

Supplementary Materials: Towards Open-vocabulary HOI Detection with Calibrated Vision-language Models and Locality-aware Queries

Anonymous Author(s)

ABSTRACT

This supplementary material presents additional information on implementation details and discussions on additional experiments: (1) Section 1: More Implementation Details ; (2) Section 2: Pretraining Details; (3) Section 3: CLIP Adapter Configuration; (4) Section 4: More Qualitative Results.

A MORE IMPLEMENTATION DETAILS

To obtain visual and spatial-wise embedding, we use two self-attention layers, each of which has 8 heads, with a dropout rate set to 0.1. The pair interaction decoder has 2 layers and the head is set to 8 for both multi-head self-attention and cross-attention layers. The Adam is adopted for the training phase, with both the learning rate and weight decay being $1e - 4$. We train the model for 50 epochs conducted on 4 Nvidia A800 GPUs, with a batch size of 120.

As mentioned in Section 4.5, we compute the hard loss \mathcal{L}_{hard} and the soft loss \mathcal{L}_{soft} . The detailed calculation can be referred as follows:

$$\mathcal{L}_{soft} = Focal(S_{clip}, S_v) \times SoftMask \quad (1)$$

$$\mathcal{L}_{hard} = Focal(S_{gt}, S_v) \times (1 - SoftMask) \quad (2)$$

where $Focal(\cdot)$, S_{clip} , S_{gt} , S_v , and $SoftMask$ denote focal loss [2], CLIP labels, ground truth labels, predicted action score and soft mask as mentioned in Section 4.4.

Base Verb	Novel Verb
ride instr	hold obj
hit instr	sit instr
hit obj	look obj
eat obj	eat instr
jump instr	lay instr
talk on phone instr	carry obj
throw obj	catch obj
cut instr	ski instr
cut obj	drink instr
work on computer instr	snowboard instr
surf instr	
kick obj	
read obj	
skateboard instr	

Table 1: The 14 base verbs and 10 novel verbs of V-COCO in our split.

B PRETRAINING DETAILS

We apply the three main interaction datasets for the first step pre-training, including HICO-DET, V-COCO, and VG150. These datasets provide bounding box coordinates, object categories, and interaction categories, which can be directly used to enhance CLIP’s alignment capabilities on HOI. To avoid introducing novel category information and leaking test dataset information, we only keep the base classes in the training sets of the three datasets for pretraining to eliminate data leakage issue. Notably, we use the same base and novel categories for the three datasets. However, due to the different annotation schemes in the three datasets, the datasets may have different label symbols but have the same meaning. Hence, given a predefined base category list, if a class label’s synonymous name is also in the list, it can be viewed as an extend base category. Otherwise, it is put into the novel list. Specifically, for HICO-DET, we follow [4] to split the base and novel categories according to different setting. For V-COCO, we provide our split as in Table 1. We use Adam as an optimizer in pre-training, with a learning rate of $5e - 5$, beta 0.9 and 0.98 for the first and second moments, and a weight decay of 0.2. For every 10 steps, the learning rate is reduced by a factor of 0.1. We conduct pretraining for 35 epochs with 4 Nvidia A800 GPUs and a batch size of 120.

C CLIP ADAPTER CONFIGURATION

As mentioned in Section 4.1 of our paper, we employ an adapter technique to calibrate CLIP, i.e. **CaCLIP**. In this section, we conduct experiments to investigate the effects of different adapter configurations on the performance of CLIP. Following the approach [1], we first introduce multi-layer perceptrons (MLP) separately in the image and language branches, denoted as MLP. Besides, we test the other two adapter variants, MLP with a residual connection and transformer-based encoder adapter [3], denoted as MLP w/res and Transformer Encoder. For MLP w/res, the calibrated embeddings are then residual fused with the original CLIP embeddings, represented as:

$$X^* = \beta \text{Adapter}(X) + (1 - \beta)(X) \quad (3)$$

where Adapter denotes the image or text adapter, X denotes an image or text feature from CLIP, X^* is calibrated embeddings and β is a hyperparameter which we set to 0.2 following [1]. For a fair comparison, the number of layers in the both MLP and transformer encoder is set to 2. The results are presented in Table 2. We observe that involving MLP in the image and text branches without residual addition operations achieves optimal performance, resulting in the best calibration of CLIP on HOI.

Furthermore, we investigate different adapter layers for the image and language branches. As shown in Table 3, having the same number of adapter layers leads to better performance than scenarios where one branch was omitted. A two-layer MLP outperformed

