Supplementary Material:
HICO: A Benchmark for Recognizing Human-Object Interactions in Images

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng
Computer Science and Engineering, University of Michigan, Ann Arbor
{ywchao,wangzhan,daranday,jiaxuan,jiadeng}@umich.edu

We present the training details of HOI classifiers, extra information on HICO dataset, and evaluation of 80 object detectors in this supplementary matrial.

1. Training HOI Classifiers

To study the usage of semanitc knowledge, we first train a set of basic classifiers. We then combine the output scores of different basic classifiers to explore the role of different semantic knowledge (Tab. 7 and Tab. 8 in the paper). All the trainings are done on the HICO training set.

Basic Classifiers We train three types of basic classifiers for each feature representation: 1) verb-object (VO) pairs, 2) verbs (V), and 3) objects (O). This gives us 600, 117, and 80 classifiers, respectively. Each classifier is trained using a linear SVM [1]. We perform 5-fold cross-validation to determine the paramter C.

Combined Classifiers For each HOI category, we train a combined classifier by linearly combining the output scores of a set of basic classifiers:

$$\overline{\phi}_j = \sum_{i \in S} w_i \phi_{ij} \tag{1}$$

ϕ_{ij} denotes the output score of basic classifier i on image sample j, and w_i denotes the learned weight. The set S is determined by the choice of semantic knowledge. For example, we take $S = \{\mathrm{V}, \mathrm{O}, \mathrm{VO}\}$ for training a V+O+VO classifier. The weights w_i are learned by maximizing the training AP using a grid search over $[0, 1]^{|S|}$. Due to the reuse of training data, applying the output scores from the trained basic classifers will lead to over-fitting. Instead, we assign ϕ_{ij} using the output scores generated during cross-validation, i.e. ϕ_{ij} is the output score from the model trained on all the splits not containing sample j.

2. Co-occurences of HOIs

As mentioned in the paper (Sec. 4.4), we exploit the co-occurences of HOIs to help the recognition of individual HOIs. For example, "eating a hot dog" often co-occurs with "holding a hot dog", but not "riding a bicycle". Thus a high confidence in "eating a hot dog" might indicate the presence of "holding a hot dog", but not "riding a bicycle". In our experiment, we first discover a fix set of co-occuring HOIs for each individual HOI. We then include the output scores from the basic classifiers of co-occuring HOIs to train a combined classifier.

To measure the level of co-occurence of two HOIs, we adopt the normalized co-occurences as in [3]. For HOI class i, the normalized co-occurence of HOI class j is defined by

$$s_{ij} = \frac{c_{ij}}{c_i} \tag{2}$$

where c_{ij} is the number of images labeled positive for both HOI class i and j, and c_i is the number of images labeled positive for HOI class i. To apply the knowledge of co-occurences, we compute the co-occurences of HOIs for each object category separately using the training annotations (Fig. 7 in the paper). We define HOI class j as a co-occuring HOI of HOI class i if $s_{ij} > 0.5$ and $i \neq j$. To train combined classifiers (VO+coocc & V+O+VO+coocc), we include only the basic classifiers of co-occuring HOIs and ignore the non-co-occuring ones.

3. HOI Categories of HICO

The complete list of HICO's 600 HOI categories, along with the 117 actions (verb senses) and 80 objects, is shown in the matrix in Fig. 1. Each row (column) corresponds to an object (verb). A blue entry indicates the presence of an HOI. Fig. 2 shows the number of positives for all 600 HOI categories. The long tail distribution highlights the presence of dominant and rare HOIs. More examples of HICO images and HOI annotations are given in Fig. 3

4. Training R-CNN Object Detectors

Our Human-Object CNN takes in detection heatmaps of 80 object categories. To obtain the detectors for 80 objects, we first take the off-the-shelf R-CNN detectors [2]

