

## Summary of the Appendix

For a better understanding of the main paper, we provide additional details in this supplementary material, which is organized as follows:

- §A provides more implementation details of GroupHOI.
- §B offers more qualitative results.
- §C provides more experimental results.
- §D presents the pseudo code of GroupHOI.
- §E discusses the societal impact of this work.

## A More Implementation Details

We follow HOICLIP [1] in converting HOI triplets and object labels into textual descriptions to generate CLIP [2] text embeddings. Specifically, each HOI triplet <human, verb, object> is converted into a sentence like “A photo of a person [verb-ing] a/an [object]”. For “no-interaction” instances, we use “A photo of a person and a/an [object]”, and object labels are converted into “A photo of a/an [object]”. These textual descriptions are then processed by the pre-trained CLIP text encoder to obtain corresponding embeddings, which initialize the weights of the interaction classifier  $\mathcal{C}^a$  and object classifier  $\mathcal{C}^o$ . During training, these classifiers are fine-tuned with a small learning rate to adapt to the specific dataset. We employ the pre-trained CLIP visual encoder to extract visual features  $V_{clip}$ . These features, along with our encoded features  $V_e$ , are independently processed by separate interaction decoders. The resulting outputs are then fused to facilitate the final reasoning. We also introduce two key modifications based on HOICLIP [1]. **First**, to address the dimensional mismatch between positional embeddings (256) and the interaction decoder (768), we expand the positional embeddings by stacking them three times. **Second**, we replace the Focal Loss with Asymmetric Loss [3] to better handle the long-tail distribution of HOI categories.

## B More Qualitative Results

We further provide qualitative examples of our approach in Fig. S1. These results highlight GroupHOI’s robust performance in HOI detection across various scenes. Notably, our model effectively captures complex interaction patterns in scenarios involving group activities such as team sports (e.g., soccer or tennis). This capability enhances interaction prediction with mutual communication. We also present several failure examples of our model, primarily due to missed object detection, as seen in Fig. S1(c). Additionally, our model encounters challenges when dealing with highly ambiguous relations. For instance, in Fig. S1(s), GroupHOI fails to detect the *look at ball* between the girl at the center and the ball, which is disrupted by her surrounding teammates.

## C More Experiments

**Ablative Experiments.** We evaluate four strategies for measuring geometric proximity between entities, using intersection-over-union (IoU), center distance (CD), and global image features (IF). As shown in Table S1, combining IoU and CD consistently yields the best performance across all splits, indicating that both spatial cues complement each other effectively. In contrast, adding global image features slightly degrades performance, suggesting that they are not essential for proximity estimation.

Table S1: Analysis of geometric proximity measurement on HICO-DET [4] test (§C).

Measurement	Full	Rare	Non-Rare
<i>IOU only</i>	36.02	32.19	37.16
<i>CD only</i>	36.14	33.60	36.90
<i>IOU + CD</i>	<b>36.70</b>	<b>34.86</b>	<b>37.26</b>
<i>IOU + CD + IF</i>	36.21	33.45	36.97

**Efficiency Comparison.** Table S2 presents a comprehensive comparison of the model scalability (i.e., number of Parameters (Params), Floating Point Operations (FLOPs) and Frames Per Second (FPS)) for various HOI detection methods. As shown, GroupHOI achieves significant performance improvements over previous models while maintaining a comparable number of parameters. We also compare FLOPs and FPS with GEN-VLKT [16] and HOICLIP [1]. Despite a marginal reduction in inference speed relative to HOICLIP, GroupHOI has lower FLOPs and yields solid mAP

Table S2: Comparison of efficiency and performance on HICO-DET [4] test and V-COCO [5] test.

