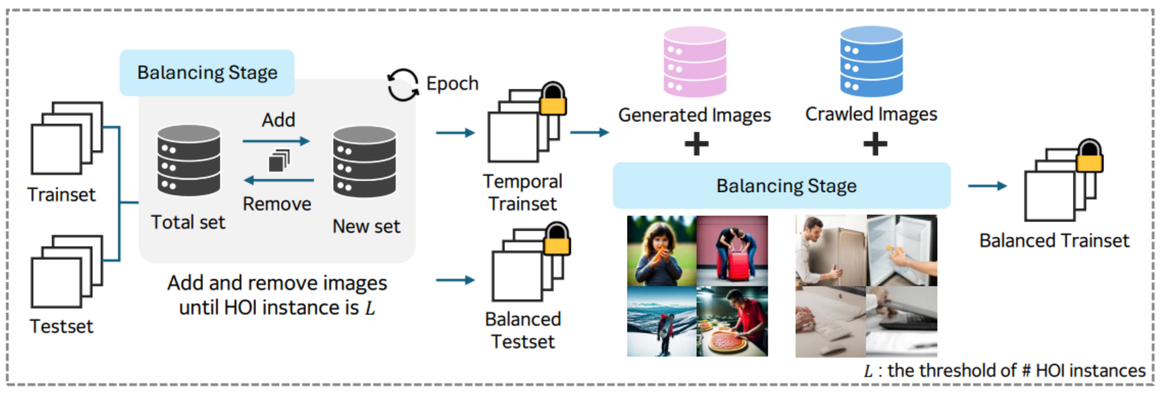


☞ **Amplifying** the impact of **each prediction on overall performance metrics**
☞ Leading the model to **inflated** or **deflated scores by chance**

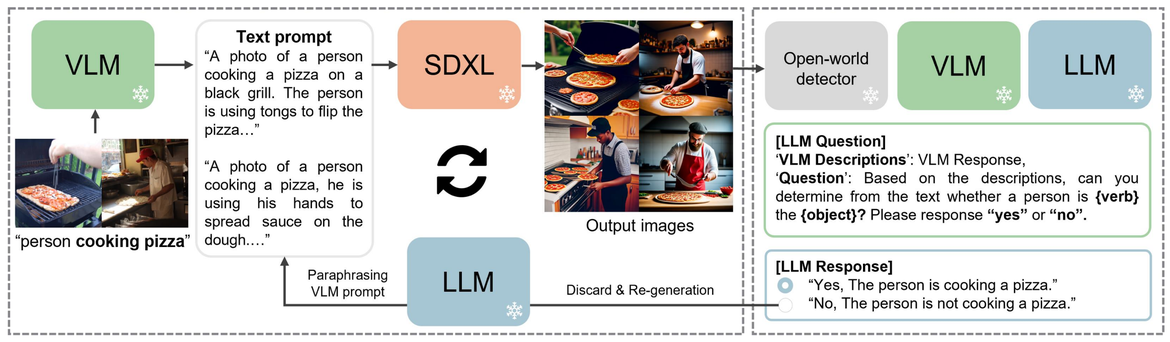
→ **A balanced dataset** is essential for fair and robust assessment of HOI models beyond dataset-induced biases

How we collected B-RIGHT?

(a) Balanced Algorithm



(b) Generation and Filtering Process



Comprehensive Analysis

(1) Ranking comparison of HOI detectors on HICO-DET and B-RIGHT

Model name	Backbone	Pretraining datasets	Auxiliary
One-stage models			
DP-HOI	DETR	HAA500, K700, F30K, VG, COCO, O365	CLIP
HOICLIP	DETR	COCO	CLIP
GEN-VLKT	DETR	COCO	CLIP
CDN	DETR	COCO	No
CQL	DETR	COCO	No
QPIC	DETR	COCO	No
Two-stage models			
RLIPv2	DAB-DDETR	VG, COCO, O365	No
PViC	DETR	COCO, HICO-DET†	No
UPT	DETR	COCO, HICO-DET†	No

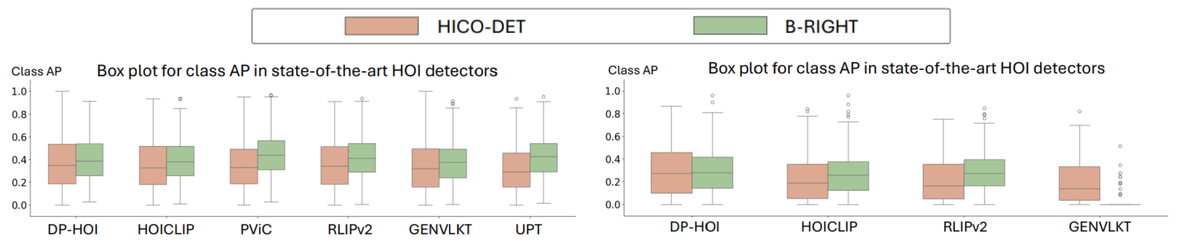
(a) HOI detectors

		Default				Zero-shot	
		HICO-DET		B-RIGHT		HICO-DET (RF-UC)	B-RIGHT
Model	Architecture	mAP	Rank	mAP	Rank	mAP	mAP
DP-HOI	one stage	36.56	1	40.85	4 (↓ 3)	30.49	31.81
RLIPv2	two stage	35.46	2	41.61	3 (↓ 1)	21.45	29.97
PViC	two stage	34.69	3	43.73	1 (↑ 2)	–	37.07
HOICLIP	one stage	34.56	4	39.14	5 (↓ 1)	25.53	28.36
GEN-VLKT	one stage	33.61	5	38.02	6 (↓ 1)	21.36	4.38
UPT	two stage	31.65	6	42.34	2 (↑ 4)	–	35.95
CQL	one stage	31.58	7	33.59	9 (↓ 2)	–	24.68
CDN	one stage	31.36	8	35.24	7 (↑ 1)	–	27.05
QPIC	one stage	29.11	9	34.00	8 (↑ 1)	–	24.05

(b) Performance comparison of models trained and evaluated on HICO-DET and B-RIGHT

☞ Two-stage models rank higher than one-stage models in both default and zero-shot settings.
☞ Using pre-training, foundation models or DAB-DDETR detection backbone alone does not guarantee consistent performance rankings in both default and zero-shot settings.

(2) Variance visualization in class-wise AP



(a) Default (b) Zero shot

☞ Balanced datasets reduce class-wise variance AP and enable more reliable evaluations.

(3) Coefficient of variation (CV) of class-wise AP under backbone scales

		HICO-DET		B-RIGHT	
	Backbone	mAP	CV	mAP	CV
RLIPv2	Resnet50	35.46	0.5984	41.61	**0.4496**
RLIPv2	Swin Large	45.18	0.4991	50.19	**0.3938**
GEN-VLKT-S	Resnet50	33.61	0.6232	38.02	**0.4831**
GEN-VLKT-M	Resnet101	34.57	0.6010	38.66	**0.4804**
GEN-VLKT-L	Resnet101	34.82	0.5976	39.37	**0.4677**

☞ Two-stage models exhibit lower variance in class-wise performance than one-stage models.
☞ Larger feature backbones not only improve overall mAP but also consistently reduce CV.

(4) Verification of utilizing synthetic data in train set

☞ Synthetic images improve performance without bias, confirming their reliability as training data.