Supplementary File

1. Full Results on Comprehension

As previous work [3, 5, 4], we show speaker's comprehension performance trained both w/o MMI and with MMI in Table. 1. For some models that have both speaker and listener, we highlight the speaker module being used for comprehension in bold. For example, "**speaker**+listener" means we use the speaker module of the joint model to do the comprehension task. Our baseline is from [3], which we denote as "baseline" and "baseline+MMI". We also show the performance of a "pure" listener and previous state-of-art results from [5, 4] as reference.

From Table. 1, we first observe all speaker models trained with MMI outperform w/o MMI. This is consistent with previous work [5, 3]. Second, as each row shows the speaker's comprehension after adding one module from listener or reinforcer during training, it is easy to observe the benefits of adding each module row by row. Our speaker jointly trained with the listener and reinforcer achieves the state-of-art results and can outperform the pure listener by $\sim 2\%$ on all three datasets.

Then we show the evaluations using variations of the listener module or ensembled listener+speaker modules for the comprehension task in Table. 2. Similarly, we highlight the listener module used in our models in bold, e.g., speaker+**listener**. We notice the joint training with speaker or reinforcer always brings additional discriminative benefits to the listener module resulting in improved performance. However the "+MMI" on speaker seems not that effecting the listener's performance. The best results are achieved by ensembling speaker and listener together.

While the above experiments analyze comprehension performance given ground-truth bounding boxes for potential comprehension objects, we also show the comprehension using object detector. We use detector trained by SSD [2] to automatically select regions for consideration. The results are shown in the bottom half of Table. 1 and 2). Overall the improvements are consistent with using ground-truth objects.

For the generation task, we evaluate variations on the speaker module. We show automatic evaluation using the METEOR and CIDEr metrics for generation in Table 3 where "+rerank" denotes models incorporating the reranking mechanism and global optimization. To computer CIDEr robustly, we collect more expressions for objects in the test sets for RefCOCO and RefCOCO+, obtaining 10.1 and 9.4 expressions respectively on average per object. For RefCOCOg we use the original expressions released with the dataset which may be limited, but we still show its performance for completeness. We choose the "speaker+tie" model in [5] as reference, which learns to tie the expression generation together and achieves state-of-art performance. Generally we find that the speaker in jointly learned models achieves higher scores than the single speaker under both metrics across datasets. Such improvements are observed under both settings without "+rerank" or with "+rerank".

2. More Examples

In this section, we show more comprehension examples in Fig. 1 and Fig. 2 using our strongest comprehension model, i.e., the ensemble of speaker and listener trained from "speaker+listener+reinforcer+MMI" model. We first show some comprehension results based on ground truth bounding boxes provided by MS COCO [1] in Fig. 1. We then show more comprehension results based on the regions detected by SSD [2] in Fig. 2.

We then show more examples on referring expression generation. In Fig. 3, we compare the generated expressions using the speaker module of different models. Our full model "speaker+listener+reinforcer+MMI+rerank" is able to achieve more discriminative expressions than the others as it considers listener's behavior. We further show the joint expression generation in Fig. 4. The expressions of every target object are considered together. Each of them is meant to be relevant to the target object and irrelevant to the other objects.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *ECCV*, 2016. 1

[3] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *arXiv preprint arXiv:1511.02283*, 2015. 1, 2

