No More Sibling Rivalry: Debiasing Human-Object Interaction Detection

Supplementary Material

In Supplementary Material, we provide additional information regarding,

A. Implementation Details

In this section, we provide detailed explanation for Figure 2, implementation details for the transformation $\mathrm{F}_{\mathrm{text}}(\cdot)$ and the human-object region mask attention (HOR mask), as mentioned in Section 3.1.2 and Section 4.3, respectively.

A.1. Detailed Explanation for Figure 2

"Sibling" Identification

We define "siblings" at two distinct levels:

- **Input-Level Siblings:** Within a single image, two HOI triplets `(human, verb, object)` are considered siblings if they share the same object or the same verb (interaction) category.
- **Output-Level Siblings:** Two distinct HOI categories are considered siblings if the cosine similarity between their respective classification head weights in the model is high. This similarity serves as a proxy for semantic or feature-space closeness.

Sub-figure Analysis

- **Figure 2(a) & (b):** These plots show the training mAP curves for representative sibling HOI pairs. The red lines highlight instances where the learning progress of one category appears to suppress the progress of its sibling, illustrating a competitive effect at both the input (a) and output (b) levels.
- **Figure 2(c):** This bar chart presents the average error rate, computed across all categories, under two conditions. It compares performance on ground-truth instances that have input-level siblings in the same image (red bar) against those that do not (yellow bar). A category's error rate is defined as the proportion of its samples that yield an mAP of 0 during evaluation.
- **Figure 2(d):** This scatter plot visualizes the relationship between a category's final performance and its inherent output-level similarity, focusing on "non-head classes". Crucially, the output-level similarity for a category (x-axis) is determined *before training* by calculating the average cosine similarity of its initial classification head weights to those of all other categories. This initial similarity is then plotted against the category's final mAP after training (y-axis).

A.2. Label Prompt in Details

In this subsection, we provide additional details about the transformation $\mathrm{F}_{\mathrm{text}}(\cdot)$. As mentioned in Section 3.1.2, for different inputs, we formalize them into distinct text descriptions, which are then fed into the CLIP [9] text encoder to obtain the corresponding text features. As shown in Table A1, we introduce the process of generating different text descriptions corresponding to different inputs following GEN-VLKT [6].

Input	Text Description
object (Object)	*A photo of a/an [Object].*
interaction (Verb)	*A photo of a person [Verb] something.*
HOI triplet	*A photo of a person [Verb-ing] a/an [Object].*

Table A1. Text descriptions corresponding to different inputs following GEN-VLKT [6].

A.3. Human-Object Region Mask Attention

In this subsection, we provide additional details about the human-object mask attention (HOR mask), as mentioned in Section 4.3. To guide interaction queries to focus on the correct human-object interaction regions within the image and avoid interference from similar sibling HOIs, we employ an attention mask for the union of each human-object pair. Specifically, for the standard mask attention [2] weights in the cross-attention of $\mathrm{Decoder}_{\mathrm{action}}$:

$$\mathbf{A} = \mathrm{softmax}\left(\mathcal{M} + W\right),$$
$$\mathcal{M}_i(x,y) = \begin{cases} 0 & \text{if } (x,y) \in \mathrm{region}(b_i^h, b_i^o) \\ -\infty & \text{otherwise} \end{cases},$$

where $\mathbf{A}$ represents the final attention weights, W denotes the initial attention weights computed by $\mathbf{Q}^a$ and $\mathbf{V}^i$, and $\mathcal{M}_i(x,y)$ is the mask applied to the attention weight of the

i-th interaction query at location (x, y) in the image features. Here, region is the bounding box that encapsulates both of the given bounding boxes b_i^h, b_i^o.

The attention mask effectively filters out most input level "Toxic Siblings" distractions, enhancing the recognition of HOI triplets. However, it cannot entirely eliminate interference caused by adjacent "Toxic Siblings" biases. For example, as illustrated in Fig 1 (a), a person standing behind a bench and another sitting on the same bench share an overlapping bounding box. To further mitigate input level "Toxic Siblings" bias, we propose the "contrastive-then-calibration" (C2C) debiasing objective. The effectiveness and necessity of the C2C debiasing objective are validated in Table 2.

