Supplementary Material for Open-Category Human-Object Interaction Pre-training via Language Modeling Framework

Sipeng Zheng
Renmin University of China
zhengsipeng@ruc.edu.cn

Boshen Xu
UESTC
xuboshen.uestc@gmail.com

Qin Jin*
Renmin University of China
qjin@ruc.edu.cn

We present more details about our model implementation in Sec 1, and additional data analysis about our pre-training data in Sec 2. Additional ablation studies on our proxy tasks and model architecture can be found in Sec 3. Finally, Sec 5 shows some qualitative examples that demonstrate the effectiveness of our proposed OpenCat.

1. Additional Implementation Details

During training, we normalize the input images before feeding them into the model. For the masked language prediction (MLP) task, we ensure that at least one human-object-interaction (HOI) triplet is masked during pre-training. However, we observe an imbalance in the distribution of pre-training HOI classes. To address this issue, we implement a weighted scheme that assigns a higher masking probability to classes with fewer training samples. Similarly, for the human-object relation prediction (HRP) task, we give a larger weight to the loss of tail classes. Regarding the human-object patch jigsaw (HPJ) task, we shuffle a patch with a 30% probability, and we limit the maximum number of shuffled patches per image to $N_s^* = 180$.

To match synonymous HOI triplets with groundtruth, we use WordNet and follow these steps: (1) we calculate the semantic similarity between each predicted HOI triplet and groundtruth HOI class; (2) we apply a threshold to retain the synonymous triplets that match with the groundtruth; (3) we manually check these matched synonymous triplets to confirm their synonymity. Note that we only use WordNet for evaluation, and it does not affect the HOI prediction results or lead to any unreasonable improvements.

2. Additional Analysis on Pre-training Data

As described in our main paper, we collect 754,001 images with 2,516 relations and 9,731 nouns. In Figure 1, we split these categories into 4 subsets based on the number of training instances in each category, i.e. > 100, $50 \sim 100$, $20 \sim 50$ and < 20. Over 30% relation and noun categories contain more than 20 instances, and around 15% of them contain over 100 instances. Additionally, we present the distribution of the top-150 relations and nouns in Figure 2. Some examples of selected relation and noun categories are provided in Table 1. The statistics reveal that our collected pre-training data includes plenty of samples for various rare classes. To address the inevitable imbalance issue, we adopt a weighted training scheme during pre-training in Sec 1.

Table 1. Examples of selected relation and noun categories for HOI pre-training.

relation categories						
dive in	hit	devour	squat	gaze	seize	prune
climb	install	fry	wrestle	disembark	dodge	hunt
rinse	trek	cry	insert	bury	hitch	wax
wrangle	pluck	pinch	kneel	urge	suck	leap
wield	cuddle	sprinkle	smash	jog	jog	caress
chew	mount	bathe	sniff	grip	grasp	...
noun categories						
flowers	garden	racket	beard	necklace	fence	tent
bat	uniform	sweater	window	face	sunset	match
desk	jeans	platform	tablet	socks	mirror	blazer
boots	vest	rock	paper	stairs	bridge	beer
gun	portrait	apron	scooter	rope	candle	net
rail	stool	flag	card	statue	lap	...

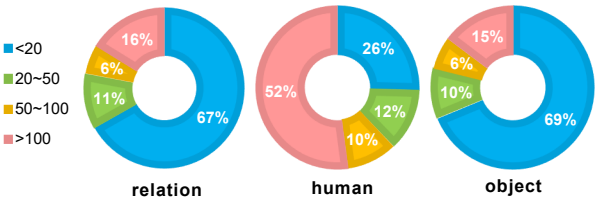

Figure 1. The proportions of the four category subsets of relation, human and object are based on the number of training instances in each category, i.e. > 100, $50 \sim 100$, $20 \sim 50$ and < 20.

*Qin Jin is the corresponding author.

