Mining Instance-Centric Vision-Language Contexts for Human-Object Interaction Detection

Supplementary Material

S1. Implementation Details

Hyperparameters. For object detection, we apply Non-Maximum Suppression (NMS) with an Intersection over Union (IoU) threshold of 0.5 and filter out detections with confidence scores below 0.05, retaining 3 to 15 humans and objects each. The VLM input resolution is set to 224 pixels for ViT-B and 336 pixels for ViT-L variants of CLIP [16]. Additional hyperparameters are detailed in Table S1.

Table S1. Hyperparameters of the InCoM-Net.

Hyperparameters	
Detector feature dim.	256
VLM feature dim.	768/1024 (ViT-B/L)
ICR / ProCA output dim.	384
HO pair feature dim.	384
Interaction Decoder feature dim.	384
Activation function	GELU

Data Augmentations. Following prior works [2, 18, 21], we use common data augmentations. Images are randomly resized with the shorter side ranging from 480–800 pixels and the longer side capped at 1333 pixels. Random cropping produces 384–600 pixels square crops with probability 0.5. Color jittering randomly adjusts brightness, contrast, and saturation with factors sampled from 0.6 to 1.4.

Training and Inference. During training, we masked out invalid actions for each object, following previous works [21, 22]. At inference time, we integrate detection confidence scores into the computation of interaction scores. The final interaction scores are computed as

$$c = (c_h \cdot c_o)^{\alpha} \cdot c_a, \tag{S1}$$

where c_h and c_o denote the detection confidence scores of humans and objects. The hyperparameter α is set to 2.8.

Masked Feature Training (MFT). Fig. S1 provides a detailed illustration of the MFT procedure. In the VLM-only configuration, detector features q^L and CNN backbone features F are masked and excluded from processing. The HO Pair Generator forms HO pair features $\hat{s}$ using the VLM context-aggregated features f^L while the detector features q^L are zeroed out. Within the Interaction Decoder, the cross-attention block associated with the masked CNN features F is zeroed out.

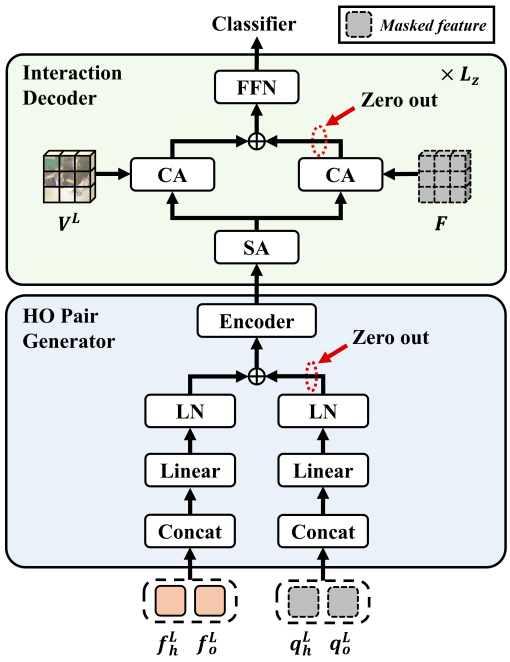

Figure S1. Illustration of MFT under the VLM-only configuration.

Conversely, in the detector-only configuration, VLM context-aggregated features f^L and VLM features V^L are masked in the same manner.

S2. Additional Zero-shot Experiments

In addition to the zero-shot HOI detection results presented in the main paper, we evaluate our model on HICO-DET [3] under three zero-shot settings—Unseen Object (UO), Unseen Verb (UV), and Unseen Combination (UC)—following previous works [9, 13, 15].

In the UO setting, 12 object categories and associated HOI classes are treated as unseen. In the UV setting, 20 verb categories and corresponding HOI classes are excluded from training. In the UC setting, all object and verb categories are seen during training, while a subset of HOI triplets is held out as unseen.

For the UV setting, we use frozen CLIP text embeddings to compute prediction scores following prior works [10, 14]. Text prompts are constructed using the template "A photo of a person [verb-ing] an object" and encoded through CLIP text encoder. We obtain pairwise and global scores by pro-

Table S2. Complete zero-shot performance comparison on HICO-DET. Models marked with † and ‡ use BLIP [11] and BLIP-2 [12] as the VLM backbone, respectively, while all other models use CLIP [16]. For object detection, all models adopt a ResNet-50 backbone [4].

