9. Supplementary

In this section, we provide more details on R@K with larger K (K = 5, K =10) and the influence of in-domain text augmentation. We also show more qualitative examples of our proposed model with and without text augmentation from image captions.

9.1. Larger K values

Remov. %	Recall	Test A		Full Test	
		Base	Text-aug.	Base	Text-aug.
50%	Rel-Obj R@5	84.54	86.17	71.35	78.18
	Rel-Obj R@10	87.35	88.57	73.95	81.38
	Obj-Loc R@5	58.58	59.90	52.18	57.37
	Obj-Loc R@10	60.36	61.71	54.38	59.87

Table 6. Results with different K values for Test A and Full Test. "Base" indicates removing different portions of training samples for the relation-object pairs in *Rel-Obj Set A* and removing all training samples for *Rel-Obj Set B*. "Text-aug." indicates adding ungrounded samples for both unseen and under-sampled relation-object pairs.

In Table 6, we report R@5 and R@10 results for Test A and Full Test. Increasing K improves both relation-object prediction and object-location coordinates prediction, but such benefits start to become marginal when K is larger. For example, from Table 2, the Rel-Obj recall and Obj-Loc recall increase to 75.51 (Rel-Obj R@3) and 55.38 (Obj-Loc R@3) from 53.48 (Rel-Obj R@1) and 39.36 (Obj-Loc R@1) when K value increases from 1 to 3. By considering two more predictions for all the relation-object pairs, Rel-Obj recall increases 22.03, and Obj-Loc recall increases 16.02. However, Obj-Loc @10 and Rel-Obj @10 can reach 81.38 and 59.87 from 75.51 (Rel-Obj R@3) and 55.38 (Obj-Loc R@3) by considering seven more predictions, showing a 5.87 improvement in Rel-Obj recall and 4.49 improvement in Obj-Loc recall.

9.2. What is the effect of the amount of grounded data seen during base training?

In Table 7, we show the numbers of different amounts of data 'removed' from triplets belonging to *Test A*. This setting simulates the presence of relation-object pairs that might not have many training samples with strong grounding annotations. As seen in Table 7, our method continues to add additional value at all levels of removal. With text augmentation, our method compares very favorably with having a lot more grounded data in the base training. For example, after text augmentation, our model with 50% data removed has a slightly better relation-object prediction and almost the same object-location coordinates prediction as the 'Base' model trained with only 25% data removed (∼24k additional fully grounded training samples). This further suggests a strong possibility of rapid expansion of the total number of unique relation-object pairs by only annotating a small amount of additional grounded data, and using a large amount of image-text data to obtain the same benefit as annotating a lot of expensive grounding data.

Remov. %	Recall	Test A		Full Test	
		Base	Text-aug.	Base	Text-aug.
25%	Rel-Obj R@1	46.61	49.28	50.45	54.14
	Rel-Obj R@3	84.75	85.34	69.80	75.83
	Obj-Loc R@1	33.22	35.06	37.18	39.80
	Obj-Loc R@3	59.51	59.92	51.13	55.64
50%	Rel-Obj R@1	39.94	43.61	49.70	53.81
	Rel-Obj R@3	82.19	83.91	69.22	75.51
	Obj-Loc R@1	29.08	31.75	36.54	39.36
	Obj-Loc R@3	57.06	58.58	50.37	55.38
75%	Rel-Obj R@1	32.37	37.90	49.04	53.09
	Rel-Obj R@3	77.56	80.43	68.80	75.18
	Obj-Loc R@1	25.05	27.01	36.26	38.84
	Obj-Loc R@3	52.23	54.32	49.84	54.55

