Supplementary Materials
Interaction Compass: Multi-Label Zero-Shot Learning of Human-Object Interactions via Spatial Relations

Dat Huynh
Northeastern University
huynh.dat@northeastern.edu

Ehsan Elhamifar
Northeastern University
e.elhamifar@northeastern.edu

1. Evaluation Metrics

1.1. Recognition

To compute the F1 score at top-10 predictions, we select 10 predictions with the highest scores as positive predictions in each image and compare these predictions with the ground-truth labels. We compute precision and recall for all interaction labels in a set $\mathcal{C}$ as,

$$P \triangleq \frac{\sum_{(a,o)\in\mathcal{C}} N^t_{a,o}}{\sum_{(a,o)\in\mathcal{C}} N^p_{a,o}}, \quad R \triangleq \frac{\sum_{(a,o)\in\mathcal{C}} N^t_{a,o}}{\sum_{(a,o)\in\mathcal{C}} N_{a,o}}, \tag{1}$$

where $N^t_{a,o}$ is the number of true positive for an interaction label (a, o), $N^p_{a,o}$ is the number of positive predictions for the same interaction label, and $N_{a,o}$ is the number of images containing the interaction label (a, o) according to ground-truth. F1 score is the harmonic mean between precision and recall and is defined as,

$$F1 = \frac{2PR}{P + R}. \tag{2}$$

Notice that precision can be computed from the reported recall and F1 scores in the main paper based on (2).

To measure mAP score, we compute the Average Precision of each label as

$$AP_{a,o} \triangleq \frac{1}{N_{a,o}} \sum_{k=1}^{N} \text{Precision}_k(a, o) \cdot \text{rel}_k(a, o), \tag{3}$$

where, $\text{Precision}_k(a, o)$ is the precision for the interaction label (a, o) when retrieving k best predictions and $\text{rel}_k(a, o)$ is the relevance indicator function that is 1 iff the interaction label (a, o) is in the ground-truth of the image at rank k. The mean Average Precision (mAP) is defined as

$$mAP = 1/|\mathcal{C}| \sum_{(a,o)\in\mathcal{C}} AP_{a,o}, \tag{4}$$

where $|\mathcal{C}|$ is the number of interaction labels.

Here, we define $\mathcal{C} = \mathit{A2} \cup \mathit{B1}$ ($\mathit{A2} \cup \mathit{B1} \cup \mathit{B2}$) for mAP and F1 scores of unseen interaction labels in $\mathit{A1} \cup \mathit{B2}$ ($\mathit{A1}$) setting. For performances on all interaction labels, we set $\mathcal{C} = \mathit{A1} \cup \mathit{A2} \cup \mathit{B1} \cup \mathit{B2}$ in both settings.

In the both HICO and Visual Genome datasets, due to the large number of interactions, each image has unannotated interaction labels. For F1 and mAP scores, we treat missing interaction annotations as negative labels, similar to [1, 2].

1.2. Localization

To quantify the localization performances, we follow [3] to measure whether action and object locations $\boldsymbol{l}_a, \boldsymbol{l}_o$ are correctly estimated within their corresponding ground-truth bounding boxes. First, we rescale their predictions from $[1, 17] \times [1, 17]$ coordinates to the original image ranges. If these predictions are within the ground-truth bounding-box regions and among the top-10 predictions, we consider these as true positive, otherwise they are false positive. Finally, we rank all interaction predictions according to their confidences and compute Average Precision to measure whether true positive predictions are ranked higher than false positive predictions. Thus, high Average Precision is achieved when a model correctly recognizes, with high confidences, and localizes present interaction labels.

In the main paper, we focus on analyzing the localization performances of each interaction label on only images having the target label following [4]. We also report the task of localizing each interaction label over all images as shown in Table 2. Since localization on all images requires a model to not only correctly attend to the bounding boxes of actions and objects within top-10 predictions but also rank images of target labels higher than images of other labels, this task is more challenging than localizing among images of target labels, leading to lower precision compared to the reported performances in the main paper. Overall, our method still surpasses other baselines at localizing actions, objects and action-object pairs in $\mathit{A1} \cup \mathit{B2}$ and $\mathit{A1}$ settings.

Method	Seen Interactions	HICO-DET				
		A1	*A2*	*B1*	*B2*	All
DEVISE	*A1* ∪ *B2*	20.4	10.3	10.9	25.3	16.9
	A1	18.5	2.2	6.3	1.7	8.1
Fast0Tag	*A1* ∪ *B2*	31.1	17.9	21.3	32.9	26.2
	A1	29.0	**4.5**	15.8	3.6	14.8
LESA	*A1* ∪ *B2*	34.1	19.8	23.9	37.2	28.3
	A1	31.3	2.5	21.4	1.7	16.2
Dual Attention	*A1* ∪ *B2*	29.6	17.3	20.4	34.6	25.8
	A1	29.1	3.8	17.5	3.4	15.1
Combined Attention	*A1* ∪ *B2*	26.1	14.1	14.5	33.0	22.1
	A1	25.6	4.0	13.7	3.5	13.0
ICompass (Ours)	*A1* ∪ *B2*	**33.4**	**23.2**	**25.3**	**39.0**	**30.4**
	A1	**32.0**	4.5	**20.9**	**4.0**	**17.2**

Table 1: Performances of zero-shot HOI recognition (mAP) on HICO-DET dataset.