	RF-UC			NF-UC			UO			UV			UC		
Method	**Unseen**	**Seen**	**Full**	**Unseen**	**Seen**	**Full**	**Unseen**	**Seen**	**Full**	**Unseen**	**Seen**	**Full**	**Unseen**	**Seen**	**Full**
VLM-based methods with ViT-B															
GEN-VLKT [13]	21.36	32.91	30.56	25.05	23.38	23.71	10.51	28.92	25.63	20.96	30.23	28.74	-	-	-
HOICLIP [15]	25.53	34.85	32.99	26.39	28.10	27.75	16.20	30.99	28.53	24.30	32.19	31.09	23.15	31.65	29.93
ADA-CM [9]	27.63	34.35	33.01	32.41	31.13	31.39	–	–	–	–	–	–	–	–	–
CLIP4HOI [14]	28.47	35.48	34.08	31.44	28.26	28.90	31.79	32.73	32.58	26.02	31.14	30.42	27.71	33.25	32.11
EZ-HOI [7]	29.02	34.15	33.13	33.66	30.55	31.17	33.28	32.06	32.27	25.10	33.49	32.32	–	–	–
CMMP [10]	29.45	32.87	32.18	32.09	29.71	30.18	33.76	31.15	31.59	26.23	32.75	31.84	29.60	32.39	31.84
SCTC++† [17]	30.36	35.06	34.12	29.68	31.89	31.45	-	-	-	-	-	-	-	-	-
HOLa [8]	30.61	35.08	34.19	35.25	31.64	32.36	36.45	33.02	33.59	27.91	35.09	34.09	–	–	–
VDRP [20]	31.29	34.41	33.78	36.45	31.60	32.57	36.13	32.84	33.39	26.69	33.72	32.73	–	–	–
LAIN [6]	31.83	35.06	34.41	36.41	32.44	33.23	37.88	33.55	34.27	28.96	33.80	33.12	31.64	35.04	34.36
InCoM-Net	**32.65**	**38.69**	**37.48**	**38.68**	**35.00**	**35.74**	**38.26**	**35.47**	**35.94**	**29.59**	**36.34**	**35.39**	**33.57**	**37.79**	**37.00**
VLM-based methods with ViT-L															
BCOM [19]	28.52	35.04	33.74	33.12	31.76	32.03	–	–	–	–	–	–	–	–	–
UniHOI‡ [1]	28.68	33.16	32.27	28.45	32.63	31.79	19.72	34.76	31.56	26.05	36.78	34.68	–	–	–
EZ-HOI [7]	34.24	37.35	36.73	36.33	34.47	34.84	38.17	36.02	36.38	28.82	38.15	36.84	–	–	–
CMMP [10]	35.98	37.42	37.13	33.52	35.53	35.13	39.67	36.15	36.74	30.84	37.28	36.38	34.46	37.15	36.56
InterProDa [5]	36.38	40.88	39.58	33.64	36.47	35.50	–	–	–	–	–	–	–	–	–
LAIN [6]	36.57	38.54	38.13	37.52	35.90	36.22	40.78	36.96	37.60	32.05	38.04	37.20	32.25	37.95	36.81
VDRP [20]	36.72	38.48	38.13	37.48	36.21	36.46	39.36	37.50	37.81	31.16	38.16	37.18	–	–	–
InCoM-Net	**37.69**	**41.46**	**40.71**	**39.45**	**38.42**	**38.62**	**42.32**	**38.97**	**39.53**	**33.31**	**40.15**	**39.19**	**36.00**	**41.32**	**40.32**

jecting HO pair features and CLIP *cls* token, then computing cosine similarity with text embeddings. Final interaction scores are obtained by integrating these scores with detection confidence, similar to Eq. S1.

Performance Comparison. We compare InCoM-Net with existing methods under the RF-UC, NF-UC, UO, UV, and UC zero-shot settings. As shown in Table S2, InCoM-Net achieves state-of-the-art performance across all settings, outperforming all previous best methods [6, 10, 20].