Method	Backbone	Params↓	FLOPs↓	FPS↑	Default			$AP_{role}^{S1}$	$AP_{role}^{S2}$
					Full	Rare	Non-Rare		
iCAN[6] <sub>[BMVC18]</sub>	R50	39.8	-	-	14.84	10.45	16.15	45.3	-
DRG[7] <sub>[ECCV20]</sub>	R50-FPN	46.1	-	-	19.26	17.74	19.71	51.0	-
PPDM[8] <sub>[CVPR20]</sub>	HG104	194.9	-	-	21.73	13.78	24.10	-	-
SCG[9] <sub>[ICCV21]</sub>	R50-FPN	53.9	-	-	31.33	24.72	33.31	54.2	60.9
HOTR[10] <sub>[CVPR21]</sub>	R50	51.2	-	-	25.10	17.34	27.42	55.2	64.4
HOITrans[11] <sub>[CVPR21]</sub>	R50	41.4	-	-	23.46	16.91	25.41	52.9	-
AS-Net[12] <sub>[CVPR21]</sub>	R50	52.5	-	-	28.87	24.25	33.14	53.9	-
QPIC[13] <sub>[CVPR21]</sub>	R50	41.9	-	-	29.07	21.85	31.23	58.8	61.0
CDN-S[14] <sub>[NeurIPS21]</sub>	R50	42.1	-	-	31.78	27.55	33.05	62.3	64.4
STIP[15] <sub>[CVPR22]</sub>	R50	50.4	-	-	32.22	28.15	33.43	<b>65.1</b>	<b>69.7</b>
GEN- <sub>[CVPR22]</sub>	R50	41.9	60.04	26.18	33.75	29.25	35.10	62.4	64.4
VLKT[16]									
HOICLIP[1] <sub>[CVPR23]</sub>	R50	66.1	104.68	19.57	34.59	31.12	35.74	63.5	64.8
CLIP4HOI[17] <sub>[NeurIPS23]</sub>	R50	71.2	-	-	35.33	33.95	35.74	-	66.3
ViPLO[18] <sub>[CVPR23]</sub>	ViT-B/32	118.2	-	-	34.95	33.83	35.28	60.9	66.6
GroupHOI (ours)	R50	79.2	83.67	16.42	<b>36.70</b>	<b>34.86</b>	<b>37.26</b>	65.0	66.0

improvements of **2.11/3.74/1.52** in HICO-DET [4], highlighting the effectiveness of our proposed framework.

## D Pseudo Code

The pseudo code for semantic and geometric group are given in Algorithm S1 and Algorithm S2.

**Algorithm S1:** Pseudo-code for Geometric Group in a PyTorch-like style.

```

"""
hs: output human/object embeddings from the instance decoder.
pos_embed: position embeddings for human/object queries.
coords: bounding boxes of humans/objects.
K_g: geometric group size.
"""
def Geometric_Group(hs, pos_embed, coords, K_g):
    # Formulate the spatial feature
    F_p = Cat([Square_distance(coords[:, None], coords[None, :]), IoU(coords[:, None], coords[None, :])])
    # Compute the proximity score
    S = Linear(F_p)
    # Select the topk neighbors
    knn_idx = TopK(S, K_g)

    # Compute the position encodings
    pos_enc = MLP(pos_embed - Gather(pos_embed, knn_idx))
    # Formulate query, key, and value
    q, k, v = Linear(hs), Linear(Gather(hs, knn_idx)), Linear(Gather(hs, knn_idx))
    # Compute dispatch matrix
    G = Softmax(q - k + pos_enc)
    # Aggregate geometric context
    C_g = Linear(G * (v + pos_enc))
    out = hs + C_g

    return out

```

**Algorithm S2:** Pseudo-code for Semantic Group in a PyTorch-like style.

```

"""
hs: interaction embeddings from the interaction decoder.
K_s: semantic group size.
"""
def Semantic_Group(hs, pos_embed, K_s):
    # Compute the similarity score
    S = CosineSimilarity(hs[:, None], hs[:, None])
    # Select the topk neighbors
    knn_idx = TopK(S, K_s)

    # Aggregate semantic context
    C_s = MLP(Max(MLP(hs[:, None], hs[:, None] - Gather(hs, knn_idx)), dim=-1))
    out = hs + C_s

    return out

```

## 52 **E Boarder Impact**

53 This work advances the recognition of human-object interactions in complex scenarios, particularly in  
54 scenes where small groups of people naturally form, which is a common occurrence in real-world set-  
55 tings. This capability holds significant promise for applications in collaborative robotics, autonomous  
56 systems, healthcare monitoring, among others. However, there are also potential downsides. Our  
57 method risks propagating irrelevant contextual information among entities that merely happen to be  
58 co-located but share no collective pattern, leading to “hallucinated” interaction predictions. Moreover,  
59 group-level clustering may inadvertently propagate systemic biases, particularly in scenarios requiring  
60 differential treatment of individuals within clusters (*e.g.*, unfair reward and punishment allocation in  
61 crowd behavior analysis). Hence, it is essential to rigorously consider legal regulations and integrate  
62 certain fairness constraints to avoid potential negative societal impacts.

