CLIP4HOI: Towards Adapting CLIP for Practical Zero-Shot HOI Detection

Supplementary Material

Yunyao Mao[1] **Jiajun Deng**[2] **Wengang Zhou**[1†] **Li Li**[1] **Yao Fang**[3] **Houqiang Li**[1†]
[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China
[2]The University of Adelaide, AIML [3]Merchants Union Consumer Finance Company Limited
myy2016@mail.ustc.edu.cn, jiajun.deng@adelaide.edu.au, zhwg@ustc.edu.cn
lil1@ustc.edu.cn, fangyao@mucfc.com, lihq@ustc.edu.cn

A More Technical Details

A.1 HO Interactor

We borrow the design of the lightweight interaction head in [6] to instantiate our HO interactor. For completeness, we elaborate on the technical details as follows:

Pairwise Positional Encodings: We first introduce the pairwise positional encodings [6] used in the HO interactor. Given a pair of bounding boxes $\mathbf{b}_h = [x_1, y_1, w_1, h_1]^\top$ and $\mathbf{b}_o = [x_2, y_2, w_2, h_2]^\top$. We first compute unary and pairwise spatial features as follows:

$$\mathbf{u} = \Big[x_1, y_1, w_1, h_1, x_2, y_2, w_2, h_2, w_1 h_1, w_2 h_2, \frac{w_1}{h_1}, \frac{w_2}{h_2}\Big]^\top, \tag{1}$$

$$\mathbf{p} = \Big[\frac{w_1 h_1}{w_2 h_2}, \texttt{IoU}(\mathbf{b}_h, \mathbf{b}_o), \texttt{ReLU}(\frac{x_1 - x_2}{w_1}), \texttt{ReLU}(\frac{x_2 - x_1}{w_1}), \texttt{ReLU}(\frac{y_1 - y_2}{h_1}), \texttt{ReLU}(\frac{y_2 - y_1}{h_1})\Big]^\top, \tag{2}$$

where IoU denotes intersection over union. The unary feature $\mathbf{u}$ consists of boxes, box areas, and aspect ratios. The pairwise feature $\mathbf{p}$ consists of the ratio of box areas, intersection over union, and directional encodings that describe the distance between boxes. A multilayer perceptron is then adopted to map the spatial features into pairwise positional encoding:

$$\mathbf{e} = \texttt{MLP}\Big(\texttt{CAT}\big(\mathbf{u}, \mathbf{p}, \log(\mathbf{u} + \epsilon), \log(\mathbf{p} + \epsilon)\big)\Big) \in \mathbb{R}^{C_r}, \tag{3}$$

, where CAT and MLP denote concatenation and multilayer perceptron, respectively.

Cooperative Layer: Taking decoded query features $\mathbf{X} \in \mathbb{R}^{N_d \times C_d}$ of detected instances and the pairwise positional encodings $\mathbf{E} \in \mathbb{R}^{N_d \times N_d \times C_r}$ between every two instances as input, the cooperative layer performs self-attention and pairwise positional information injection as follows:

$$\mathbf{X}' = \texttt{Linear}(\mathbf{X}) \in \mathbb{R}^{N_d \times C_r}, \quad \mathbf{E}' = \texttt{Linear}(\mathbf{E}) \in \mathbb{R}^{N_d \times N_d \times C_r}, \tag{4}$$

$$\dot{\mathbf{X}}' \in \mathbb{R}^{N_d \times N_d \times C_r}, \quad \text{where } \dot{\mathbf{X}}'^{[i]} \triangleq \mathbf{X}' \in \mathbb{R}^{N_d \times C_r}, \tag{5}$$

$$\ddot{\mathbf{X}}' \in \mathbb{R}^{N_d \times N_d \times 2C_r}, \quad \text{where } \ddot{\mathbf{X}}'^{[i,j]} \triangleq \texttt{CAT}(\mathbf{X}'^{[i]}, \mathbf{X}'^{[j]}) \in \mathbb{R}^{2C_r}, \tag{6}$$

$$\mathbf{V} = \dot{\mathbf{X}}' \cdot \mathbf{E}', \quad \mathbf{W} = \texttt{Softmax}\Big(\texttt{Linear}\big(\texttt{CAT}(\ddot{\mathbf{X}}', \mathbf{E})\big)\Big) \in \mathbb{R}^{N_d \times N_d \times 1}, \tag{7}$$

$$\mathbf{X}_{\text{attn}} = \texttt{LN}\Big(\mathbf{X} + \texttt{Linear}\big(\texttt{Mean}(\mathbf{W} \cdot \mathbf{V}, \texttt{dim} = 1)\big)\Big) \in \mathbb{R}^{N_d \times C_r}, \tag{8}$$

$$\hat{\mathbf{X}} = \texttt{LN}\big(\mathbf{X}_{\text{attn}} + \texttt{FFN}(\mathbf{X}_{\text{attn}})\big) \in \mathbb{R}^{N_d \times C_r}, \tag{9}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

where Linear, Softmax, FFN, LN denote linear projection, softmax function, feed-forward network, and layer normalization, respectively.