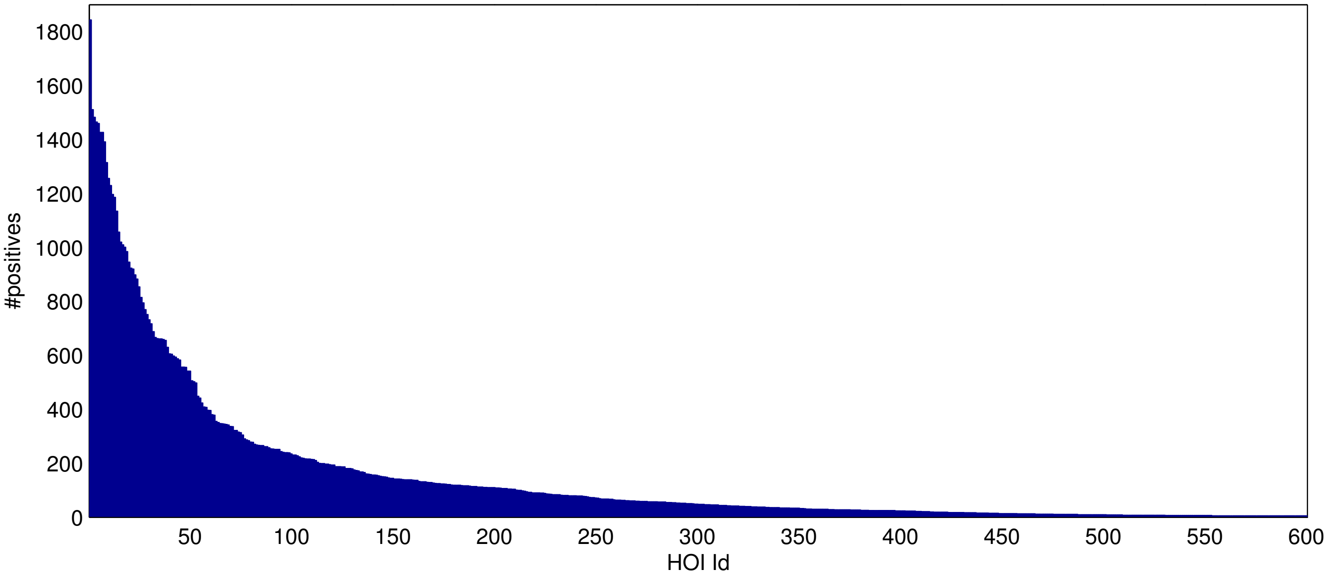

Figure 2: Number of positives per HOI category. The long tail distribution highlights the presence of dominant and rare HOI categories.

	truck	tf light	hydrant	sp sign	pk meter	bench	elephant	bear	zebra	giraffe	backpack	umbrella	handbag	tie	suitcase
R-CNN fc_7	22.9	13.2	56.0	61.4	23.2	10.8	48.7	50.3	56.9	56.7	9.7	19.1	1.9	16.1	13.4

	frisbee	skis	snowbd	sp ball	kite	bb bat	bb glove	skatebd	surfbd	racket	wn glass	cup	fork	knife	spoon
R-CNN fc_7	23.7	11.6	11.2	17.2	14.7	13.4	18.4	19.6	14.6	28.2	13.3	15.0	9.4	9.2	9.6

	bowl	banana	apple	sandwich	orange	broccoli	carrot	hot dog	pizza	donut	cake	bed	toilet	laptop	mouse
R-CNN fc_7	23.4	14.6	13.1	24.6	20.3	16.8	9.4	24.4	41.8	17.5	11.5	27.4	39.4	35.1	23.6

	remote	keyboard	phone	microwave	oven	toaster	sink	fridg	book	clock	vase	scissors	td bear	hr drier	tbrush
R-CNN fc_7	8.1	21.5	13.4	27.3	17.5	11.1	15.1	24.4	1.0	48.0	17.6	15.8	36.3	0.3	0.6

Table 1: Detection average precision (%) of 60 non-PASCAL VOC object categories on MS-COCO validation set.

for 20 PASCAL VOC object classes (a subset of 80 MS-COCO object classes). For the remaining 60 object classes, we train 60 R-CNN detectors using the training set of MS-COCO. We use the Alex's Net pre-trained on ILSVRC 2012 without fine-tuning. All the features are obtained from the output of layer fc_7. To validate our trained models, we evaluate the 60 trained object detectors using the MS-COCO validation set. The detection average precisions (AP) are reported in Tab. 1. We see the APs are generally higher for larger objects such as elephants and trucks, while the APs are lower for smaller objects such as forks and remotes. Overall, more than half of the object classes have AP below 20%, showing the limitation of using object detection as a mid-level representation for HOI recognition.

References

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *JLMR*, 9:1871–1874, 2008. 1