B. More Experimental Results

In this section, we provide additional experimental results, including the performance gains of our method over the baseline on categories affected by class imbalance bias and the model's performance without the HOR mask module. Following LOGICHOI [5], we conducted further ablation experiments on V-COCO [3] to analyze the effects of loss weights and hyperparameters.

B.1. Evaluation Metrics

We use mean Average Precision (mAP) as the primary evaluation metric. A detection is considered a true positive (TP) if (1) the intersection over union (IoU) between the predicted human and object bounding boxes exceeds 0.5, and (2) the predicted action/interaction category is correct. For HICO-DET, mAP is computed in two settings: (i) the default setting, using all test images, and (ii) the Known Object setting, where average precision (AP) is computed for each object on a subset of images containing specific objects. For V-COCO, we use two evaluation settings: Scenario 1, where the detector must report an empty box when the interaction excludes an object, and Scenario 2, where object box detection can be ignored.

B.2. Performance on bias Categories

In Fig A1, we present a comparison of Average Precision (AP) for several HOI categories affected by "Toxic Siblings" bias. Due to this bias, the baseline model struggles to recognize certain HOI categories, such as "waving a bus" and "exiting a train" (see the line in Fig A1). The merge learning objective partially addresses this by leveraging shared features among similar HOI categories, enabling the recognition of some challenging categories, such as "waving a bus" (AP +55.4%). However, "Toxic Siblings" bias arises not only from long-tail issues but is often exacerbated by shared interactions or objects across categories, making it even more challenging to address. While

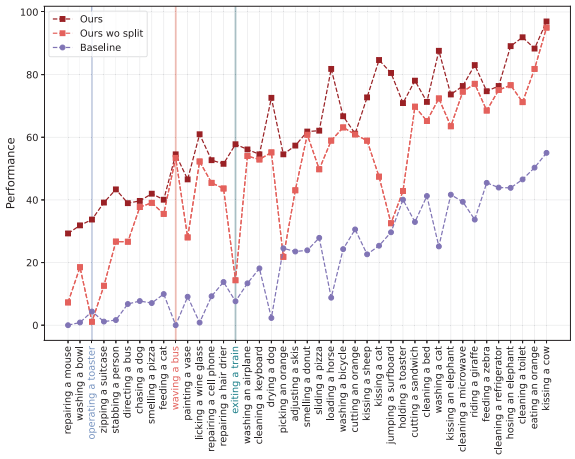

Figure A1. Experimental results of our method compared to the baseline on categories affected by class imbalance bias.

the merge learning objective facilitates the learning of generalized features, it is insufficient for resolving ambiguities among overly similar categories, such as "exiting a train", and may even intensify confusion between certain categories, such as "operating a toaster". To overcome this limitation, we introduce the split learning objective, which enables the model to more effectively distinguish between these closely related categories (see the line in Fig A1).

Overall, our combined "merge-then-split" learning objective effectively mitigates "Toxic Siblings" bias, resulting in a significant mAP improvement of +18.93% for HOI categories affected by class imbalance bias.

B.3. Experimental Result On Zero-shot

Method	Type	Unseen	Seen	Full
GEN-VLKT	UV	20.96	30.23	28.74
UniHOI	UV	26.05	36.78	34.68
Ours	UV	**30.47**	**41.96**	**38.05**
GEN-VLKT	UO	10.51	28.92	25.63
UniHOI	UO	19.72	34.76	31.56
Ours	UO	**31.01**	**41.59**	**37.23**
GEN-VLKT	NF	25.05	23.38	23.71
UniHOI	NF	28.45	32.63	31.79
SICHOI	NF	34.52	36.06	35.75
Ours	NF	**36.12**	**37.58**	**37.01**
GEN-VLKT	RF	21.36	32.91	30.56
UniHOI	RF	28.68	33.16	32.27
SICHOI	RF	34.24	41.58	40.11
Ours	RF	**35.48**	**42.91**	**40.96**

Table A2. Zero-Shot Results On HICO-DET

Table A2 serves as an extension of Table 2, illustrating the zero-shot performance in greater detail.

B.4. Applying C2C/M2S to HOICLIP

Method	HICO-DET (Default)		
	Full	Rare	Non-Rare
HOICLIP	34.54	30.71	35.70
HOICLIP (with C2C and M2S)	**43.11**	**42.69**	**43.58**
	+8.42	+11.57	+7.84

Table A3. Experimental results with applying C2C and M2S to HOICLIP on HICO-DET [1] under the default setting.

