Locality-Aware Zero-Shot Human-Object Interaction Detection

Sanghyun Kim Deunsol Jung Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
{sanghuyn.kim, deunsol.jung ,mscho}@postech.ac.kr
http://cvlab.postech.ac.kr/research/LAIN

A. Appendix

In this supplementary material, we provide additional experimental results and analysis to support our method, including qualitative results.

A.1. Feature map analysis

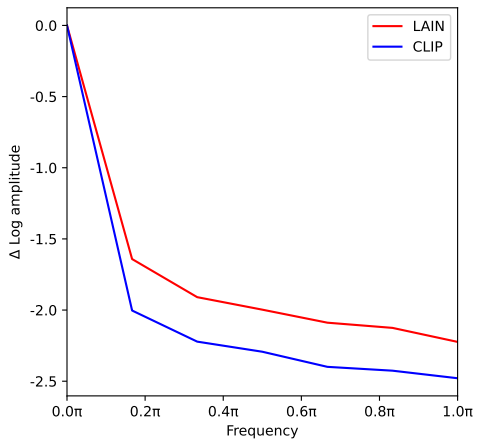

Figure S1. Relative log amplitudes of Fourier transformed feature maps.

The recent work [3, 4] shows that ViT structure learns low-frequency signals, which capture the global information, then high-frequency signals which capture the local spatial information, *i.e.*, fine-grained details. Following the previous work [3, 4], we visualize the result of the Fourier transform on feature maps to investigate the impact of the Locality Adapter. Specifically, we reshape the patch tokens from the output of the last layer into their original spatial structure and then apply the Fourier transform. As shown in Figure S1, the CLIP struggles to capture high-frequency signals. In contrast, we observe that our LAIN captures more high-frequency signals compared to the CLIP. These results indicate the effectiveness of LAIN in capturing fine-grained details for an HO pair by incorporating locality awareness, thus leading to a performance increase in the zero-shot HOI detection where the fine-grained details are crucial.

A.2. Ablation study on zero-shot settings

Setting	LA	IA	Unseen	Seen	Full
RF-UC	-	-	28.15	32.99	32.02
	-	✓	30.96	35.01	34.20
	✓	-	29.44	34.90	33.80
	✓	✓	**31.83**	**35.06**	**34.41**
NF-UC	-	-	34.78	30.20	31.11
	-	✓	35.93	31.97	32.77
	✓	-	35.65	31.43	32.29
	✓	✓	**36.41**	**32.44**	**33.23**
UO	-	-	34.26	31.33	31.84
	-	✓	36.82	33.13	33.75
	✓	-	36.39	32.47	33.12
	✓	✓	**37.88**	**33.55**	**34.27**

Table S1. Ablation studies on each adapter under various zero-shot settings. The 'Setting' column indicates different evaluation setups: RF-UC, NF-UC, and UO. LA and IA denote locality and interaction adapter, respectively.

To further validate the effectiveness of our approach, we conduct a comprehensive ablation study across various zero-shot settings. Similar to the ablation study on UV setting in Table 3 in the main paper, we gradually add each component and report the results under the RF-UC, NF-UC, and UO settings in Table S1. Similar to the UV setting, both the Locality Adapter (LA) and the Interaction Adapter (IA) independently improve performance for both unseen and seen classes in the RF-UC, NF-UC, and UO settings. When both LA and IA are applied together, they provide further improvement by jointly enhancing locality and interaction awareness in the CLIP representation. These demonstrate the effectiveness of the proposed method and its generalization ability across seen and unseen classes.

