

Supplemental: A Simple Baseline for Weakly-Supervised Human-centric Relation Detection

Raghav Goyal¹²
rgoyal14@cs.ubc.ca

Leonid Sigal¹²³
lsigal@cs.ubc.ca

¹ University of British Columbia
Vancouver, BC, Canada

² Vector Institute for AI

³ CIFAR AI Chair

A Training details and Hyperparameters

HICO-DET. We use batch size of 24 frames distributed across two NVIDIA Tesla T4 GPUs with 16 GB memory, and trained for 10 epochs. Note that we are able to use a higher batch size because we use the features from a pretrained detector instead of training it end-to-end in order to make fair comparisons to [2]. We use SGD optimizer with a learning rate of 0.001, and a scheduler which divides the learning by a factor of 10 after 5th and 8th epoch. As mentioned in Section 4.1, we use 15% of train set as the validation set to pick the best model.

Action Genome. We use batch size of 12 frames distributed across four NVIDIA Tesla T4 GPUs with 16 GB memory, and trained for 8 epochs. We use SGD optimizer with a learning rate of 0.001, and a scheduler which divides the learning by a factor of 10 after 4th and 6th epoch. As mentioned in Section 4.2, we use 400 videos of train set as the validation set to pick the best model.

B Pipeline for training and testing

In this section we describe the details of training and testing pipelines both for Action Genome and HICO-DET datasets. We refer to Action Genome as AG and HICO-DET as HICO for the rest of the section. Any detail, unless specified, applies both to AG and HICO. This section is supplemental to Implementation Details in Section 4.1 and 4.2 and mainly describes pathways a sample take during training and testing procedures.

B.1 Training

Inputs, feature extraction and proposal generation. The input to our model is an image, ground truth bounding boxes and relations between them. For HICO, instead of an image we are provided with object proposals and their features from a pretrained detector for which we use average pooling to obtain RoI features (4096-dimensional) for every proposal. For

AG we use a feature pyramid network as backbone to obtain image features, an RPN to obtain object proposals and an RoI head along with RoIAlign pooling method to obtain 4096-dimensional RoI features for every proposal.

Relation candidate pair generation. For AG we pick most confident person among the object proposals, and form all relation candidate pairs with that person as the subject. This is possible because AG dataset [15] is only one person-centric dataset. We subsample the obtained object proposals by using a threshold of 0.5 on their objectness score. This is done to ensure only high quality object proposals are used to form relations and use weak supervision as effectively as possible. The relation candidate pairs are then formed by enumerating all the subsampled object proposals with the selected person proposal mentioned above.

For HICO, as mentioned previously, we are given object proposals from a pretrained detector and whether a proposal belongs to a person or not. We pick all the persons in one set and rest of the non-person objects in the other set, and form relation candidate pairs by taking cross product between the person and non-person set. In that way we also account for multiple-persons and their interactions in our model.

Object and Relation feature extraction. As mentioned in Section 4.1 and 4.2, we use a 2-layered MLP to obtain object features. For relations, we concatenate the object features of objects involved in a relation and use a linear layer to obtain relation features. We use union features in the case of AG since it’s commonly used setup in Scene-graph literature [42], however we skip that in the case of HICO to maintain fairness of comparison with [2].

Loss. We form weak object and relation supervision by forming one-hot vectors where 1’s indicate presence of an object or relation class in an image and 0’s otherwise using their ground truth information. We include background class as the 0^{th} index in one-hot vectors and set it to 1 as we always assume that background class is present among the object and relation candidates for any image.

Fully-supervised version. The fully-supervised version of AG and HICO models differ from weakly-supervised models in the following ways: (1) **Sampling of relation candidate pairs.** We use the default relation sampling procedure from Scene Graph literature [41] where relation candidate pairs are assigned labels based on the degree of IoU match with ground truth relation pairs, and then are subsampled for a mini-batch based on foreground and background subsampling to maintain the ratio of background and non-background relation candidates for training, and (2) **Loss criteria.** The label assignment described above is then used as ground truth to supervise object and relation detection branches where cross entropy loss is used for object and relation classification, and smooth L1 loss is used for bounding box regression for object branch.

B.2 Testing

For testing, the details described in Training section are applicable up until the loss computation. The relation predictions are scored as a product of object scores involved and the relation score itself. However, for HICO we use only the relation score for scoring relation predictions. The obtained predictions are then sorted according to their scores and evaluated using standard evaluation protocols for AG [41] and HICO [2, 40]. For AG we also use frequency prior to modulate final predictions as done in prior related works [15, 42].