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1

[3] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 1

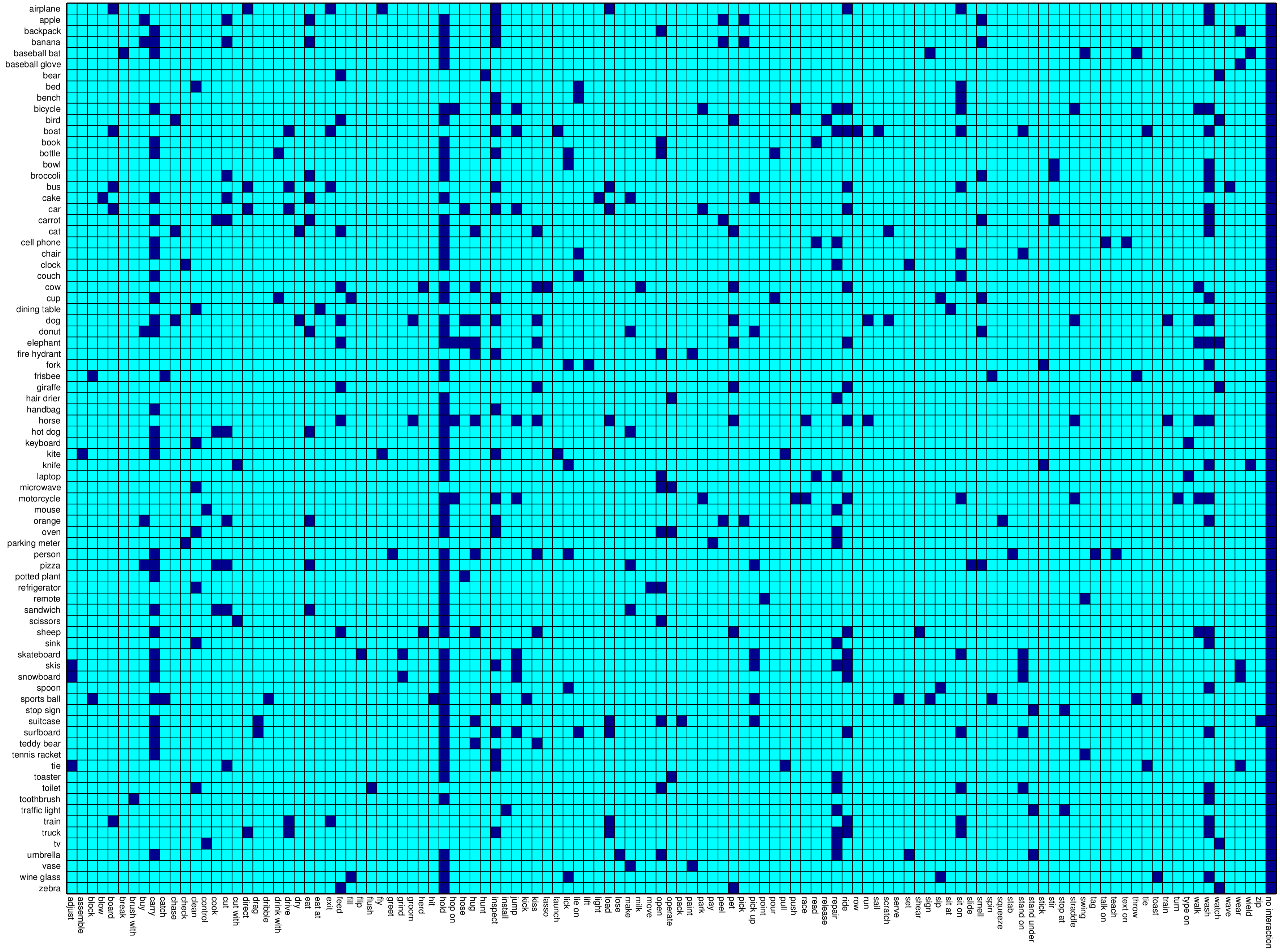

Figure 1: Our HICO dataset consists of 600 HOI categories over 117 verbs (including the “no interaction” class) and 80 object classes.