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174

175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

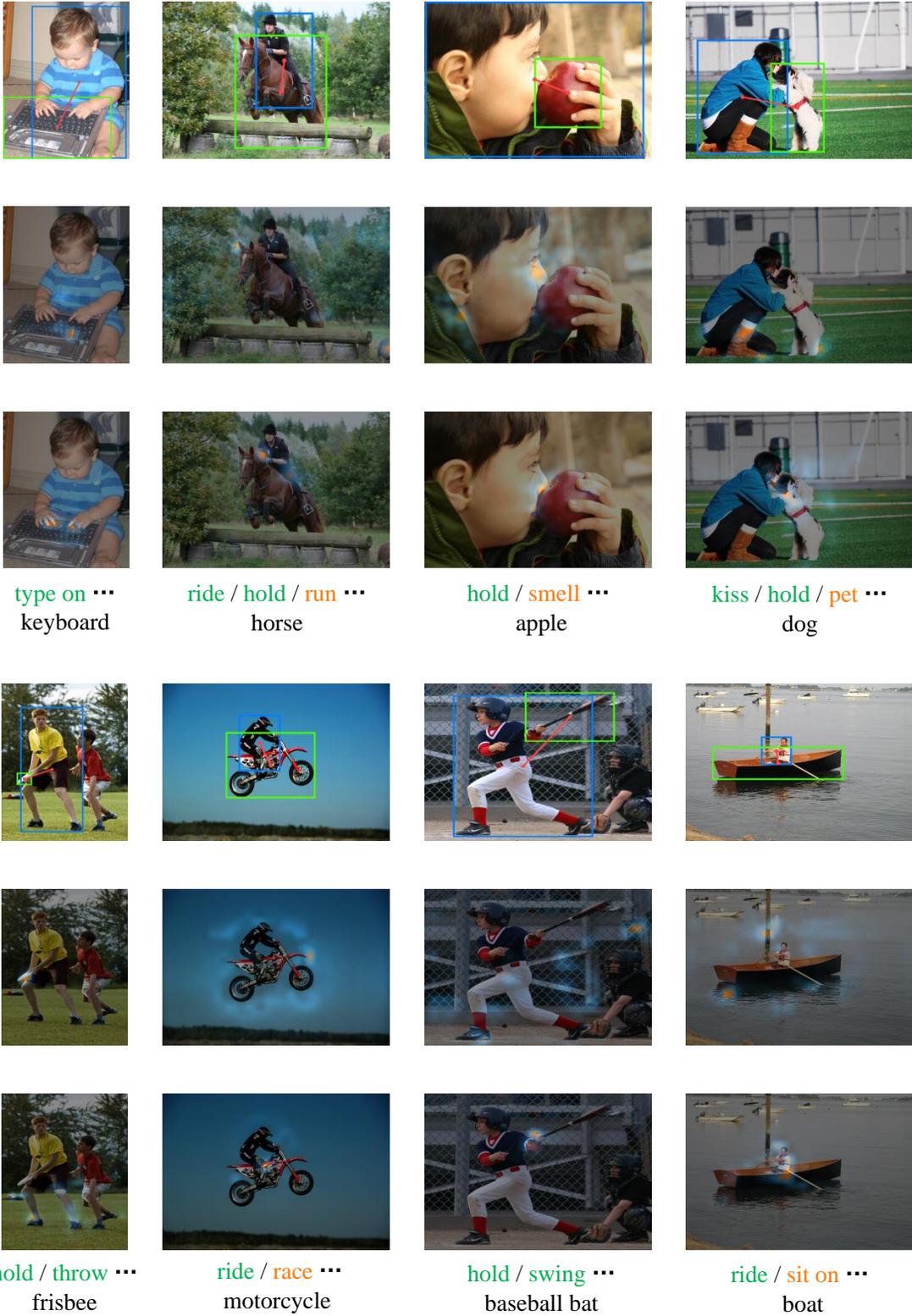


Figure 1: More qualitative results of our CaM-LQ. We show localization results, attention maps of the decoder w/o and w spatial priors respectively. **Green**: correctly detected base category. **Orange**: correctly detected novel category.

Variants	Full	Seen	Unseen
MLP	38.00	38.90	34.39
MLP w/ res	35.21	35.71	33.19
Transformer Encoder	36.12	37.09	32.24

Table 2: The performance with different CLIP adapter variants.

N_{img}	N_{txt}	Full	Seen	Unseen
0	1	33.96	34.29	32.62
1	0	35.30	35.69	33.76
1	1	35.96	37.25	30.81
2	2	38.00	38.90	34.39
3	3	36.22	37.12	32.64

Table 3: The performance with different numbers of adapter layers in CLIP image and text branch. N_{img} denotes the number of layers of the image branch while N_{txt} refer to that of the text branch

a single-layer MLP by a large margin of 2.04 mAP, and introducing the third linear layer does not yield improved results. Hence, we choose two MLP layers as our default configuration for our adapter module.

D MORE QUALITATIVE RESULTS

We showcase more qualitative results in Figure 1. As presented, CaM-LQ demonstrates outstanding localization capabilities. With the integration of spatial embedding, our approach can more accurately attend to regions of human-object interaction. In contrast, models without spatial embedding introduce more irrelevant areas of interference. By incorporating CaCLIP, our method can predict novel categories beyond predicting base categories, showcasing robust Ov-HOI capability.

REFERENCES

- [1] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2023. Clip-adapter: Better vision-language models with feature adapters. *ICCV* (2023), 1–15.
- [2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [4] Mingrui Wu, Jiaxin Gu, Yunhang Shen, Mingbao Lin, Chao Chen, and Xiaoshuai Sun. 2023. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2839–2846.