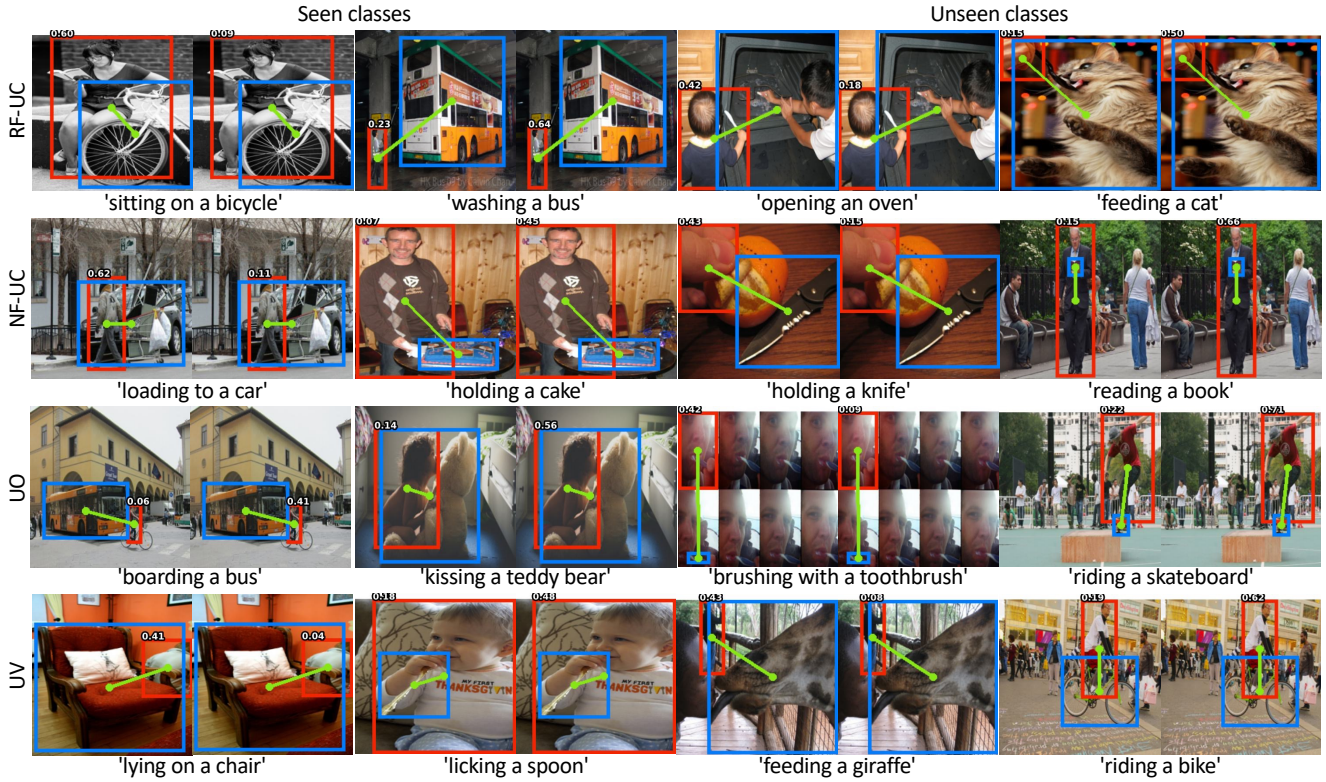

Figure S2. Qualitative comparison of non-interactive pairs between our model and the baseline, without LA and IA, on the HICO-DET under various zero-shot settings (RF-UC, NF-UC, UO, UV). We represent a human with a red box and an object with a blue box, along with similarity scores with an HO pair and text embedding. The first and second columns visualize interactions for seen classes only, while the third and fourth columns focus exclusively on unseen classes.

$\mathbb{K}$	Unseen	Seen	Full
$\emptyset$	24.88	31.06	30.19
{1}	26.17	**33.84**	**32.76**
{3}	27.10	32.10	31.41
{5}	27.26	32.33	31.63
{3,5}	**27.71**	32.55	31.95
{1,3,5}	27.56	32.42	31.74

Table S2. The impact of combining different kernel sizes under the UV settings.

A.3. The impact of combining different kernel sizes.