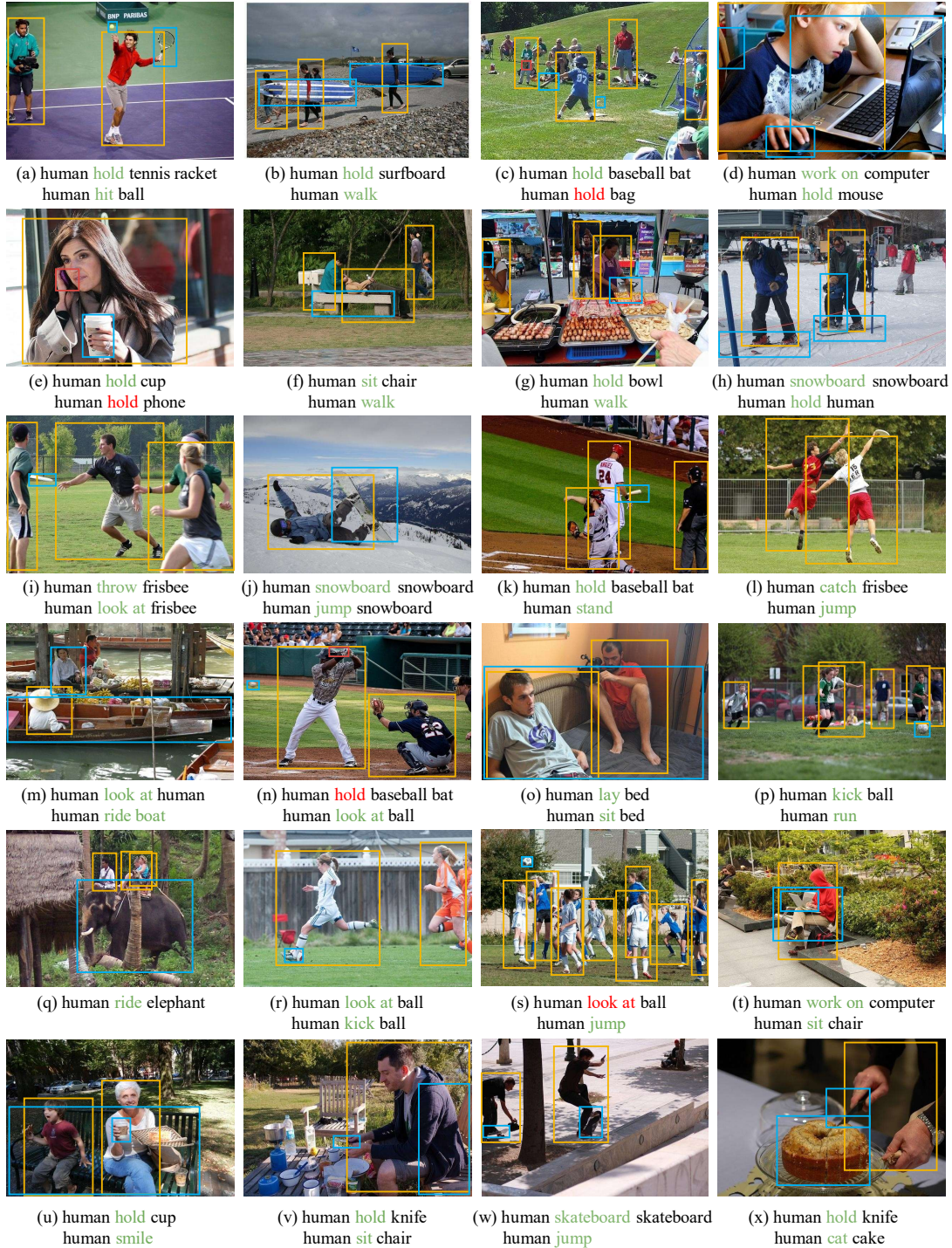


Figure S1: Visualization of GroupHOI results on V-COCO [5] test. Detected interactions are marked in **Green**, while missed interactions and objects are in **Red**.



## References

- [1] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 1, 2
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Aspell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1
- [3] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, 2021. 1
- [4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2
- [5] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 4
- [6] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2
- [7] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 2
- [8] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 2
- [9] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 2
- [10] Bumsoo Kim, Junhyun Lee, Jaewoo Kan, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 2
- [11] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 2
- [12] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 2
- [13] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 2
- [14] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*, 2021. 2
- [15] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 2
- [16] 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 *CVPR*, 2022. 1, 2
- [17] Yunyao Mao, Jiajun Deng, Wengang Zhou, Li Li, Yao Fang, and Houqiang Li. Clip4hoi: Towards adapting clip for practical zero-shot hoi detection. In *NeurIPS*, 2023. 2
- [18] 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. 2