		RefCOCO			RefCOCO+			RefCOCOg
		val	TestA	TestB	val	TestA	TestB	val
1	listener	77.48%	76.58%	78.94%	60.50%	61.39%	58.11%	71.12%
2	previous state-of-art[4][5]	76.90%	75.60%	78.00%[4]	58.94%	61.29%	56.24%[5]	65.32%[5]
3	baseline[3]	64.56%	63.20%	66.69%	47.78%	51.01%	44.24%	56.81%
4	**speaker**[5]	69.95%	68.59%	72.84%	52.63%	54.51%	50.02%	59.40%
5	**speaker**+listener	71.20%	69.98%	73.66%	54.23%	56.22%	52.46%	61.83%
6	**speaker**+reinforcer	71.88%	70.18%	73.01%	53.38%	56.50%	51.16%	61.91%
7	**speaker**+listener+reinforcer	72.46%	71.10%	74.01%	55.54%	57.46%	53.71%	64.07%
8	baseline+MMI[3]	72.28%	72.60%	73.39%	56.66%	60.01%	53.15%	63.31%
9	**speaker**+MMI[5]	76.18%	74.39%	77.30%	58.94%	61.29%	56.24%	65.32%
10	**speaker**+listener+MMI	79.22%	77.78%	79.90%	61.72%	64.41%	58.62%	71.77%
11	**speaker**+reinforcer+MMI	78.38%	77.13%	79.53%	61.32%	63.99%	58.25%	67.06%
12	**speaker**+listener+reinforcer+MMI	**79.56%**	**78.95%**	**80.22%**	**62.26%**	**64.60%**	**59.62%**	**72.63%**
		RefCOCO (detected)			RefCOCO+ (detected)			RefCOCOg (detected)
		val	TestA	TestB	val	TestA	TestB	val
1	listener	-	71.63%	61.47%	-	57.33%	47.21%	56.18%
2	previous state-of-art[5]	-	72.03%	63.08%	-	58.87%	47.70%	58.26%
3	baseline[3]	-	64.42%	56.75%	-	52.84%	42.68%	53.13%
4	**speaker**[5]	-	67.69%	60.16%	-	54.37%	45.00%	53.83%
5	**speaker**+listener	-	68.27%	61.00%	-	55.41%	45.65%	54.96%
6	**speaker**+reinforcer	-	69.12%	60.47%	-	55.45%	44.96%	55.64%
7	**speaker**+listener+reinforcer	-	69.15%	61.96%	-	55.97%	46.45%	57.03%
8	baseline+MMI[3]	-	68.73%	59.56%	-	58.15%	46.63%	57.23%
9	**speaker**+MMI[5]	-	72.03%	63.08%	-	58.87%	47.70%	58.26%
10	**speaker**+listener+MMI	-	**72.95%**	63.10%	-	60.23%	48.11%	58.57%
11	**speaker**+reinforcer+MMI	-	72.34%	63.24%	-	59.36%	48.72%	58.70%
12	**speaker**+listener+reinforcer+MMI	-	72.88%	**63.43%**	-	**60.43%**	**48.74%**	**59.51%**

Table 1. Ablation study using the speaker module for the comprehension task (indicated in **bold**). Top half shows performance given ground truth bounding boxes for objects, bottom half performance using automatic object detectors to select potential objects. We find that adding listener and reinforcer modules to the speaker increases performance.

		RefCOCO			RefCOCO+			RefCOCOg
		val	TestA	TestB	val	TestA	TestB	val
1	listener	77.48%	76.58%	78.94%	60.50%	61.39%	58.11%	71.12%
2	previous state-of-art [4][5]	76.90%	75.60%	78.00% [4]	58.94%	61.29%	56.24% [5]	65.32% [5]
3	speaker+**listener**	77.84%	77.50%	79.31%	60.97%	62.85%	58.58%	72.25%
4	speaker+**listener**+reinforcer	78.14%	76.91%	80.10%	61.34%	63.34%	58.42%	71.72%
5	**speaker+listener**+reinforcer (ensemble)	78.88%	78.01%	80.65%	61.90%	64.02%	59.19%	72.43%
6	speaker+**listener**+MMI	78.42%	78.45%	79.94%	61.48%	62.14%	58.91%	72.13%
7	speaker+**listener**+reinforcer+MMI	78.36%	77.97%	79.86%	61.33%	63.10%	58.19%	72.02%
8	**speaker+listener**+reinforcer+MMI (ensemble)	**80.36%**	**80.08%**	**81.73%**	**63.83%**	**65.40%**	**60.73%**	**74.19%**
		RefCOCO (detected)			RefCOCO+ (detected)			RefCOCOg (detected)
		val	TestA	TestB	val	TestA	TestB	val
1	listener	-	71.63%	61.47%	-	57.33%	47.21%	56.18%
2	previous state-of-art[5]	-	72.03%	63.08%	-	58.87%	47.70%	58.26%
3	speaker+**listener**	-	72.23%	62.92%	-	59.61%	48.31%	57.38%
4	speaker+**listener**+reinforcer	-	72.65%	62.69%	-	58.68%	48.23%	58.32%
5	**speaker+listener**+reinforcer (ensemble)	-	72.78%	**64.38%**	-	59.80%	49.34%	**60.46%**
6	speaker+**listener**+MMI	-	72.95%	62.43%	-	58.68%	48.44%	57.34%
7	speaker+**listener**+reinforcer+MMI	-	72.94%	62.98%	-	58.68%	47.68%	57.72%
8	**speaker+listener**+reinforcer+MMI (ensemble)	-	**73.78%**	63.83%	-	**60.48%**	**49.36%**	59.84%

