Supplementary for Paper: Bilateral Adaptation for Human-Object Interaction Detection with Occlusion-Robustness

Table 1. Results with different CLIP encoders.

Method	Encoder Size	Full	Rare	Non-rare
ADA-CM [1]	B	33.80	31.72	34.42
BCOM (Ours)	B	**36.04**	**34.96**	**36.36**
ADA-CM [1]	L	38.40	37.52	38.66
BCOM (Ours)	L	**39.34**	**39.90**	**39.17**

Table 2. Results with ResNet-101 as backbone.

Method	Full	Rare	Non-rare
UPT [6]	32.31	28.55	33.44
SDT [5]	32.97	28.49	34.31
GEN-VLKT [2]	34.63	30.04	36.01
ViPLO [3]	37.22	35.45	37.75
BCOM (Ours)	**39.52**	**42.26**	**38.71**

1. More Experimental Results

Results with Different CLIP Encoders. Following [1], our work adopts CLIP-L [4] as the visual encoder for extracting semantic features. In fact, there are CLIP of different sizes, we thus provide a comparison with different CLIP encoders in Table 1. The results show that our method is consistently better than the baseline [1]. This indicates the necessity of adapting spatial features, although the performance gap is narrowed when shifting the CLIP encoder from Base (B) to Large (L) size.

Results with Different Backbones. Besides using ResNet-50 as the detector backbone, Previous work also utilizes ResNet-101 as one important backbone for the detector. We provide the performance comparison in Table 2. The results show that our BCOM still achieves superior performance with a stronger backbone.

CCM for Far-apart Human-Objects. We visualized the attention in the proposed Conditional Contextual Mining (CCM) Module of two examples when a human and an object are far apart in the figure above. For the people `flying kites`, our method can roughly attend to the most informative context (*i.e.*, arm and head) in (a). However, the attention mechanism does not focus on the most informative human parts in (b). This may be due to 1) the person and object are too far away and their interaction clue is very difficult to identify. 2) the person and object are too small in the image and their ROI feature is less informative.

2. Structure of the Adapters

Structure of Spatial Adapter. The spatial adapter that is appended to each block of the detector backbone is two linear layers with ReLU activation in between. The dimension of the middle representation is one-fourth of the feature dimension dim. The pseudo-code is as follows:

```
class SpatialAdapter(nn.Module):
    def __init__(self, in_dim, out_dim):
        self.adapter = nn.Sequential(
            nn.Linear(in_dim, in_dim // 4),
            nn.ReLU(),
            nn.Linear(in_dim // 4, out_dim)
        )

    def forward(self, x):
        return self.adapter(x)
```

Structure of Semantic Adapter. The semantic adapter follows the structure of [1], where a Transformer Decoder

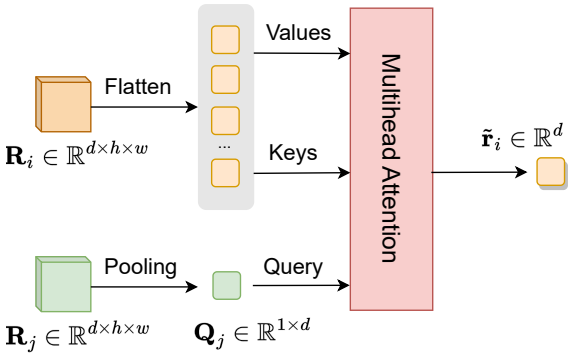

Figure 1. The Conditional Contextual Mining (CCM) module. We use the pooled feature of $\mathbf{R}_j$ as a query and obtain the most informative feature in $\mathbf{R}_i$ via multi-head attention.

Layer is adopted. The instance information including the bounding box and confidence score are taken as the keys and values, while the visual features are used as queries.

Structure of CCM. For better understanding, we visualize the architecture of our proposed Conditional Contextual Mining (CCM) module in Figure 1.

References

[1] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Ji, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. 2023. 1

[2] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. 2021. 1

[3] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *CVPR*, 2023. 1

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[5] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli. Distance matters in human-object interaction detection. In *ACM MM*, 2022. 1

[6] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 1