2. Recognition Performances on HICO-DET

In addition to localization performances, we also present recognition performances on different interaction label sets on HICO-DET in Table 1.

Overall, recognition performances follow similar trends as localization performances where our method significantly outperforms the state of the art in both *A1* ∪ *B2* and *A1* settings. While HICO-DET has additional bounding-box annotations compared to HICO, these datasets share the same set of images and image-level labels, leading to similar recognition performances on all interactions.

3. Visualization

3.1. Visual Queries

Figure 1 visualizes the word embeddings from the pre-trained GloVe model and our proposed visual queries for a subset of actions and objects on HICO dataset in *A1* ∪ *B2* setting. We apply t-SNE [7] to project word embeddings and visual queries into 2D space for visualization.

Notice that the original word embeddings form two distinct clusters of actions and objects, thus when used for recognition, these queries cannot leverage semantic similarity between actions and objects. Our method learns to modify the word embeddings into visual queries that reflect the relationship between actions and objects in terms of affordance, e.g., "jump" and "flip" can be performed on "skis", thus these queries are close together. This is due to our ability to share knowledge among actions and objects by using the same function $r(\cdot)$ to construct their visual queries.

3.2. Interaction Localization

Figure 3 shows that our method is capable of not only recognizing but also localizing multiple interaction labels in each image. Moreover, cross attention can correct localization errors for objects via action information. Although the object attention mis-localizes "surfboard" because of its small visual appearance, we observe that cross attention correctly attends to "surfboard" regions by leveraging the spatial relation from "ride" action.

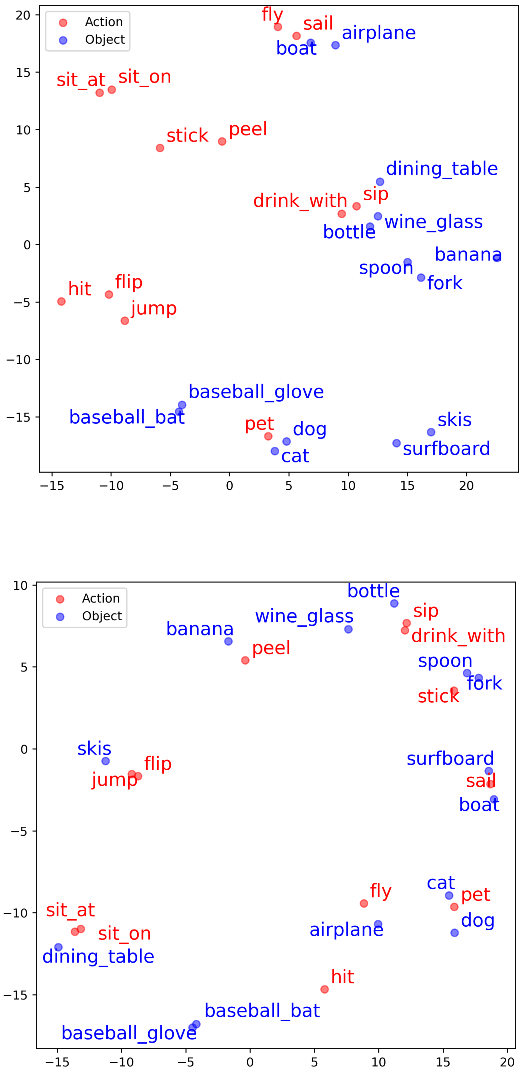

Figure 1: Visualization of word embeddings from the GloVe model (top) and visual queries learned by our method (bottom).

3.3. Relational Direction Statistic

Fig. 2 shows the percentages of samples in each action having the relational direction pointing upward from

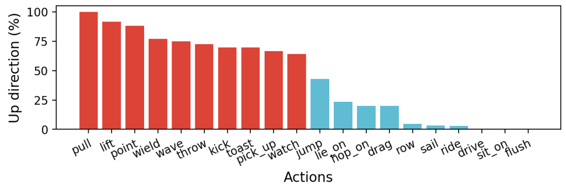

Figure 2: Percentage of samples from 10 actions with most/least upward relational directions from actions to objects in HICO-DET.

