Exploring Context and Visual Pattern of Relationship for Scene Graph Generation

Wenbin Wang[1,2], Ruiping Wang[1,2], Shiguang Shan[1,2,3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Peng Cheng Laboratory, Shenzhen, 518055, China
wenbin.wang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

In the following parts, we will firstly provide more details of the models and training process. Then we provide qualitative results on VG dataset for comparing the method **IMP**+Isc** with **IMP**** under the setting of PREDCLS metric to further demonstrate the superiority of our intersection region. Moreover, more qualitative samples for comparing **Mem** with **IMP**** under the setting of PREDCLS metric and SGGEN metric respectively are also provided which show the effectiveness of relationship context.

For convenience, here we briefly recall models mentioned in the main paper again.

- **IMP**** [5] (*i.e.* Ref. [42] in the main paper): Our *baseline* which uses union region to extract relationship features. We reimplement this model and re-train it using our object detector. *IMP*** is our reimplemented version.
- **IMP**+Isc**: Based on the reimplemented model *IMP***, it replaces the union region with intersection region.
- **Mem**: Our context-utilized model. It uses union region to extract relationship features.

1. Models and Training Details

Usage of Intersection Region. In the experiments, we compare the results between union region and intersection region. Besides, in order to explore for a better performance, we further try to combine these two types of features. The combination version of relationship features is computed as:

$$\mathbf{f}^{R_{comb}} = \mathrm{Conv}(\mathrm{Conv}([\mathbf{f}^{R_{uni}}, \mathbf{f}^{R_{isc}}])), \qquad (1)$$

where $\mathbf{f}^{R_{uni}}$ and $\mathbf{f}^{R_{isc}}$ represent relationship feature maps extracted by union regions and intersection regions respectively. $[\cdot]$ denotes channel-wise concatenation. The consecutive convolutions have filters of size 3×3 and remain the spatial size of feature maps unchanged. Besides, to avoid that the area of intersection region is too small, we enlarge its area with factor σ. We empirically set σ to 1.2.

Detector. We use Faster-RCNN [3] with VGG-16 [4] backbone as our front-end object detector. The VGG-16 is pretrained on ImageNet [1]. The detector is further finetuned on Visual Genome objects of 150 categories and optimized with SGD algorithm on a single Titan Xp GPU. During the whole 90k iterations, the learning rate is initialized as 1.0×10^{-3} and divided by 10 at 35k and 70k iterations respectively. The detector gets 25% mAP on Visual Genome. Once the detector is trained, we freeze its layers.

End-to-End Training and Attention Assembling. With our pretrained detector, the whole framework is then trained on ground truth scene graph annotations. For each image, we firstly sample 128 RoIs (region of interest), of which 25% are foreground. The foreground RoIs are sequentially sampled according to their degrees in the scene graph, which means that if an RoI is related with more other RoIs, it is more probable to be sampled. Next we sample relationships among the sampled RoIs, of which 75% are positive. The loss is the sum of cross entropy for predicate classification, cross entropy for object classification, and smooth L1 loss for object location regression in each round of predictions. We optimize the model with SGD on a single Titan Xp GPU with an initial learning rate 1.0×10^{-3}. Furthermore, we also try to assemble the predictions from each iteration with attention mechanism [2]. Therefore, when making a prediction in each iteration, an extra attention weight is predicted at the same time.

2. Qualitative Results of Intersection Region

In Fig.1, we show qualitative samples for comparing *IMP*** and *IMP**+Isc*. The results are generated under the setting of PREDCLS metric. In each pair of images, the left one is generated by *IMP*** while the right one is by *IMP**+Isc*. The orange and yellow triplet items below each pair are those missed and wrongly detected by *IMP*** respectively but detected correctly by *IMP**+Isc*.

With the intersection region, the model is able to learn better representations which are closer to their real visual patterns. Interestingly, we find some triplets relevant to animals shown in last several samples, such as *dog-on-skateboard*, *cat-on-towel*, which do not frequently appear. Our intersection region still helps detect them successfully. It proves that our intersection region reduces distraction from object information and pays attention to the visual pattern of relationship itself.

3. Qualitative Results of Relationship Context

In Fig.2 we demonstrate qualitative results for comparing *IMP*** and *Mem* under the setting of PREDCLS metric. Similarly, the orange triplets are those missed by *IMP*** but detected correctly by *Mem*. With relationship context, the model obtains higher recall especially for images in which lots of predicates repeat.