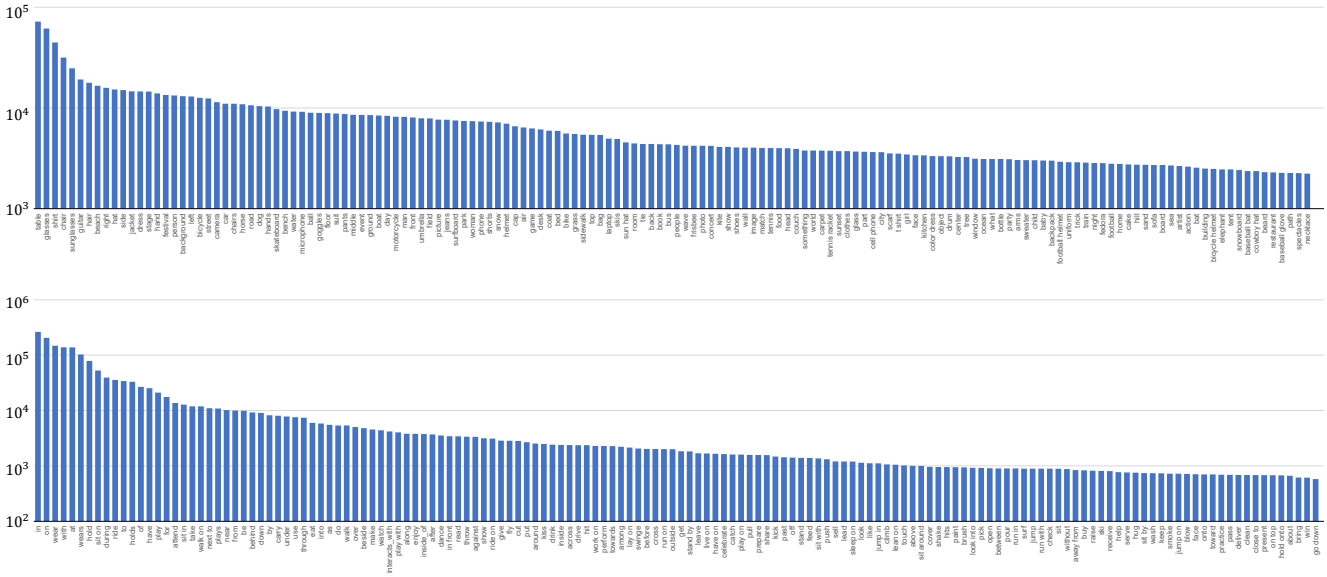

Figure 2. Distribution of the relations (bottom) and nouns (top) in our pre-training data. For clarity, we display the top 150 most frequent relations and nouns.

Table 2. Ablation of masked language prediction (MLP) on HICO-DET.

		mAP (*Full*)
1	w/o MLP	29.61
2	+ auto-reg with 10% mask	30.25
3	+ auto-reg with 30% mask	30.87
4	+ auto-reg with 50% mask	32.68
5	+ auto-reg with 70% mask	32.41

Table 3. Ablation of human-object patch jigsaw (HPJ) on HICO and the HICO-DET *Full* set.

		HICO	HICO-DET *Full*
1	w/o patch shuffle	51.7	24.85
2	+ patch shuffle 10%	51.9	24.97
3	+ patch shuffle 20%	52.5	25.48
4	+ patch shuffle 30%	53.2	25.80
5	+ patch shuffle 40%	52.9	25.64
6	++patch rotation	53.7	26.12

3. Additional Ablation

Ablation of Masked Language Prediction. Table 2 presents the evaluation of various setups for the masked language prediction (MLP) task on the HICO-DET *Full* set [1]. The results show that a masking ratio of 50% yields the best performance, leading to 32.68 mAP in row 4.

