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

Supplementary Material

1. Validation Set Generation

The HICO-DET doesn't provide a official validation set. When deciding our hyperparameters, we split a validation set from the training set, and also generate a new training set specially for hyperpatameter selection from the remaining examples. We evaluate the performances of each hyperparameters choice on the separated validation set, to ensure fair and reliable results. We train models for hyperparameter selection on the new training set, thus model will not see the training example of validation set during training stage. To guarantee a proper function of the validation set and the new training set, we generate the the two set following the criterion: 1) all training instances are randomly picked from the training set; 2) there is at least one training instance for any class included in the training set. The final validation set contains 12892 images and the new training set contains 18250 images.

2. Supplementary Analysis

Zero-shot HOI Enhancement for GEN-VLKT We conduct experiments to validate the effectiveness of zero-shot HOI enhancement in previous HOI detector e.g. GEN-VLKT. We report the result of different top k selection on test set in Table 1. The training-free enhancement also works for GEN-VLKT and achieve a improvement of +0.56 mAP gain when k equals to 5. We also provide the result of HOICLIP with different k on test set to explore the upper bound for zero-shot HOI enhancement in Table 2. We observe the maximum improvement is achieved when k equals to 20 and we report result with k equals to 5, which is selected by validation set.

Visualization of Improvement We visualize the performance of HOICLIP with zero-shot HOI enhancement and without zero-shot HOI enhancement. The result is showed in Figure 1. We observe the enhancement benefit the tail classes more compared with head classes. We conclude the enhancement is a complement of CLIP knowledge to learned knowledge from HOI detection training data and model with worse performance benefits more from enhancement.

Additional Qualitative Analysis We present the fail cases of conventional HOI detector in Figure 2. The ground truth interaction category for first row is training a dog and

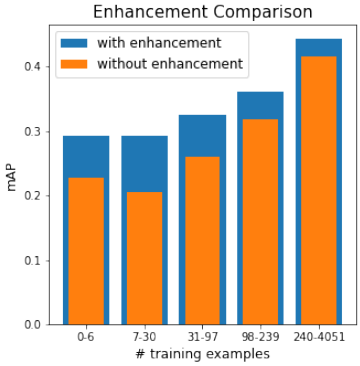


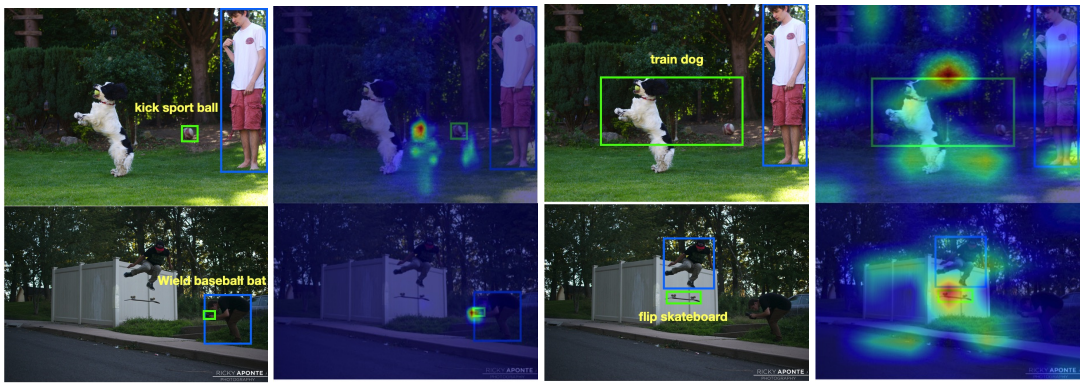
Figure 1. **Enhancement Analysis**: We split all 600 HOI categories into 5 part by the number of training examples, then evaluate the performance of model with/without the training free enhancement method on each part and show the results. The performance gain is distinct especially on the classes with fewer training example, i.e. the tail classes.