To investigate the impact of the combination of different kernel sizes, we conduct experiments varying the combination of kernel size in LA as shown in Table S2. When utilizing $\{1\}$, *i.e.*, without locality awareness, the model tends to overfit to the seen classes, as it struggles to capture the fine-grained details of the object. In contrast, using $\{3\}$ or $\{5\}$ demonstrates better generalization to unseen classes by providing locality awareness to CLIP representation. This result indicates that considering locality awareness is crucial for zero-shot HOI detection. We observe that combining different kernel sizes yields further performance improvements, with the best results achieved when using $\{3, 5\}$ by providing the local information across multiple spatial scales.

A.4. Additional qualitative results

We provide additional qualitative results under the various zero-shot settings in Figures S2 and S3 to show the model's effectiveness.

In Figure S2, we present image pairs to investigate the impact of incorporating IA and LA: the left displays the results of the baseline model without these components, while the right illustrates the results of LAIN. As shown in the first column of Figure S2, we observe that our LAIN effectively distinguishes interactive pairs. For example, in the RF-UC setting, the baseline model assigns a high similarity score to 'sitting on a bicycle' for a scenario where a person is actually sitting on a bench, indicating confusion in distinguishing interactive pairs. In contrast, our model assigns a significantly lower score to non-interactive pairs. Similarly, as shown in the second column of Figure S2, LAIN assigns higher similarity scores to the correct interactive pairs, demonstrating its enhanced understanding of the