Table 4. Ablation of the number of layers in the encoder and decoder on HICO-DET.

enc layers	mAP	dec layers	mAP
3	30.43	3	30.76
4	31.62	4	31.77
5	32.57	5	32.60
6	32.68	6	32.68
7	32.64	7	32.56

Ablation of Human-object Patch Jigsaw. We have evaluated different ratios of shuffled patches and the necessity of patch rotation for the HPJ task, and the results are shown in Table 3. The table indicates that there is only limited improvement when the shuffling ratio is low. For instance, shuffling patches with a probability of 10% only yields an increase of +0.2% mAP on HICO [2] (row 1 vs. row 2). This is primarily because a low shuffling ratio reduces the difficulty of restoring patch positions, causing the model to learn "shortcuts" (e.g., texture in the image) instead of the relative relationship of distinguishable local information between human and potential interacting objects. Furthermore, our results show that using patch rotation is also beneficial, leading to a gain of +0.8% mAP (row 5 vs. row 6), since it increases the task complexity of HPJ pre-training.

Ablation of Model Architecture. We further investigate the impact of the number of layers in the encoder and decoder, as shown in Table 4. Note that we keep the number of layers in the decoder fixed at 6 while altering the number

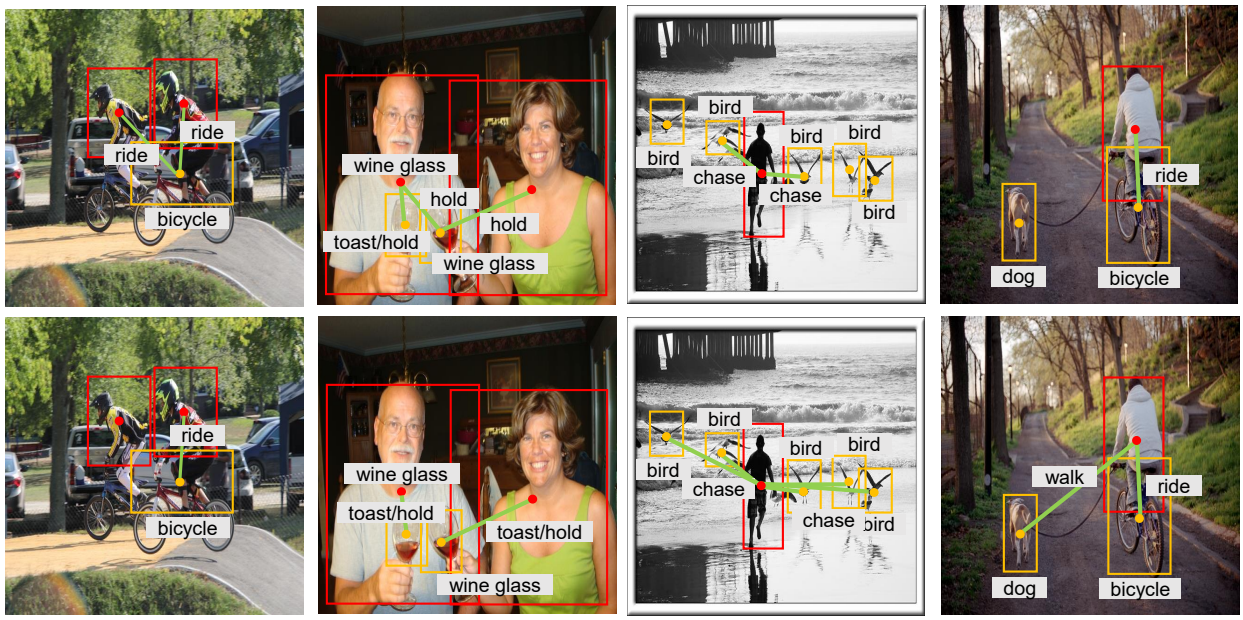

Figure 3. Examples of HOI detection on the HICO dataset. The top row shows the HOI detection obtained using the iCAN model [3], while the bottom row shows the results obtained using our model. Despite not following a one-stage unified pipeline, our model still benefits from the global context of the entire image as well as other contextual HOI triplets.

of layers in the encoder, and vice versa. The findings suggest that our proposed OpenCat achieves the highest performance when the encoder and decoder each have 6 layers for cross-modal encoding.