With ViT-B, InCoM-Net surpasses previous best methods on RF-UC, NF-UC, UO, UV, and UC splits by 0.82, 2.23, 0.38, 0.63, and 1.93 unseen mAP, respectively. With ViT-L, the improvements are 0.97, 1.93, 1.54, 1.26, and 1.54 unseen mAP on the same five splits.

Notably, InCoM-Net achieves the best results in both zero-shot and fully supervised settings. Unlike previous methods, which often suffer from a trade-off between the two regimes, InCoM-Net preserves strong accuracy in both without compromising either.

S3. Computational Efficiency

We compare the computational complexity of our method with representative approaches in Table S3. All compared methods utilize CLIP ViT-B backbone. InCoM-Net achieves 39.53 full mAP with 165.4M parameters and 121 GFLOPs, operating within a comparable complexity range to existing methods. The results demonstrate that our approach attains substantial performance improvements while maintaining competitive computational efficiency.

Table S3. Comparison of model complexity on HICO-DET.

Method	Params (M)	GFLOPs	mAP
CMMP [10]	193.4	114	33.24
HOICLIP [15]	193.3	179	34.69
CLIP4HOI [14]	262.4	186	35.33
LAIN [6]	145.4	110	36.02
InCoM-Net	165.4	121	39.53

S4. Additional Qualitative Results

We present additional qualitative comparisons between InCoM-Net and the baseline on the HICO-DET test set. As shown in Fig. S2, the baseline frequently misidentifies interacting pairs in scenes with multiple objects. It often predicts common actions in the scene or confuses which person is interacting when multiple people are present. For example, in the boat scene, the baseline incorrectly assigns rowing to the man due to surrounding rowers. In contrast, InCoM-Net better captures contextual cues and human-object relationships to correctly identify true interactions.

Fig. S3 compares activation maps from InCoM-Net and the baseline. While the baseline focuses on instance regions, InCoM-Net selectively captures interaction-relevant context. For example, InCoM-Net highlights tools on the ground in bicycle repair, fragments in orange peeling, and loading surfaces in truck scenes. This shows InCoM-Net effectively captures relevant context for better HOI understanding.