Method	Seen Interactions	Action					Object					Action & Object				
		A1	*A2*	*B1*	*B2*	All	*A1*	*A2*	*B1*	*B2*	All	*A1*	*A2*	*B1*	*B2*	All
DEVISE [2] + CAM [5]	$A1 \cup B2$	5.3	1.3	2.4	5.3	3.6	6.6	2.7	4.0	9.2	5.6	2.5	0.8	1.3	3.6	2.1
	$A1$	3.0	0.4	1.0	0.1	1.1	3.9	0.5	1.7	0.1	1.5	1.6	0.3	0.5	0.0	0.6
LESA [6]	$A1 \cup B2$	13.4	5.7	7.8	12.2	9.8	13.5	7.5	8.6	13.7	10.8	7.7	4.5	4.8	8.3	6.3
	$A1$	12.4	0.4	6.4	0.3	4.9	12.3	0.4	6.7	**0.4**	4.9	6.6	0.2	3.9	0.3	2.8
Dual Attention	$A1 \cup B2$	11.5	6.1	7.2	12.2	9.3	11.6	7.6	8.0	13.0	10.1	5.5	5.0	4.5	8.3	5.8
	$A1$	11.2	**0.9**	4.9	0.1	4.3	12.0	0.5	4.8	0.0	4.3	5.8	0.5	2.8	0.0	2.3
Combined Attention	$A1 \cup B2$	10.0	2.2	5.6	10.8	7.1	6.8	2.8	3.5	6.8	5.0	3.1	1.1	1.9	4.0	2.5
	$A1$	8.9	0.7	3.6	0.2	3.4	6.4	0.3	2.4	0.0	2.3	2.4	0.2	1.4	0.0	1.0
ICompass (Ours)	$A1 \cup B2$	**14.8**	**9.2**	**12.0**	**15.3**	**12.8**	**15.5**	**10.8**	**11.3**	**18.2**	**13.9**	**8.4**	**7.5**	**6.9**	**11.4**	**8.5**
	$A1$	**12.6**	0.7	**7.7**	**0.3**	**5.3**	**14.9**	**0.8**	**7.9**	0.3	**6.0**	**7.2**	**0.5**	**4.8**	**0.3**	**3.2**

Table 2: Performances of zero-shot HOI localization (mAP) on all images in HICO-DET dataset.

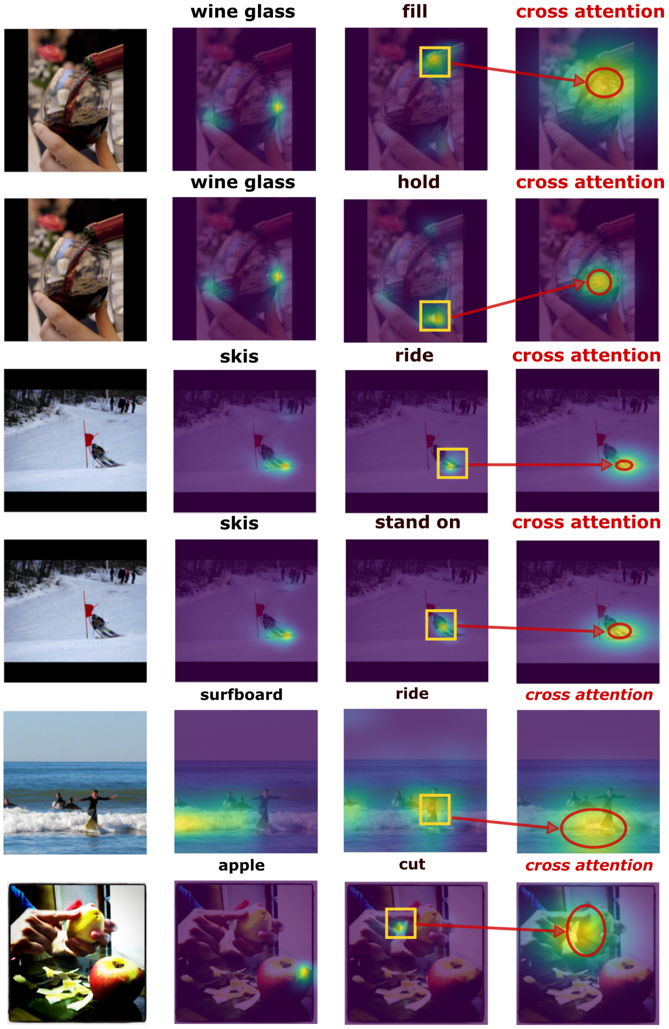

Figure 3: Visualization of action/object attention maps and cross-attention predictions in $A1 \cup B2$ setting on HICO-DET. Yellow boxes indicate regions with highest attention weights and red ellipses highlight most probable object locations.

actions to objects. We observe that upward actions such as ‘lift’ or ‘throw’ often result in object locations higher than action locations while ‘sit on’ or ‘lie on’ actions mostly follow downward relational directions. Please notice that we demonstrated the effectiveness of the estimated relational direction in Fig. 4 (left) in main paper, which shows cross attention significantly outperforms random guesses.

References

[1] Y. W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, “Hico: A benchmark for recognizing human-object interactions in images,” *IEEE International Conference on Computer Vision*, 2015. 1

[2] K. Kato, Y. Li, and A. Gupta, “Compositional learning for human object interaction,” *European Conference on Computer*

Vision, 2018. 1, 3

[3] I. L. M. Oquab, L. Bottou and J. Sivic, “Is object localization for free? – weakly-supervised learning with convolutional neural networks,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1

[4] T. Deselaers, B. Alexe, and V. Ferrari, “Localizing objects while learning their appearance,” *European Conference on Computer Vision*, 2010. 1

[5] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[6] D. Huynh and E. Elhamifar, “A shared multi-attention framework for multi-label zero-shot learning,” *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[7] L. V. D. Maaten and G. E. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, 2008. 2