Competitive Layer: The competitive layer takes the output of cooperative layer $\hat{\mathbf{X}}$, the global visual feature $\mathbf{X}_{\text{glob}}$ from the backbone of DETR [1], and the pairwise positional encodings $\mathbf{E}$ as inputs. First, the valid pairwise query features $\mathbf{X}_{\text{pair}}$ and positional encodings $\mathbf{E}_{\text{pair}}$ are extracted according to the index set of valid human-object pairs $\mathbf{idx}_{\text{valid}}$:

$$\mathbf{idx}_{\text{valid}} = \{(u, v) | u \neq v, c_u = \text{“human”}\}, \tag{10}$$

$$\mathbf{X}_{\text{pair}} \in \mathbb{R}^{N_{\text{pair}} \times 2C_r}, \quad \text{where } \mathbf{X}_{\text{pair}}^{[i]} = \texttt{CAT}(\hat{\mathbf{X}}^{[u]}, \hat{\mathbf{X}}^{[v]}),\ (u, v) = \mathbf{idx}_{\text{valid}}^{[i]}, \tag{11}$$

$$\mathbf{E}_{\text{pair}} \in \mathbb{R}^{N_{\text{pair}} \times C_r}, \quad \text{where } \mathbf{E}_{\text{pair}}^{[i]} = \mathbf{E}^{[u,v]},\ (u, v) = \mathbf{idx}_{\text{valid}}^{[i]}. \tag{12}$$

Then, the valid pairwise query features $\mathbf{X}_{\text{pair}}$ and global visual feature $\mathbf{X}_{\text{glob}}$ are respectively fused with the valid pairwise positional encodings $\mathbf{E}_{\text{pair}}$. The resulting $\mathbf{X}'_{\text{pair}}$ and $\mathbf{X}'_{\text{glob}}$ are concatenated together and processed by a transformer encoder layer to generate the final pairwise HOI tokens $\mathbf{Q}$:

$$\mathbf{X}'_{\text{pair}} = \texttt{Linear}\Big(\texttt{ReLU}\big(\texttt{Linear}(\mathbf{X}_{\text{pair}}) \cdot \texttt{Linear}(\mathbf{E}_{\text{pair}})\big)\Big) \in \mathbb{R}^{N_{\text{pair}} \times C_r}, \tag{13}$$

$$\mathbf{X}'_{\text{glob}} = \texttt{Linear}\Big(\texttt{ReLU}\big(\texttt{Linear}(\mathbf{X}_{\text{glob}}) \cdot \texttt{Linear}(\mathbf{E}_{\text{pair}})\big)\Big) \in \mathbb{R}^{N_{\text{pair}} \times C_r}, \tag{14}$$

$$\mathbf{Q} = \texttt{TransformerEncoderLayer}\big(\texttt{CAT}(\mathbf{X}'_{\text{pair}}, \mathbf{X}'_{\text{glob}})\big) \in \mathbb{R}^{N_{\text{pair}} \times 2C_r}, \tag{15}$$

A.2 Positional Distribution Discrepancy

In this section, we first give the definition of positional distribution discrepancy and then provide some visualizations of the distribution discrepancy between seen and unseen HOI categories.

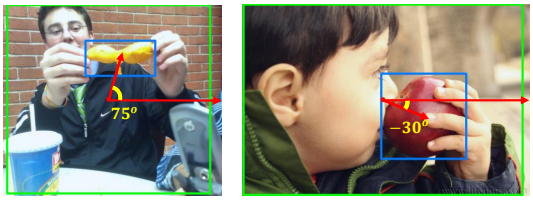


Figure 1: An illustration of the distribution statistics, *i.e.*, the angles between the line from the person to the object and the x-axis.

Definition: To verify the robustness of our approach to the distribution gap between seen and unseen HOI categories, we introduce the concept of positional distribution discrepancy. We first compute the positional distribution of the seen and unseen HOI categories corresponding to each object. The distribution statistics are the angles (quantized into 90 discrete bins) between the line from the person to the object and the x-axis, as shown in Figure 1. Then, we compute the KL divergence between the distributions of seen and unseen categories to measure the distribution discrepancy.

Visualization: In Figure 2, we provide some visualizations of the positional distribution of seen and unseen HOI categories corresponding to each object class.