Finally, we give scene graph generation samples in Fig.3 for comparing *IMP*** and *Mem* under the setting of SGGEN metric. Our *Mem* model uses both object and relationship context, which makes great contributions to object and relationship inference.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1

[2] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2204–2212, 2014. 1

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 1

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[5] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017. 1

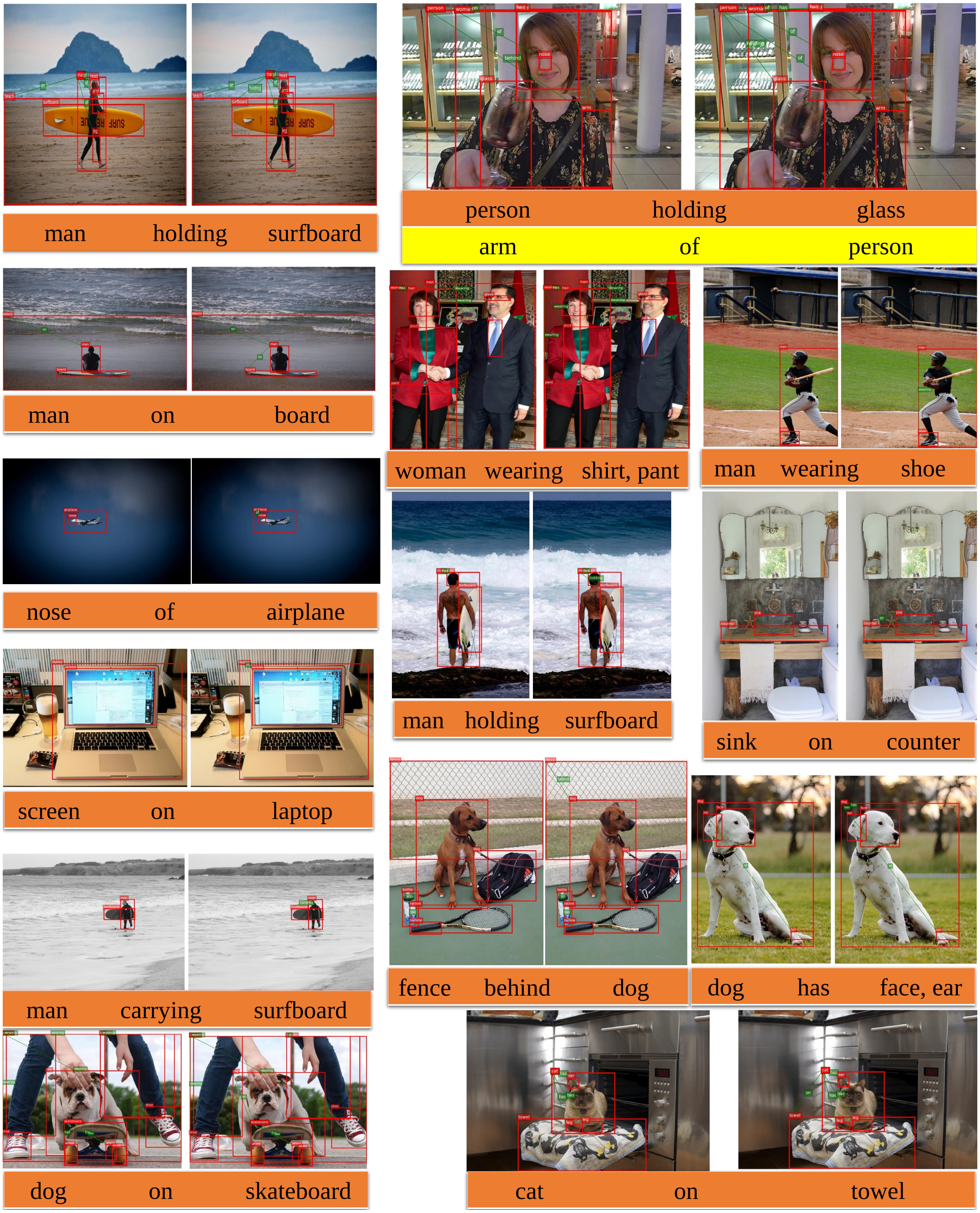

Figure 1. **Qualitative results of *IMP*** vs. *IMP**+Isc***. The results are generated under the setting of PREDCLS metric. In each pair of results, the left one is generated by *IMP*** while the right one is by *IMP**+Isc*. The orange and yellow triplet items below each pair are those missed and wrongly detected by *IMP*** respectively but detected correctly by *IMP**+Isc*.