We integrate our C2C and M2S learning objectives into HOICLIP [7], as shown in Table A3. This increases HOICLIP's HICO-DET mAP to 43.11% (Full), 42.69% (Rare) and 43.58% (Non-Rare), achieving substantial gains of +8. 42% (Full), +11.57% (Rare) and +7.84% (Non-Rare), respectively.

B.5. Experimental Result without HOR Mask

Method	HICO-DET (Default)		
	Full	Rare	Non-Rare
Baseline (GEN-VLKT-s)	33.75	29.25	35.10
ViPLO [8]	34.95	33.83	35.28
ADA-CM [4]	38.40	37.52	38.66
BCOM [10]	39.34	39.90	39.17
Ours (w/o HOR mask)	42.01	41.55	42.21
	+2.67	+1.65	+3.04
Ours (with HOR mask)	**42.93**	**42.41**	**43.11**
	+3.59	+2.51	+3.94

Table A4. Experimental results with and without HOR mask on HICO-DET [1] under the default setting.

Table A4 presents the performance of our method without the HOR mask module, using ResNet-50 (R50) as the backbone and CLIP as the VLM. Notably, the proposed Contrastive-then-Calibration (C2C) and Merge-then-Split (M2S) learning objectives independently mitigate both types of bias, achieving an 8.26% improvement over the baseline and a performance of 42.01% (+2.67% compared to the state-of-the-art BCOM) on the HICO-Det dataset.

B.6. Ablation Studies on Hyper-Parameters

Ablation on loss weights. To evaluate the effectiveness of the proposed learning objectives, we conducted ablation experiments by varying their respective weights, as summarized in Table A5. Specifically, λ_1 denotes the weight of the original detection loss, while λ_2, λ_3, λ_4, and λ_5 represent the weights of the contrastive loss $\mathcal{L}_{\text{con}}$, calibration loss $\mathcal{L}_{\text{cal}}$, merge loss $\mathcal{L}_{\text{merge}}$, and split loss $\mathcal{L}_{\text{split}}$, respectively.

Ablation on loss weights	V-COCO AP^{S1}_{role}	V-COCO AP^{S2}_{role}
$\lambda_2 = 1, \lambda_3 = 0.5, \lambda_4 = 1, \lambda_5 = 0.5$, while $\lambda_1 =$		
0.1	69.3	71.8
1	**69.8**	**72.1**
10	69.5	72.0
$\lambda_1 = 1, \lambda_3 = 0.5, \lambda_4 = 1, \lambda_5 = 0.5$, while $\lambda_2 =$		
0.1	69.1	71.4
1	**69.8**	**72.1**
10	69.2	71.6
$\lambda_1 = 1, \lambda_2 = 1, \lambda_4 = 1, \lambda_5 = 0.5$, while $\lambda_3 =$		
0.05	69.5	71.7
0.5	**69.8**	**72.1**
5	69.4	71.8
$\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.5, \lambda_5 = 0.5$, while $\lambda_4 =$		
0.1	69.2	71.3
1	**69.8**	**72.1**
10	69.4	71.6
$\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 0.5, \lambda_4 = 1$, while $\lambda_5 =$		
0.05	69.6	71.8
0.5	**69.8**	**72.1**
5	69.4	71.7

Table A5. Ablation study on loss weights.

Ablation on # of k_1 and k_2	V-COCO AP^{S1}_{role}	V-COCO AP^{S2}_{role}
$k_1 = 1, k_2 = 10$	69.0	71.2
$k_1 = 2, k_2 = 10$	**69.8**	**72.1**
$k_1 = 3, k_2 = 10$	69.5	71.9
$k_1 = 5, k_2 = 10$	69.3	71.6

Table A6. Ablation study on k_1.

Ablation on k_1. As a supplement to Table 3 in the main text, we conduct further ablation experiments on the hyper-parameter k_1 mentioned in Section 3.3.2. The corresponding results are shown in Table A6.