C Ablation: Balancing terms for object and relation loss

Section 3.3 describes the total loss which is the sum of two terms \mathcal{L}_{weak}^{obj} from equation 3 and \mathcal{L}_{weak}^{rel} from equation 4. Here we ablate over balancing terms λ_1 and λ_2 such that the total loss becomes $\mathcal{L} = \lambda_1 \mathcal{L}_{weak}^{obj} + \lambda_2 \mathcal{L}_{weak}^{rel}$. Table 1 shows the results where we observe that the default configuration of $\lambda_1 = 1, \lambda_2 = 1$ performs best.

Table 1: Ablation for balancing terms λ_1 and $\lambda_2 = 1$ corresponding to weak object and relation loss respectively.

Method	Full (600)	Rare (138)	Non-Rare (462)
$\lambda_1 = 1, \lambda_2 = 1$	28.77	24.64	30.00
$\lambda_1 = 1, \lambda_2 = 2$	25.16	22.26	26.02
$\lambda_1 = 2, \lambda_2 = 1$	26.19	22.78	27.21

D Qualitative results on HICO-DET and Action Genome

We include some qualitative results on HICO-DET and Action Genome dataset in this section.

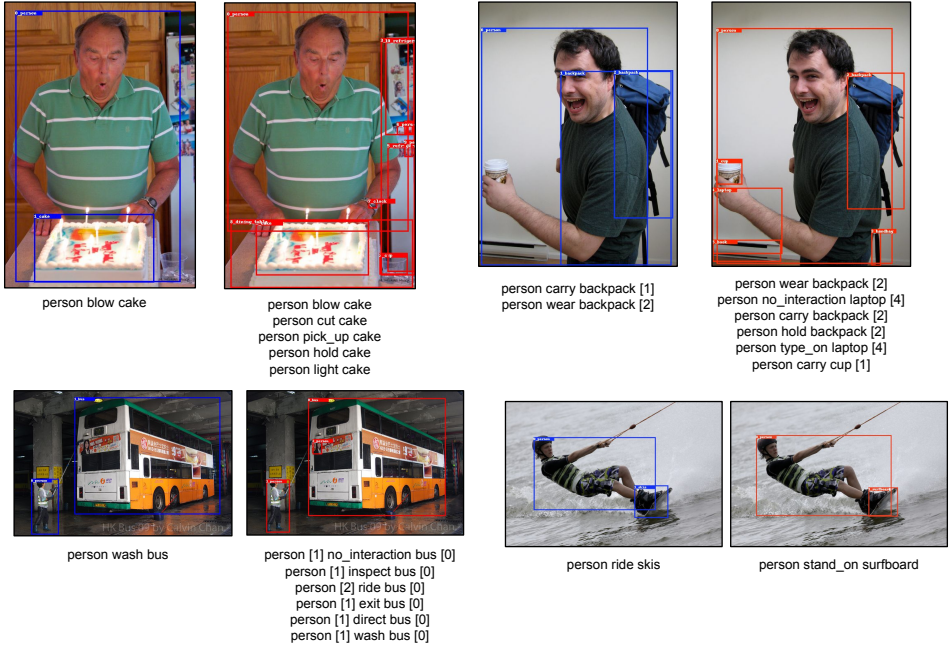


Figure 1: **Qualitative results on HICO-DET.** The figure shows pairs of images where the left image displays ground-truth objects and relations (in blue), and the right image displays predictions (in red). The predictions are sorted in descending order according to their scores, and optionally a number is referenced for an object, e.g. `backpack [2]`, which denotes the object number displayed in the image. Notably, in some cases many objects are predicted which increases the number of predicted relations (such as row 1 - left, row 1 - right), others have incorrect objects (row 2 - right), and incorrect relations (row 2 - left).