A.3 Training Details

We adopt the AdamW [3] optimizer with a weight decay of 1e-4. We train the network for 100 epochs with an effective batch size of 16. The learning rate is linearly increased to 1e-4 from 0 in the first 5 warm-up epochs and then decreased to 1e-7 by the cosine decay schedule. The experiments are conducted on four NVIDIA RTX 3090 GPUs.

B Border Impact and Limitations

Border Impact: The proposed CLIP4HOI framework significantly improves the performance and robustness of zero-shot HOI detection. With the strong generalization capability, CLIP4HOI can empower many practical downstream tasks, such as surveillance, augmented reality, and human-computer interaction. Human-object interaction detection also acts as a cornerstone of the recently popular embodied artificial intelligence.

Limitations: Unlike previous top-performing methods [2, 5] that only adopt the CLIP [4] model as the teacher during training, our CLIP4HOI retains the CLIP visual encoder in the test phase to

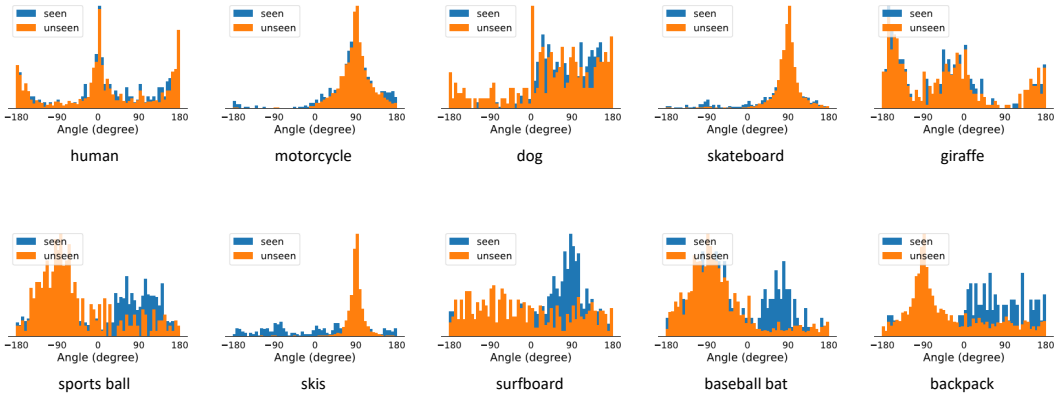


Figure 2: Visualization of positional distribution. For each object class, there exist small (first row) or large (second row) distribution discrepancies between the corresponding seen and unseen HOI categories, posing challenges to the generalization of the algorithm.

provide contextual visual clues for the HOI decoder. This allows CLIP4HOI to achieve superior performance compared to the previous methods but brings more computational overhead to a certain extent. In addition, although the CLIP model exhibits a strong generalization capability during proposal discrimination, the diversity of proposal generation is still limited by the categories that the detector can identify. Fortunately, thanks to the two-stage design of CLIP4HOI, in the future, we can easily integrate advanced open-vocabulary object detectors into our framework to provide more diverse HOI proposals.

C More Qualitative Results

In Figure 3 and Figure 4, we provide more qualitative visualizations of our CLIP4HOI.

References

[1] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020) 2

[2] Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: GEN-VLKT: Simplify association and enhance interaction understanding for hoi detection. In: CVPR. pp. 20123–20132 (2022) 2

[3] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 2

[4] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 2

[5] Wu, M., Gu, J., Shen, Y., Lin, M., Chen, C., Sun, X., Ji, R.: End-to-end zero-shot hoi detection via vision and language knowledge distillation. In: AAAI (2023) 2

[6] Zhang, F.Z., Campbell, D., Gould, S.: Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In: CVPR. pp. 20104–20112 (2022) 1