C. Qualitative Results

C.1. Qualitative Comparisons with the Baseline

Qualitative Comparisons of input level Bias. Fig A2 and A3 present qualitative comparison results between the baseline [6] and our model under the influence of input level "Toxic Siblings" bias on HICO-DET [1]. From left to right, the images depict the negative triplet causing bias, the predictions of the baseline, our predictions, and the ground truth.

Fig A2 illustrates the first type of input level "Toxic Siblings" bias: interaction actions are misclassified due to the influence of nearby, similar negative triplets (see column 1), leading to incorrect predictions (see column 2). For instance, in the first row, the ground truth (see column 4) is "a person riding a train". However, due to the presence of a similar negative HOI "a person driving a train" (see column 1), which shares the same object "train" with the ground truth, the baseline misclassifies the interaction as "driving", even though the human and object are correctly predicted (see column 2). By identifying negative triplets in the surrounding context and differentiating them through the proposed C2C learning objective, our model accurately predicts the interaction (see column 3).

Fig A3 illustrates the second type of input level "Toxic Siblings" bias: unrelated and unpaired humans and objects are mistakenly predicted to have interactions due to the influence of nearby triplets (see column 1 and 4) that share the same human or object in the context. For example, in the first row, the baseline incorrectly matches a black motorcycle with a person wearing a blue vest (see column 2), influenced by the nearby interactions of "a person riding a motorcycle". The proposed C2C learning objective guides the model's attention to the correct spatial regions, mitigating the effects of input level "Toxic Siblings" bias and enabling accurate predictions (see column 3).

Qualitative Comparisons of Class Imbalance Bias. The qualitative results in Fig A4 demonstrate the effectiveness of our Merge-then-Split (M2S) learning objective in mitigating class imbalance bias. From top to bottom, the rows represent the baseline predictions, our model's predictions, and the ground truth. Due to output level "Toxic Siblings" bias, the baseline tends to misclassify long-tail HOI categories, such as "eating an orange" or "zipping a suitcase" into semantically similar sibling head classes, like "eating a cake" or "and a suitcase" (see rows 1 and 4). In contrast, our debiased model successfully predicts the correct long-tail HOI categories (see rows 2 and 5). This improvement is attributed to the M2S learning objective, which effectively reduces misclassifications for long-tail classes.

C.2. Qualitative Examples for Super Interaction-Object Categories

Tables A7 and A8 present selected category clusters obtained through the merge process. The clusters in Table A7 primarily encompass actions involving the manipulation of everyday objects using hands or mouth, while those in Table A8 focus on interactions that reflect the diverse forms of emotional contact, interaction, and specific actions between humans and animals or other individuals. Table A9 illustrates the multimodal interaction behaviors between humans and food objects through actions such as eating, inspecting, and smelling.

HOI categories in Super Category 1
holding a bowl, stirring a bowl
washing a bowl, licking a bowl
cutting a cake, holding a cake
drinking with a cup, holding a cup
pouring a cup, sipping a cup
washing a cup, holding a pizza
holding a remote, cutting a sandwich

Table A7. Super-category of manipulating everyday objects using hands or mouth.

HOI categories in Super Category 2
holding a cow, hugging a cow
kissing a cow, holding a dog
hosing a dog, hugging a dog
kissing a dog, washing a dog
holding a horse, hugging a horse
kissing a horse, washing a horse
holding a person, hugging a person
kissing a person, stabbing a person

Table A8. Super-category of interactions between humans and animals.

HOI categories in Super Category 3
eating an apple, inspecting an apple
smelling an apple, and an apple
eating a banana, inspecting a banana
smelling a banana, and a banana
eating a broccoli, smelling a broccoli
and a broccoli, eating a carrot
smelling a carrot, and a carrot
eating a donut, smelling a donut
and a donut, eating a hot dog
and a hot dog, eating an orange
inspecting an orange, and an orange

Table A9. Super-category of interaction between humans and food objects.