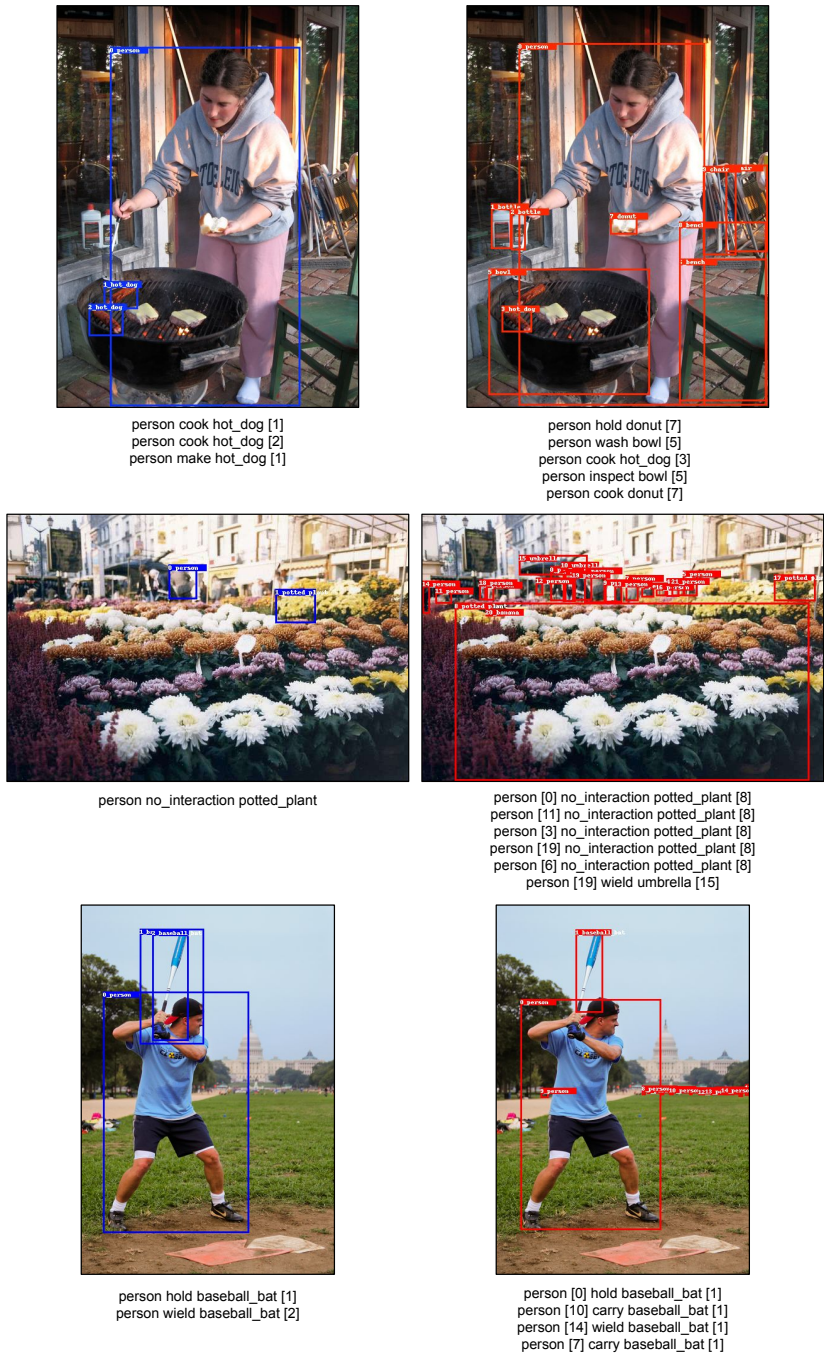
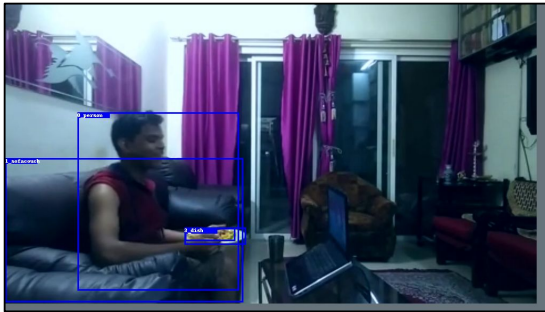


Figure 2: **Qualitative results on HICO-DET.** Same instructions as in Fig 1, but here we show predictions where multiple objects and persons are detected, which leads to some interesting object interactions (such as row 1), or in some cases persons present in background results in incorrect relation predictions (such as row 3).



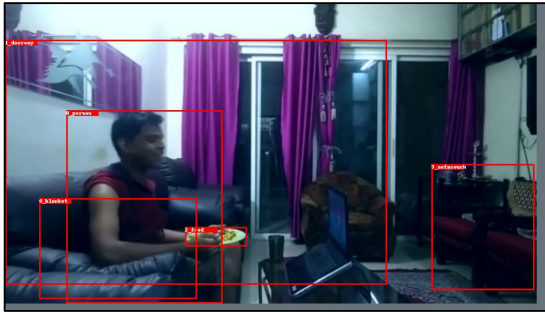
person not looking at sofacouch [1]
 person sitting on sofacouch [1]
 person leaning on sofacouch [1]

 sofacouch [1] behind person
 sofacouch [1] beneath person

 person holding food [2]

 person not looking at dish [3]
 person holding dish [3]

 dish [3] in front of person



person not looking at doorway [1]
 person not contacting doorway [1]
 doorway [1] in person