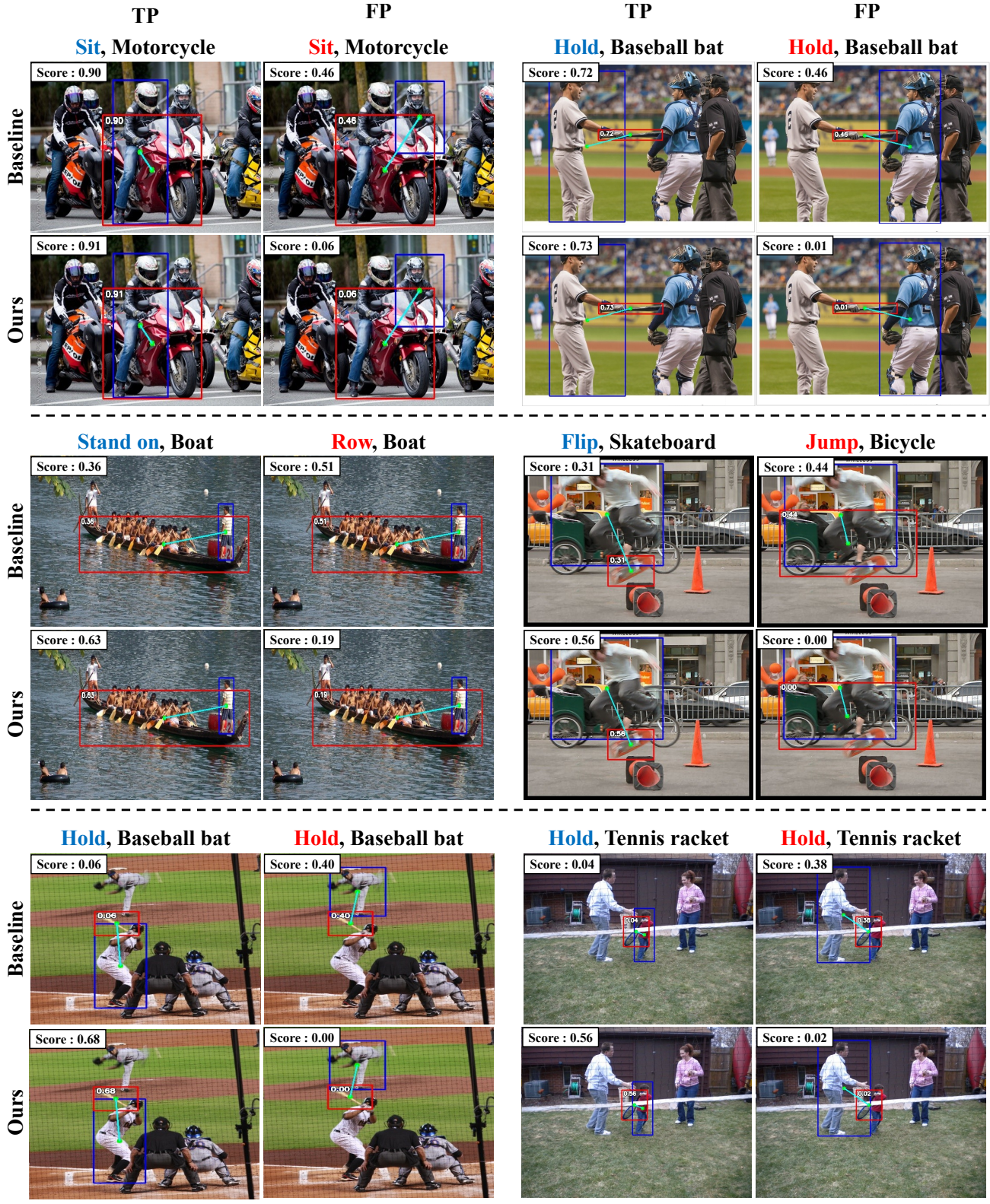

Figure S2. Qualitative comparisons between InCoM-Net and the baseline on HICO-DET. The TP column shows human–object pairs assigned the correct interaction label, while the FP column shows false-positive human–object pairs with incorrect labels.