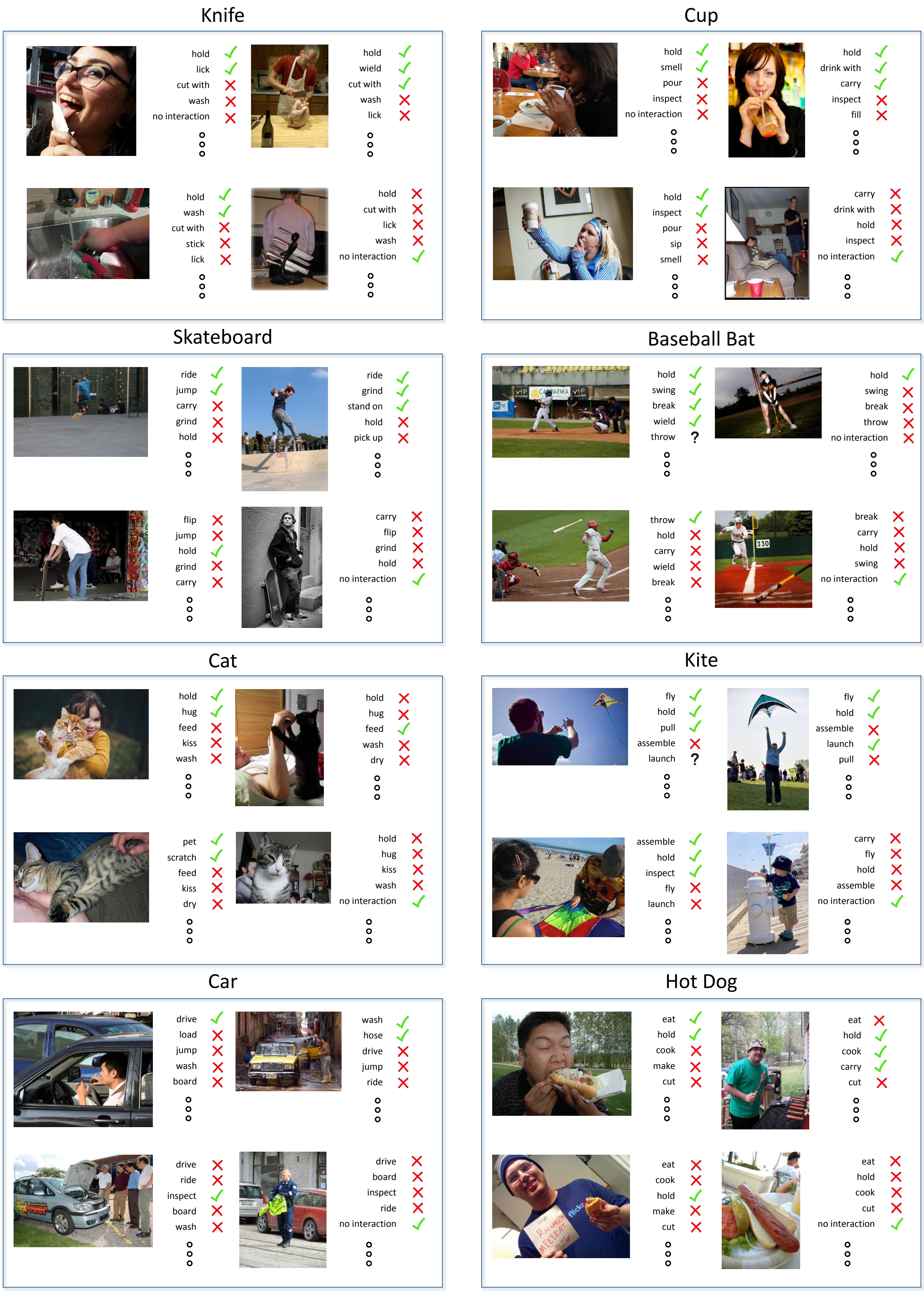
Knife
hold ✓ lick ✓ cut with ✗ wash ✗ no interaction ✗
hold ✓ wield ✓ cut with ✓ wash ✗ lick ✗
hold ✓ wash ✓ cut with ✗ stick ✗ lick ✗
hold ✗ cut with ✗ lick ✗ wash ✗ no interaction ✓
Cup
hold ✓ smell ✓ pour ✗ inspect ✗ no interaction ✗
hold ✓ drink with ✓ carry ✓ inspect ✗ fill ✗
hold ✓ inspect ✓ pour ✗ sip ✗ smell ✗
carry ✗ drink with ✗ hold ✗ inspect ✗ no interaction ✓
Skateboard
ride ✓ jump ✓ carry ✗ grind ✗ hold ✗
ride ✓ grind ✓ stand on ✓ hold ✗ pick up ✗
flip ✗ jump ✗ hold ✓ grind ✗ carry ✗
carry ✗ flip ✗ grind ✗ hold ✗ no interaction ✓
Baseball Bat
hold ✓ swing ✓ break ✓ wield ✓ throw ?
hold ✓ swing ✗ break ✗ throw ✗ no interaction ✗
throw ✓ hold ✗ carry ✗ wield ✗ break ✗
break ✗ carry ✗ hold ✗ swing ✗ no interaction ✓
Cat
hold ✓ hug ✓ feed ✗ kiss ✗ wash ✗
hold ✗ hug ✗ feed ✓ wash ✗ dry ✗
pet ✓ scratch ✓ feed ✗ kiss ✗ dry ✗
hold ✗ hug ✗ kiss ✗ wash ✗ no interaction ✓
Kite
fly ✓ hold ✓ pull ✓ assemble ✗ launch ?
fly ✓ hold ✓ assemble ✗ launch ✓ pull ✗
assemble ✓ hold ✓ inspect ✓ fly ✗ launch ✗
carry ✗ fly ✗ hold ✗ assemble ✗ no interaction ✓
Car
drive ✓ load ✗ jump ✗ wash ✗ board ✗
wash ✓ hose ✓ drive ✗ jump ✗ ride ✗
drive ✗ ride ✗ inspect ✓ board ✗ wash ✗
drive ✗ board ✗ inspect ✗ ride ✗ no interaction ✓
Hot Dog
eat ✓ hold ✓ cook ✗ make ✗ cut ✗
eat ✗ hold ✓ cook ✓ carry ✓ cut ✗
eat ✗ cook ✗ hold ✓ make ✗ cut ✗
eat ✗ hold ✗ cook ✗ cut ✗ no interaction ✓

Figure 3: Samples of images and HOI annotations in the HICO dataset.