K	Full	Rare	Non-Rare
0	33.75	29.25	35.10
5	**34.31**	**30.50**	**35.44**
10	34.15	29.95	35.41
15	34.16	30.01	35.40
20	34.17	29.94	35.43
25	34.13	29.88	35.40

Table 1. **GEN-VLKT with zero-shot HOI enhancement on HICO-DET.**

K	Full	Rare	Non-Rare
0	34.55	30.71	35.70
5	34.54	30.71	35.70
10	34.69	31.18	35.74
15	34.71	31.23	35.74
20	**34.75**	**31.30**	**35.79**
25	34.70	31.11	35.78

Table 2. **HOICLIP with zero-shot HOI enhancement on HICO-DET.**



(a) (b) (c) (d)

Figure 2. **Fail cases of conventional methods.** In (a) and (b), we present the fail cases of GEN-VLKT and corresponding attention map from interaction decoder. In (c) and (d), we visualize the prediction from HOICLIP and corresponding attention map from CLIP spatial feature in interaction decoder.

flipping a skateboard for second row. In the first row, conventional methods wrongly predict that human is interacting with a ball instead of the dog. Meanwhile, in the second row, GEN-VLKT wrongly predict the interaction category where a man is flipping the skateboard instead of the wield a baseball bat. As discussed in the main manuscript, we conclude the difference lies in the focus point of GEN-VLKT and HOICLIP. We observe the attention map of HOICLIP covers more informative regions and aggregate more accurate interaction information. In the other hand, GEN-VLKT simply focus on the object region which is inconsistent with region required for interaction prediction.

Justification for Visual Semantic Arithmetic. In contrast to previous interpretation where the verb representation is interpreted as features of the union region minus the object features, our VSA design aims to reduce the noisy cues from the object regions due to the variation in object classes. We verify the effectiveness of this design in Table 3, which shows our method **Frac** outperforms previous representation **Union.** 2) We use a verb representation extracted from whole dataset in main paper experiments. To investigate the impact of using partial data, we conduct experiments in Table 3 where **Frac** indicates verb representation extracted from only partial data settings. By comparing with HOICLIP, we can see that our model is robust w.r.t different amount of data for VSA.

Ablation Study under Low-data Regime. We conduct ablation study under low-data settings in Table 3 to better demonstrate the characteristic of proposed methods. The result shows that the model relies more on the **CLIP** features with fewer data. However, all the modules work collaboratively to achieve the best performance.

The necessity of CLIP. To verify the necessity of CLIP, we replace the CLIP visual encoder with imagenet pre-trained ViT. The strong performance of HOICLIP is achieved by leveraging the alignment between CLIP's text and visual encoders instead of using more parameters. As shown in Table 3, replacing CLIP ViT with imagenet pre-trained ViT (denoted as **HOICLIP-ViT**) breaks the alignment and the performance degrades.

Method	100%	50%	15%
Union	33.03	30.79	26.80
Frac	34.69	31.11	26.84
Base	32.09	25.54	21.57
+CLIP	32.72	29.80	25.20
+integration	34.13	30.28	25.63
+verb	34.54	30.33	26.25
HOICLIP-ViT	33.03	28.28	23.91
HOICLIP	**34.69**	**30.88**	**27.07**

Table 3. Fractional data performance on HICO-DET.

3. Limitation Discussion

We notice the training parameter number for conventional methods are different with HOICLIP since HOICLIP include CLIP visual encoder as a indivisible part of its network architecture. Specifically, during training under default setting, GEN-VLKT leverage CLIP as a teacher model for knowledge distillation and finetune CLIP with a smaller learning rate while HOICLIP freeze all of CLIP parameters during training and inference. However, during inference, HOICLIP require CLIP spatial feature which leads to additional cost in CLIP visual encoder. In summary, HOICLIP require less training cost but more inference cost compared with GEN-VLKT. We regard the cost in inference is inevitable for better integrating CLIP knowledge and achieving more generalized and data efficient HOI detectors.