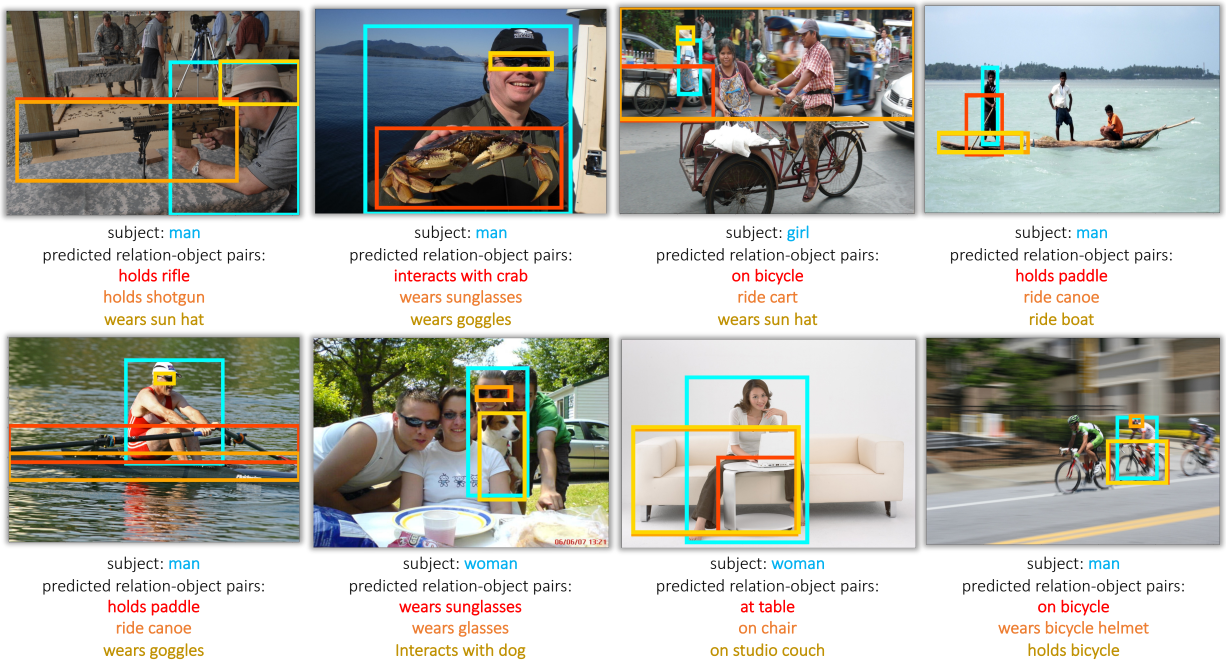
Table 7. Results for Test A and Full Test. "Base" indicates removing different portions of training samples for the relation-object pairs in *Rel-Obj Set A* and removing all training samples for *Rel-Obj Set B*. "Text-aug." indicates adding ungrounded samples for both unseen and under-sampled relation-object pairs.

9.3. In-domain Text Augmentation

Data	Rel-Object		Object-Loc	
	R@1	R@3	R@1	R@3
ATS + OutDomain	20.96	37.59	12.20	20.74
ATS + InDomain	48.24	86.51	23.93	43.73

Table 8. Effect of additional text augmentation from in-domain data (OIv6+VG+Flickr) and out-of-domain data (COCO+CC), starting from a base training on the Ablation Training Set (ATS) (See Sec. 7.2). Considerable gains are obtained by text-augmented data without any additional box information. Results are reported for 104 relation-object pairs.

In Table 8, we report results for 104 relation-object pairs from Test A and Test B. For both experiments, we remove all the samples for these 104 pairs from VG, Flickr30k and OIv6. Then, we add ungrounded samples for these 104 relation-object pairs from VG+Flickr30k+OIv6 (in-domain data) and from COCO+CC (out-of-domain data) separately.



subject: man
predicted relation-object pairs:
holds rifle
holds shotgun
wears sun hat

subject: man
predicted relation-object pairs:
interacts with crab
wears sunglasses
wears goggles

subject: girl
predicted relation-object pairs:
on bicycle
ride cart
wears sun hat

subject: man
predicted relation-object pairs:
holds paddle
ride canoe
ride boat

subject: man
predicted relation-object pairs:
holds paddle
ride canoe
wears goggles

subject: woman
predicted relation-object pairs:
wears sunglasses
wears glasses
Interacts with dog

subject: woman
predicted relation-object pairs:
at table
on chair
on studio couch

subject: man
predicted relation-object pairs:
on bicycle
wears bicycle helmet
holds bicycle

Figure 4. Qualitative results from the model trained with fully grounded data. Subjects and predicted relation-object pairs are shown under images, and bounding boxes correspond to subjects and objects with the same colors. During inference time, all subjects and bounding boxes for subjects are provided as inputs, and we show three predicted relation-object pairs with highest scores.

In-domain data contains clean subjects and relation-object pairs for images, and shows higher improvement than out-of-domain data for both text and box predictions as expected.

9.4. Qualitative Examples

First, we show predicted relation-object pairs and bounding boxes for objects using fully grounded data. In this case, the model is trained by Visual Genome, Flickr30k and OIv6 with all the text and bounding boxes. In Figure 4, our model can successfully predict correct relation-object pairs given a specific subject and the bounding box for the subject. For example, in the top right figure, by given the text and location of the leftmost man, the model can successfully predict "holds paddle" and "ride canoe" with corresponding bounding boxes for "paddle" and "canoe" for the given man, even there is another paddle in the same figure. These examples also show the future direction to improve the grounding performance. For example, in the second row, the third image shows the predicted boxes for "chair" and "studio couch", but these boxes are not accurate enough because of the incorrect prediction for the bottom-right x-coordinate. Since our model localize objects by generating box coordinates directly, the generated boxes can be improved by other post-processing steps to refine box predictions.