 person holding food [2]
 person looking at food [2]
 food [2] in front of person

 person not looking at sofacouch [3]
 sofacouch [3] beneath person

 person sitting on blanket [4]
 person not looking at blanket [4]
 blanket [4] beneath person



person looking at vacuum [1]
 person holding vacuum [1]
 vacuum [1] in front of person

 person not looking at floor [2]
 person standing on floor [2]
 floor [2] beneath person



person looking at cup/glass/bottle [1]
 person holding cup/glass/bottle [1]
 cup/glass/bottle [1] in front of person

 person not looking at sofacouch [2]
 sofacouch [2] behind person

 person looking at vacuum [3]
 person holding vacuum [3]
 vacuum [3] in front of person

 person not looking at clothes [4]
 person holding clothes [4]
 clothes [4] in front of person

Figure 3: **Qualitative results on Action Genome.** We sort the relations according to their scores and for each person-object pair, and we look at top 3 predicted relations in order to account for three different types of relations - attention, spatial and contact. The predictions contain objects that are not present in ground-truth, which gives interesting results such as doorway in row 2, and cup/glass/bottle in row 4.

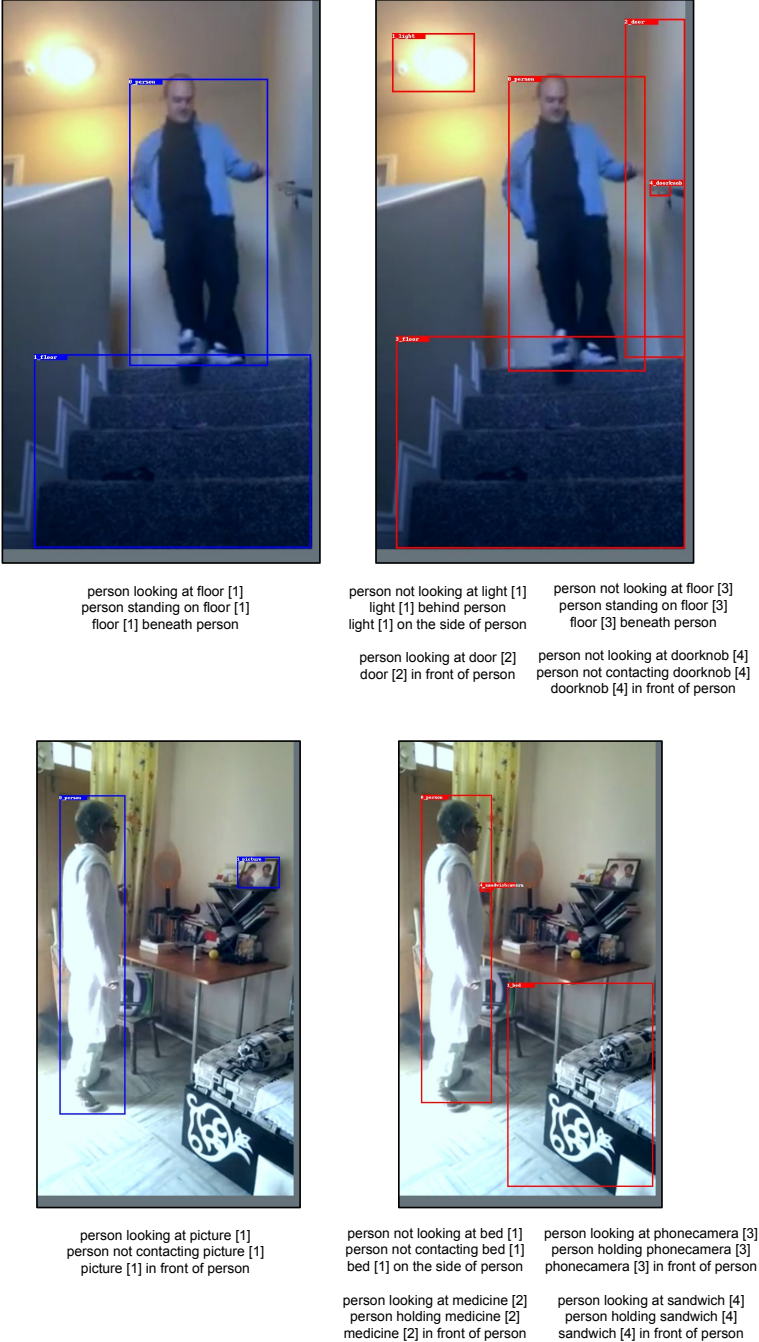


Figure 4: **Qualitative results on Action Genome.** Same instructions as in Fig 3. We again see extra objects being predicted and the relation predictions the model makes on those objects.