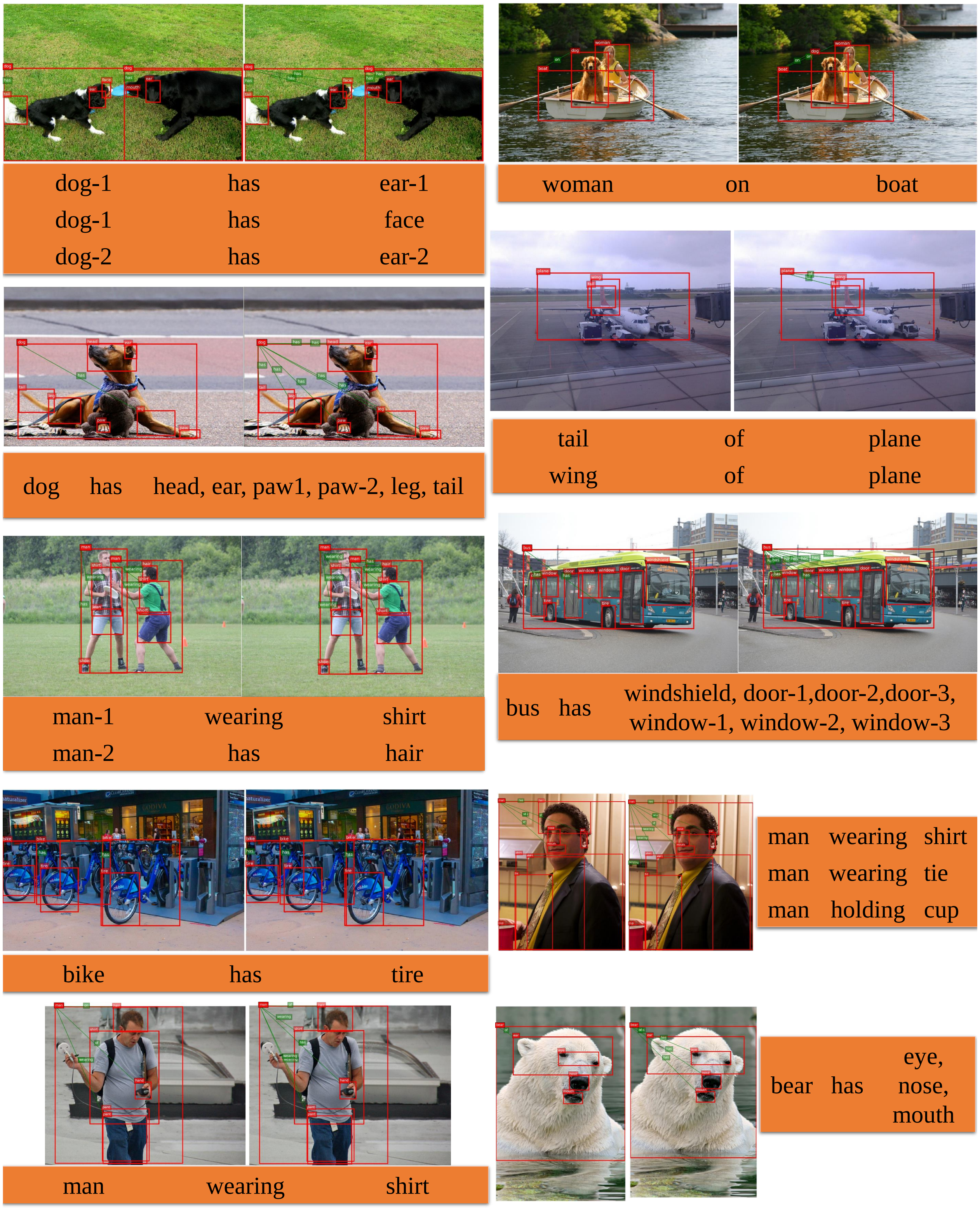

Figure 2. **More qualitative results of *IMP*** vs. *Mem***. The results are generated under the setting of PREDCLS metric. In each pair of results, the left one is generated by *IMP*** while the right one is by *Mem*. The orange triplet items are those missed by *IMP*** but detected correctly by *Mem*.