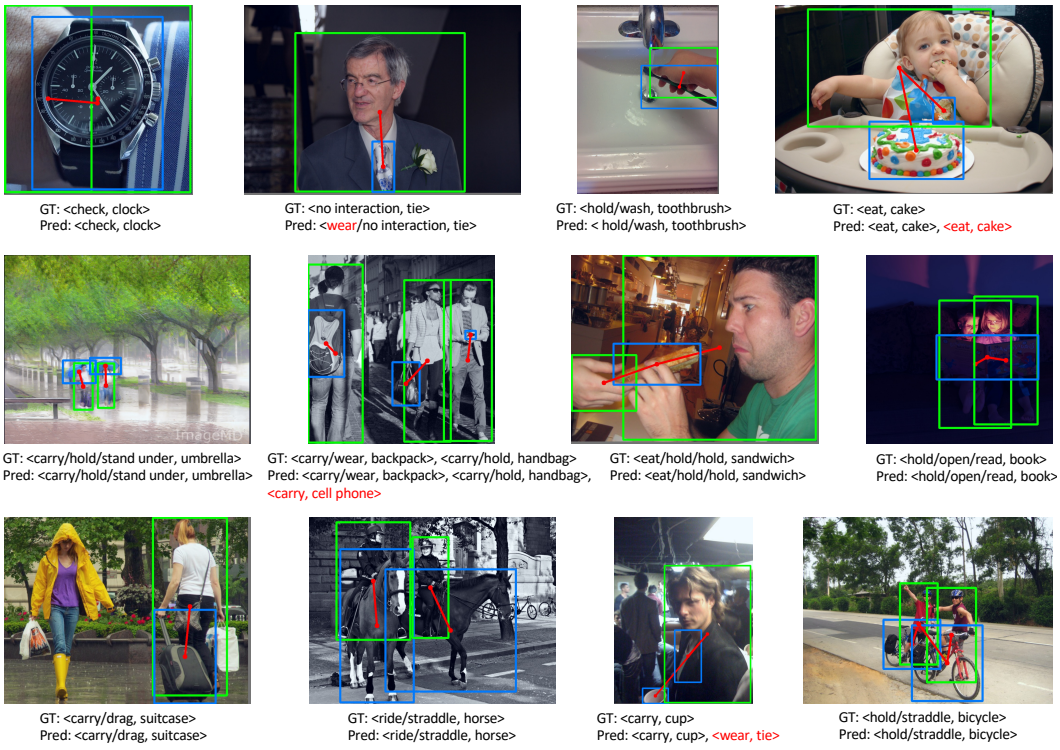
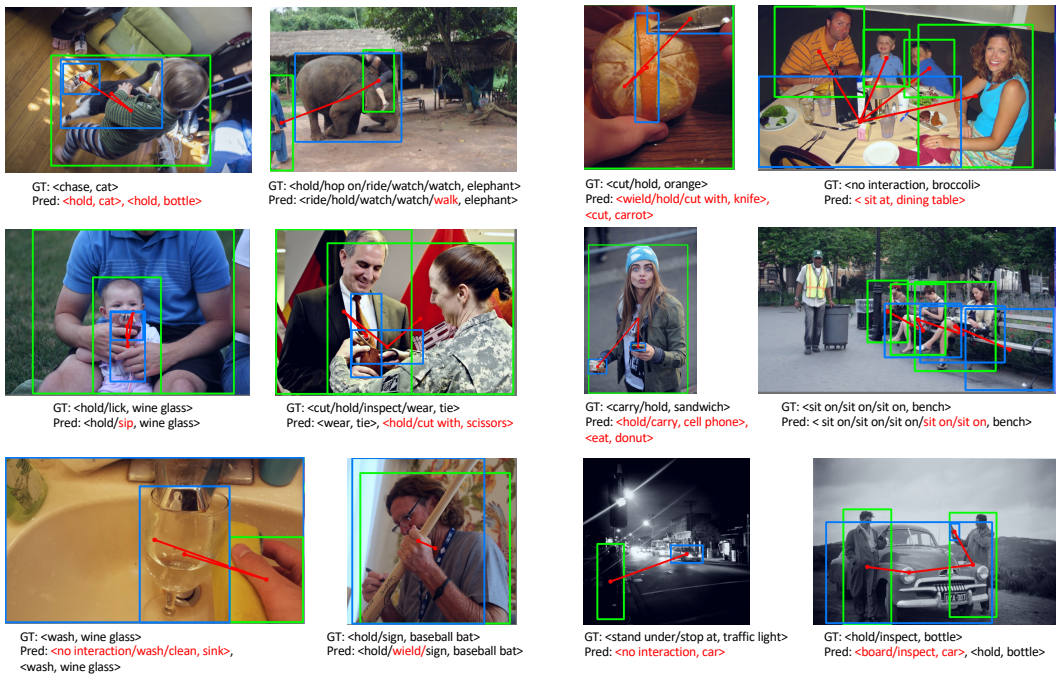


Figure 3: Additional qualitative results for detected human–object pairs on the HICO-DET test set. We can find that our CLIP4HOI can cope well with multiple instances, incomplete subjects/objects, motion blur, etc. Moreover, it can also detect some human-object interactions that are not labeled but correct (marked in red).



(a) Failure cases caused by HOI discrimination.

(b) Failure cases caused by object detection.

Figure 4: Failure cases. Through observation, we found that the failure cases can be roughly divided into two groups, one is the detection error of the humans and the objects, and the other is the error in the discrimination of the interaction category for each human-object pair.