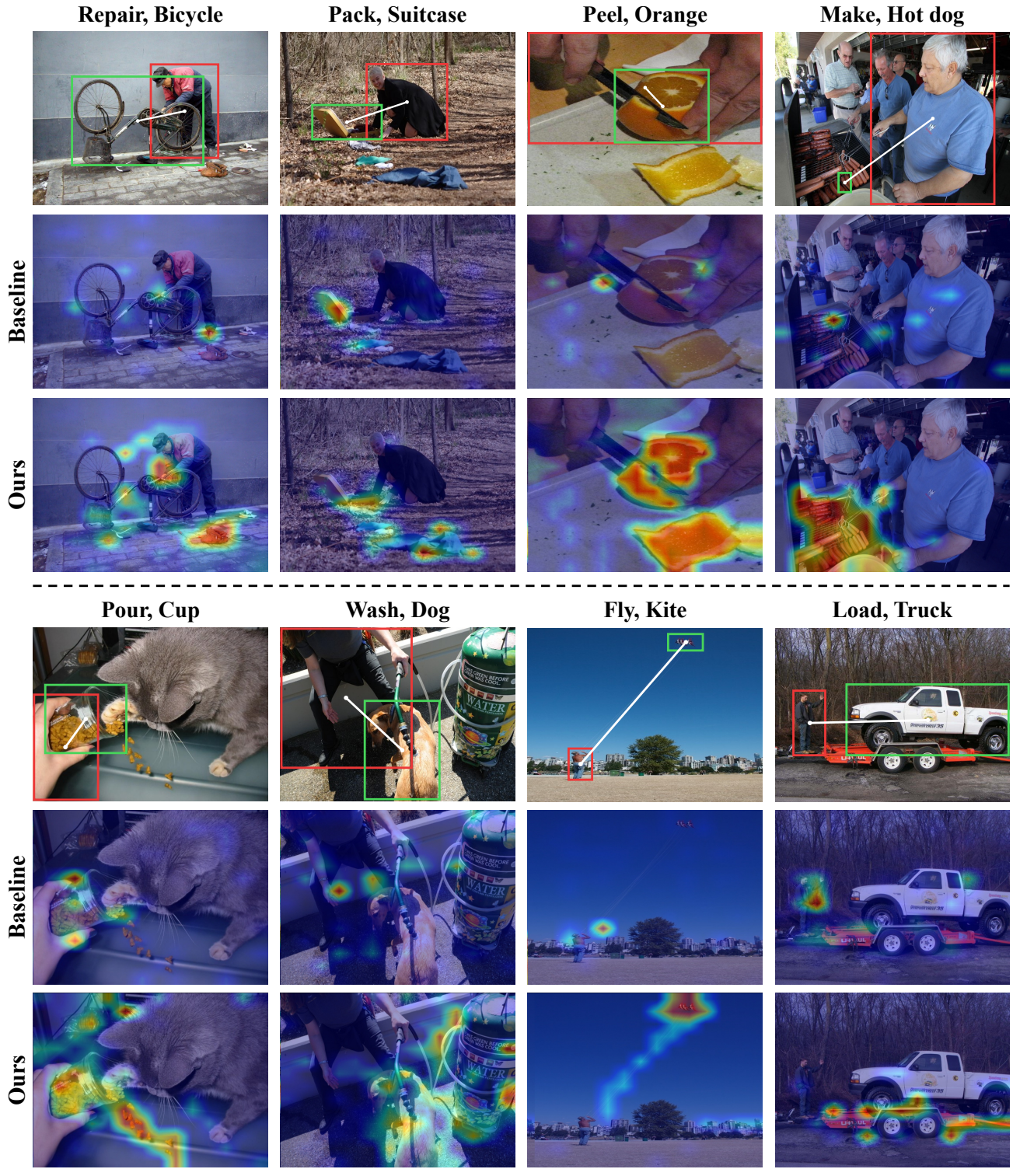

Figure S3. Comparison of activation maps between InCoM-Net and the baseline on HICO-DET.

References

[1] Yichao Cao, Qingfei Tang, Xiu Su, Song Chen, Shan You, Xiaobo Lu, and Chang Xu. Detecting any human-object interaction relationship: Universal HOI detector with spatial prompt learning on foundation models. In *Advances in Neural Information Processing Systems*, pages 739–751, 2023. 2

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229, 2020. 1

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 381–389, 2018. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[5] Mingda Jia, Liming Zhao, Ge Li, and Yun Zheng. Orchestrating the symphony of prompt distribution learning for human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3940–3948, 2025. 2

[6] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Locality-aware zero-shot human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20190–20200, 2025. 2

[7] Qinqian Lei, Bo Wang, and Robby T. Tan. EZ-HOI: Vlm adaptation via guided prompt learning for zero-shot HOI detection. In *Advances in Neural Information Processing Systems*, pages 55831–55857, 2024. 2

[8] Qinqian Lei, Bo Wang, and Robby T Tan. HOLa: Zero-shot HOI detection with low-rank decomposed VLM feature adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1825–1835, 2025. 2

[9] Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, and Yang Liu. Efficient adaptive human-object interaction detection with concept-guided memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6480–6490, 2023. 1, 2

[10] Ting Lei, Shaofeng Yin, Yuxin Peng, and Yang Liu. Exploring conditional multi-modal prompts for zero-shot HOI detection. In *Proceedings of the European Conference on Computer Vision*, pages 1–19, 2024. 1, 2

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2

[13] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. GEN-VLKT: Simplify association and enhance interaction understanding for HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 1, 2

[14] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. CLIP4HOI: Towards adapting CLIP for practical zero-shot HOI detection. In *Advances in Neural Information Processing Systems*, pages 45895–45906, 2023. 1, 2

[15] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. HOICLIP: Efficient knowledge transfer for HOI detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 1, 2

[16] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2

[17] Weihong Ren, Jinguo Luo, Weibo Jiang, Liangqiong Qu, Zhi Han, Jiandong Tian, and Honghai Liu. Learning self- and cross-triplet context clues for human-object interaction detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9760–9773, 2024. 2

[18] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1

[19] Guangzhi Wang, Yangyang Guo, Ziwei Xu, and Mohan Kankanhalli. Bilateral adaptation for human-object interaction detection with occlusion-robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27970–27980, 2024. 2

[20] Chanhyeong Yang, Taehoon Song, Jihwan Park, and Hyunwoo J Kim. Visual diversity and region-aware prompt learning for zero-shot HOI detection. In *Advances in Neural Information Processing Systems*, 2025. 2

[21] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 1

[22] Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting of human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10411–10421, 2023. 1