In Figure 5, results for "Base" are generated by the model trained with VG+Flickr30k+OIv6 by removing 50% of training samples for relation-object pairs in *Rel-Obj Set A* and removing all training samples for pairs in *Rel-Obj Set B*. The model generating results for "Text-aug" is trained with the same amount of grounded data and additional ungrounded data from COCO+CC for the relation-object pairs in *Rel-Obj Set A* and *Rel-Obj Set B*. For both sets, additional text augmentation helps the text prediction and grounding. For example, in the top left example, without additional text augmentation, the relation-object pair "wears goggles" cannot be predicted correctly. Text augmentation introduces more images to help the model to recognize relation-object pairs better.

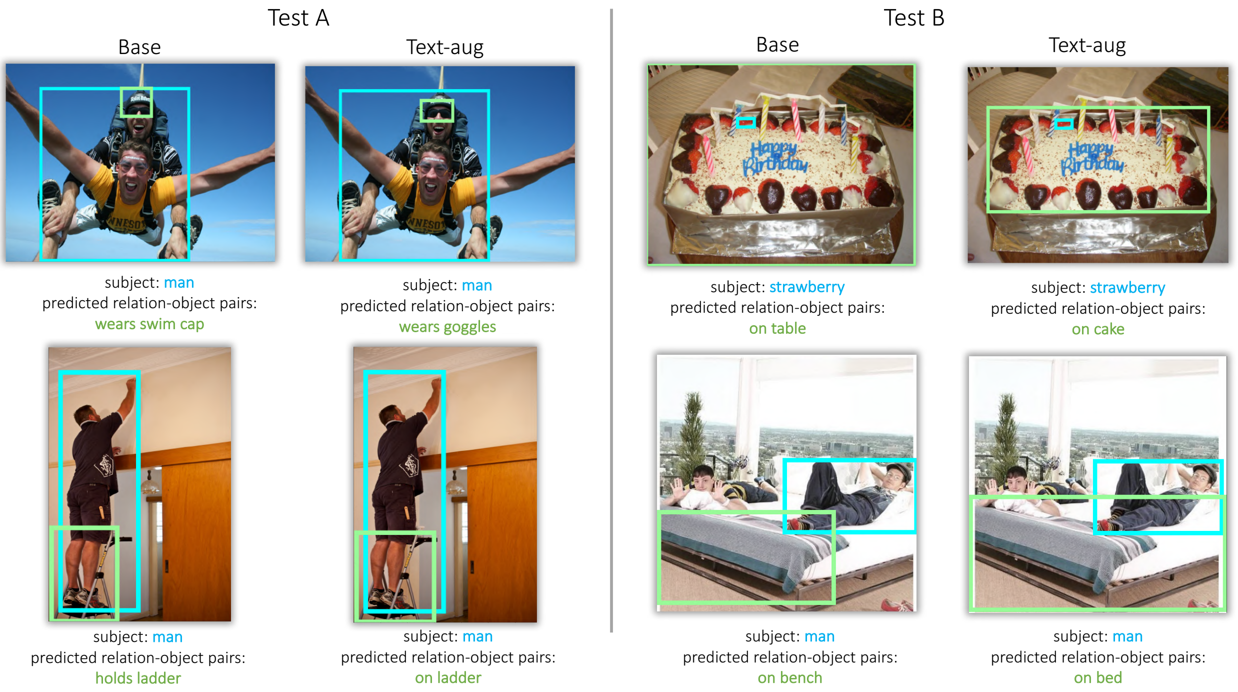

Test A
Base
Text-aug
subject: man
predicted relation-object pairs:
wears swim cap
subject: man
predicted relation-object pairs:
wears goggles
subject: man
predicted relation-object pairs:
holds ladder
subject: man
predicted relation-object pairs:
on ladder
Test B
Base
Text-aug
subject: strawberry
predicted relation-object pairs:
on table
subject: strawberry
predicted relation-object pairs:
on cake
subject: man
predicted relation-object pairs:
on bench
subject: man
predicted relation-object pairs:
on bed

Figure 5. Qualitative results from the models trained with and without text augmentation from COCO and CC. “Base” indicates results generated by the model trained with 50% training samples for the relation-object pairs in *Rel-Obj Set A* and no samples for the relation-object pairs in *Rel-Obj Set B*. “Text-aug” indicates results generated by the model trained with additional ungrounded data for both relation-object pairs in *Rel-Obj Set A* and *Rel-Obj Set B* from COCO and CC.