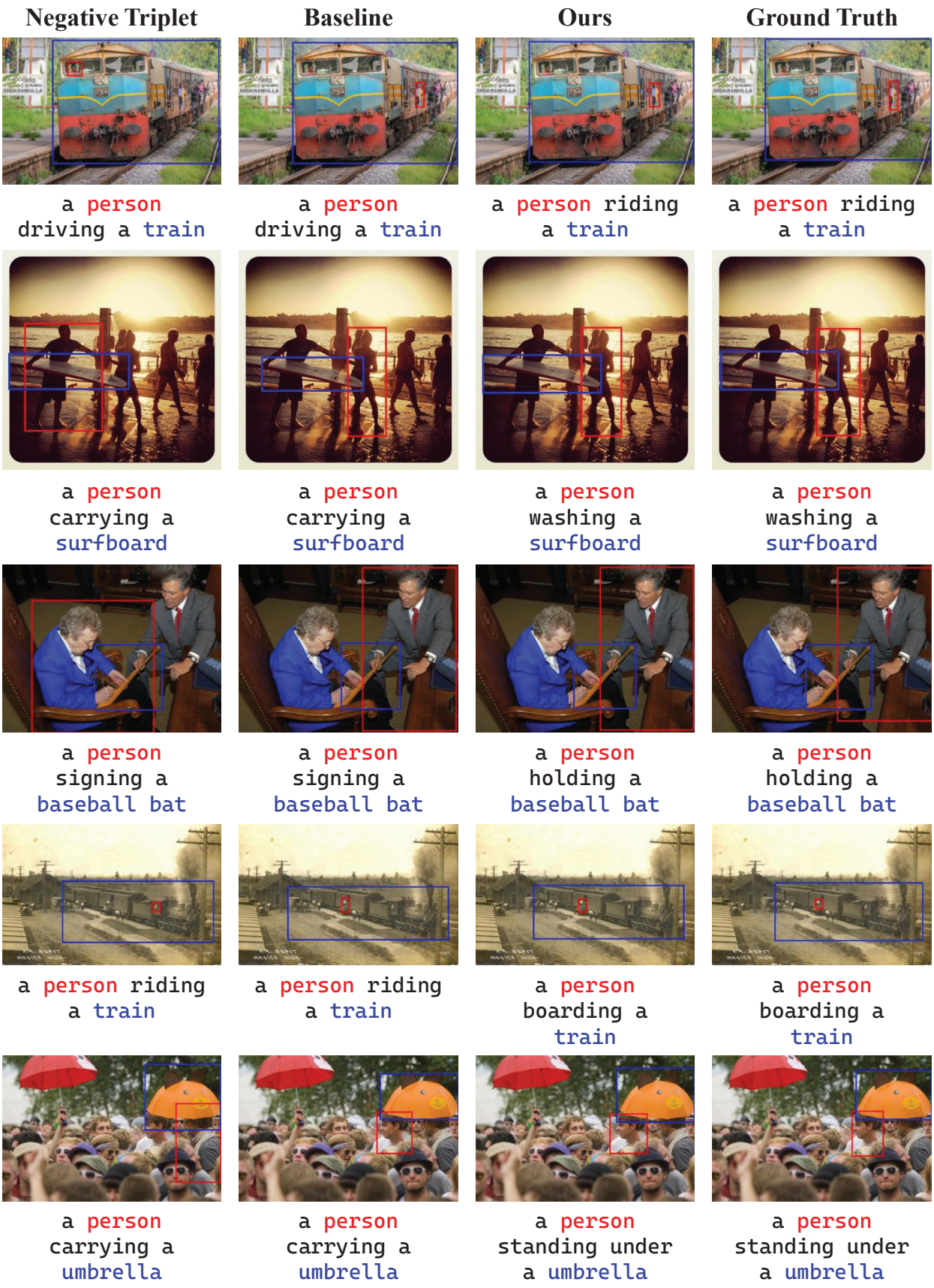

Figure A2. Qualitative comparisons of the first type of input level "Toxic Siblings" bias between our method and the baseline.

Figure A3. Qualitative comparisons of the second type of input level "Toxic Siblings" bias between our method and the baseline.

Figure A4. Qualitative comparisons of output level "Toxic Siblings" bias between our method and the baseline.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. 2018. 3

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 1

[3] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[4] 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. 3

[5] Liulei Li, Jianan Wei, Wenguan Wang, and Yi Yang. Neural-logic human-object interaction detection. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[6] 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, 3

[7] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 3

[8] Jeeseung Park, Jin-Woo Park, and Jong-Seok Lee. Viplo: Vision transformer based pose-conditioned self-loop graph for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17152–17162, 2023. 3

[9] 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. 2021. 1

[10] 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. 3