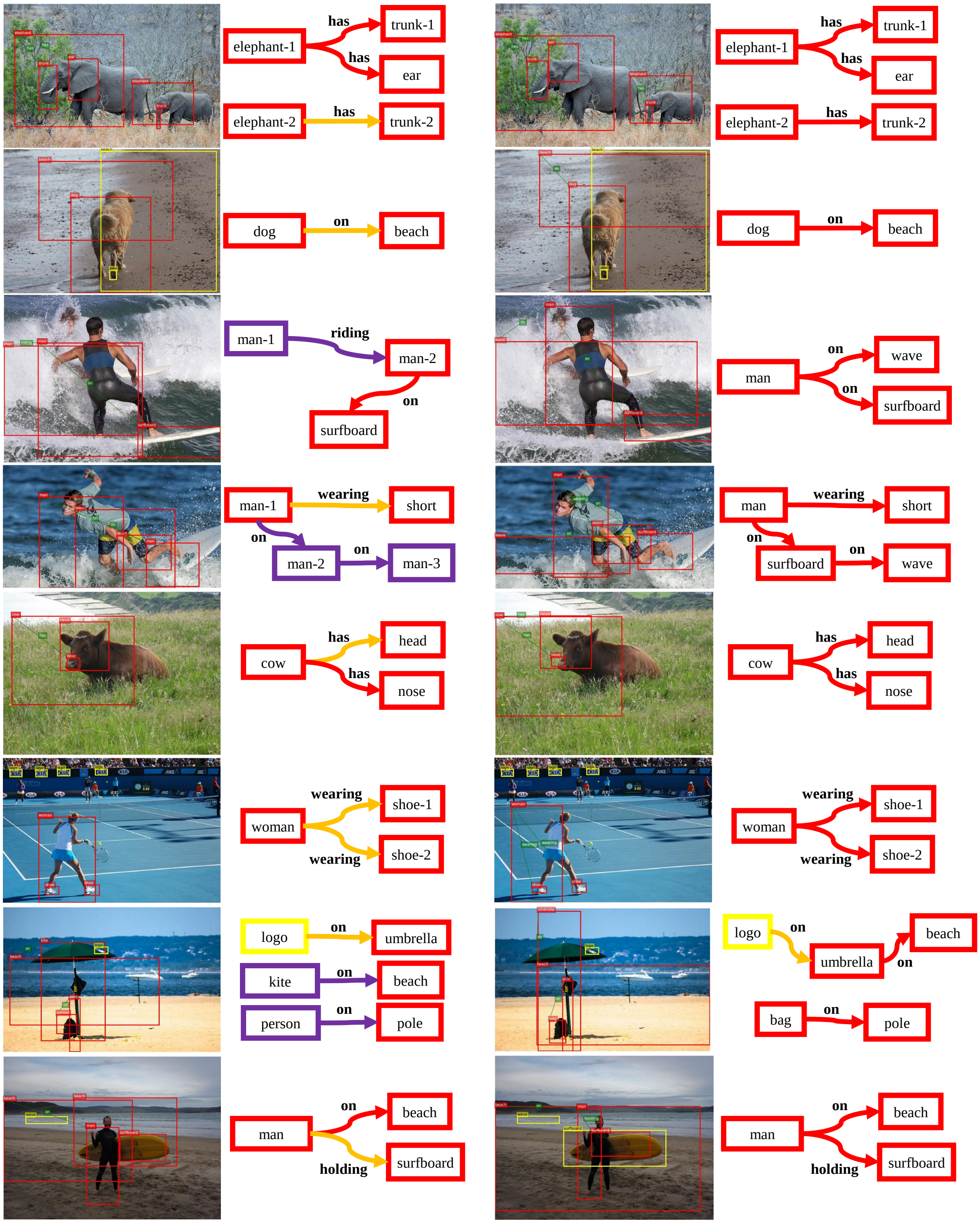

Figure 3. **More scene graph generation samples of *IMP*** vs. *Mem*.** The results are generated under the setting of SGGEN metric. In each row, the left image and scene graph are generated by *IMP*** while the right ones are generated by *Mem*. In images and scene graphs, red boxes are predicted and overlap with the ground truth (but the classes of red boxes in images may be wrong), yellow boxes are ground truth with no match. In scene graphs, red edges are true positives, orange edges are false negatives, purple boxes and edges are false positives. Some yellow boxes in scene graphs which do not exist in images mean that they are detected correctly but the model fails in detecting their relationships with any other objects.