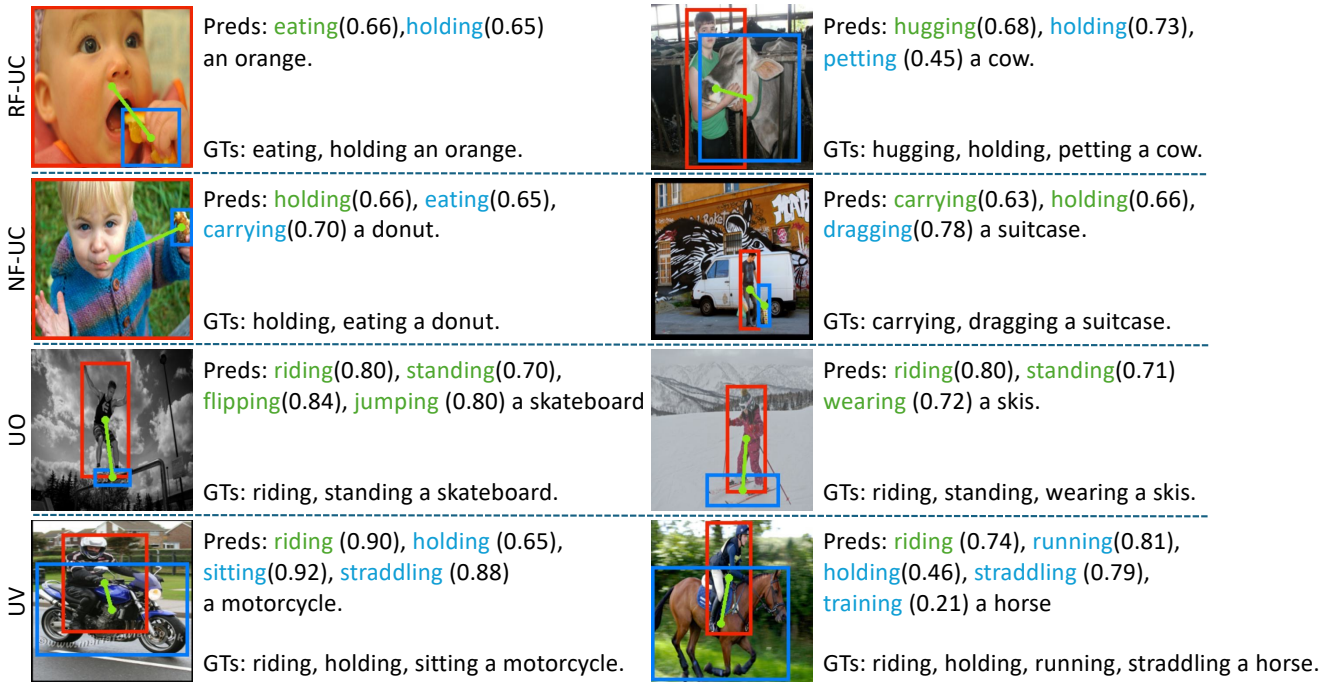

Figure S3. Qualitative results comparing our model's predictions (denoted as "Preds") with the ground truths (denoted as "GTs") are presented. Seen classes are highlighted in blue, while unseen classes are highlighted in green, along with their corresponding similarity scores. Humans are represented by red bounding boxes, and objects are represented by blue bounding boxes.

interactive pairs by providing fine-grained details about an HO pair with locality and interaction awareness. Similar results were also observed not only across different zero-shot settings but also for unseen classes presented in the third and fourth columns of Figure S2, further confirming its strong generalization ability. These results indicate the importance of integrating locality and interaction awareness into CLIP representations, enabling the model to capture fine-grained details of HO pairs for zero-shot HOI detection.

Figure S3 shows qualitative results comparing LAIN's predictions ('Preds') with the ground truths ('GTs'), along with similarity scores, under various zero-shot settings. We observe that LAIN accurately identifies interactions for both seen (highlighted in blue) and unseen classes (highlighted in green) in the HO pairs. Furthermore, LAIN identifies interactions for both seen and unseen classes that are absent in the ground truths. For example, in the NF-UC setting, LAIN successfully identifies the seen class 'carrying,' which is not annotated in the ground truths. Similarly, in the UO setting, LAIN accurately predicts the unseen classes 'flipping' and 'jumping,' demonstrating its ability to generalize to interactions beyond those explicitly annotated.

A.5. Implementation details

In this section, we provide additional implementation details about LAIN to facilitate reproducibility and understanding. We adopt the pre-trained DETR [1] as the detector, utilizing ResNet-50 as the CNN backbone. Following previous work [7], we discard detection results with a confidence score under 0.2 and retain between 3 and 15 instances each of humans and objects, choosing those with the highest confidence scores. In all experiments, we leverage the ViT-B/16 backbone of CLIP. During the training, the CLIP and the detector are kept frozen except for LA, IA, and learnable tokens inserted in the text description. The initial learning rate is set to 1e-3 and is multiplied by 0.1 after the first 10 epochs. The model is trained for 20 epochs. We optimize our network using AdamW [2]. All experiments are trained with a batch size of 8 on a single RTX 4090 GPU. The dimensionalities of the key components in our model are as follows: D_{det}, D_{clip}, and D_a are 512, 768, and 32, respectively. For the feed-forward network (FFN), we use a simple two-layer MLP with ReLU activation to implement Eq. 2, while a fully-connected layer is applied for others. The cross-attention layer is composed of two attention heads. Following prior work [6, 7], The value of λ, used to suppress overconfidence in object predictions, is set to 1.0 during training and 2.8 during inference. γ_{PA} and γ_{RA} are initially set to zero to prevent drastic changes in the CLIP representations during the early stages of training.

Implementation details for CLIP baseline. We design the HOI detection process using CLIP as follows. Given an image, we first detect object candidates using a pre-trained object detector, DETR [1]. For all valid HO pairs (u, v), we

create a union box of the human b_u and object box b_v, then crop the union box region from the image to classify the interaction label. The cropped images are fed into the pre-trained CLIP [5] visual encoder to obtain image features. Then, we use the template 'A photo of a person [verb]-ing a [object]' to obtain text descriptions for the 600 HOI categories in HICO-DET. The text descriptions are fed into the pre-trained CLIP text encoder to obtain text features. We then calculate the similarity between the image and text features. Finally, we compute the final score for each HOI category for evaluation. The CLIP baseline is not fine-tuned and is utilized with the original parameters provided by [5], being evaluated in a training-free manner.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[3] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022. 1

[4] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1

[5] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[6] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 3

[7] 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. 3