4. Additional Analysis

Can the performance improvement be attributed solely to the additional pre-training data? The improvement achieved by OpenCat cannot be solely attributed to "more data." In Table.2 (main paper), we compare OpenCat with HAKE†, another pre-trained model that is based on additional fine-grained annotations (e.g., bodypart-level actions). Despite relying only on low-cost pre-training data, OpenCat outperforms HAKE†. We suggest that the performance gains of OpenCat are mainly due to (1) effective pre-training strategies, including the designed proxy tasks presented in Table 6 (main paper), and (2) the auto-regressive structure that allows our model to be pre-trained by weakly-supervised data and predict open-set HOI categories.

Does OpenCat provide only a marginal improvement over the state-of-the-art methods in HOI detection task? Although OpenCat performs slightly better on the HICO-DET *Full* set, it significantly outperforms state-of-the-art CDN on the *Rare* subset, demonstrating its ability to adapt to rare classes caused by the long-tailed distribution. Furthermore, OpenCat directly utilizes an off-the-shelf object detector to obtain object bounding boxes and is thus subject to the bottleneck of object detection. Therefore, we also compare OpenCat with HAKE† on the HOI recognition task, which does not rely on object detection. As shown in Table.2 (main paper), OpenCat outperforms the previous state-of-the-art approach by a considerable margin.

How can we prevent text items with high frequency from affecting the recognition of novel visual concepts? One way is by transferring knowledge from similar HOI compositions, which can lead to better recognition of low frequency concepts. For instance, although the phrase "tear a book" is rare, the verb "tear" is associated with 102 samples in our pre-training data, and most triplets containing "tear" exhibit similar appearance and semantic information. Thus, our model can recognize the act of a man tearing a book apart, even if the majority of training data only depicts a man reading a book. Additionally, we observe that only a few object classes (4.3%) and relation classes (3.6%) contain excessive samples (≥ 1000). To address this imbalance between high and low frequency items, we adopt a sample re-weighting strategy.

More experimental results of the bare model w/o pre-training. As an example, We further present the results of bare model w/o pre-training in weakly-supervised HOI detection as shown in Table 5. As can be seen, the pre-training plays an important role for the final performance.

Computation complexity analysis. OpenCat's 6-layer transformer encoder and decoder have a parameter size that is similar to traditional DETR-based models such as DETR and MDETR. However, the inference speed of OpenCat

Table 5. Results of bare model w/o pre-training in weakly-supervised HOI detection.

Method	HICO-DET			V-COCO	
	Full	Rare	Non-rare	S1	S2
Align-Former	20.85	18.23	21.64	15.8	16.3
OpenCat w/o pre-train	19.72	14.56	21.01	14.6	15.2
OpenCat	**25.82**	**24.35**	**26.19**	**34.4**	**36.1**

is primarily determined by the length of the generated sequences. Our statistics indicate that only 1.2% of the 9545 HICO testing images have ground-truth sequences longer than 256 tokens. Consequently, OpenCat's runtime is approximately 2X faster than other sequence models, such as **Pix2Seq** with 500 token prediction. Moreover, the inference speed of OpenCat can be further increased by limiting the sequence length (e.g., 100 token prediction).

5. Qualitative Analysis

OpenCat does not follow the one-stage HOI pipeline. Also, our model differs from traditional two-stage models in that it can utilize contextual information for HOI prediction. Figure. 3 provides some comparison examples with a two-stage model named iCAN [3]. The top and bottom rows display the results of iCAN and our model, respectively. In column 1, our model infers that the bicycle on which a man is riding should not have any interaction with another person. Similarly, in columns 2 and 3, our model successfully predicts the missing interactions "toast" and "chase" based on contextual human and object information, whereas iCAN fails to do so. Another example in column 4 demonstrates that our model can detect the challenging HOI triplet "person walk dog" with the help of contextual information about the "rope".

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 2

[2] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015. 2

[3] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 3, 4