Table 2. Ablation study using listener or ensembled listener+speaker modules for the comprehension task (indicated in **bold**). Top half shows performance given ground truth bounding boxes for objects, bottom half performance using automatic object detectors to select potential objects. We find that jointly training with the speaker improves listener's performance and that adding the reinforcer module in an ensembled speaker+listener prediction performs the best.

[4] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 1, 2

[5] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. *ECCV*, 2016. 1, 2, 3

	RefCOCO				RefCOCO+				RefCOCOg	
	Test A		Test B		Test A		Test B		val	
	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr
speaker+tie [5]	0.283	0.681	0.320	1.273	0.204	0.499	0.196	0.683	-	-
baseline+MMI	0.243	0.615	0.300	1.227	0.199	0.462	0.189	0.679	0.149	0.585
speaker+MMI	0.260	0.679	0.319	1.276	0.202	0.475	0.196	0.683	0.147	0.573
speaker+listener+MMI	0.268	0.704	0.327	1.303	**0.208**	**0.496**	0.201	0.697	0.150	0.589
speaker+reinforcer+MMI	0.266	0.702	0.323	1.291	0.204	0.482	0.197	0.692	0.151	**0.602**
speaker+listener+reinforcer+MMI	**0.268**	0.697	**0.329**	**1.323**	0.204	0.494	**0.202**	**0.709**	**0.154**	0.592
baseline+MMI+rerank	0.280	0.729	0.329	1.285	0.204	0.484	0.205	0.730	**0.160**	0.654
speaker+MMI+rerank	0.287	0.745	0.334	1.295	0.208	0.490	0.213	0.712	0.156	0.653
speaker+listener+MMI+rerank	0.293	0.763	0.337	1.306	0.211	0.500	**0.221**	0.734	0.159	0.650
speaker+reinforcer+MMI+rerank	0.291	0.748	0.337	1.311	0.207	0.499	0.215	0.729	0.158	0.653
speaker+listener+reinforcer+MMI+rerank	**0.296**	**0.775**	**0.340**	**1.320**	**0.213**	**0.520**	0.215	**0.735**	0.159	**0.662**

Table 3. Ablation study for generation using automatic evaluation.

Figure 1. Example comprehension results from each dataset using ground truth bounding boxes. Green box shows the ground-truth region and red box shows incorrect comprehension. We show some correct comprehension results in the top two rows and the incorrect ones in the bottom two rows.

Figure 2. Example comprehension results from each dataset using detection. Green box shows the ground-truth region, blue box shows correct comprehension using our speaker+listener+reinforcer+MMI model on detection, and red box shows incorrect comprehension. We show some correct comprehension results in the top two rows and the incorrect ones in the bottom two rows.

Figure 3. Example generation results from each dataset. From top to bottom showing: speaker+MMI, speaker+listener+MMI, speaker+reinforcer+MMI, speaker+listener+reinforcer+MMI, speaker+listener+reinforcer+MMI+rerank.

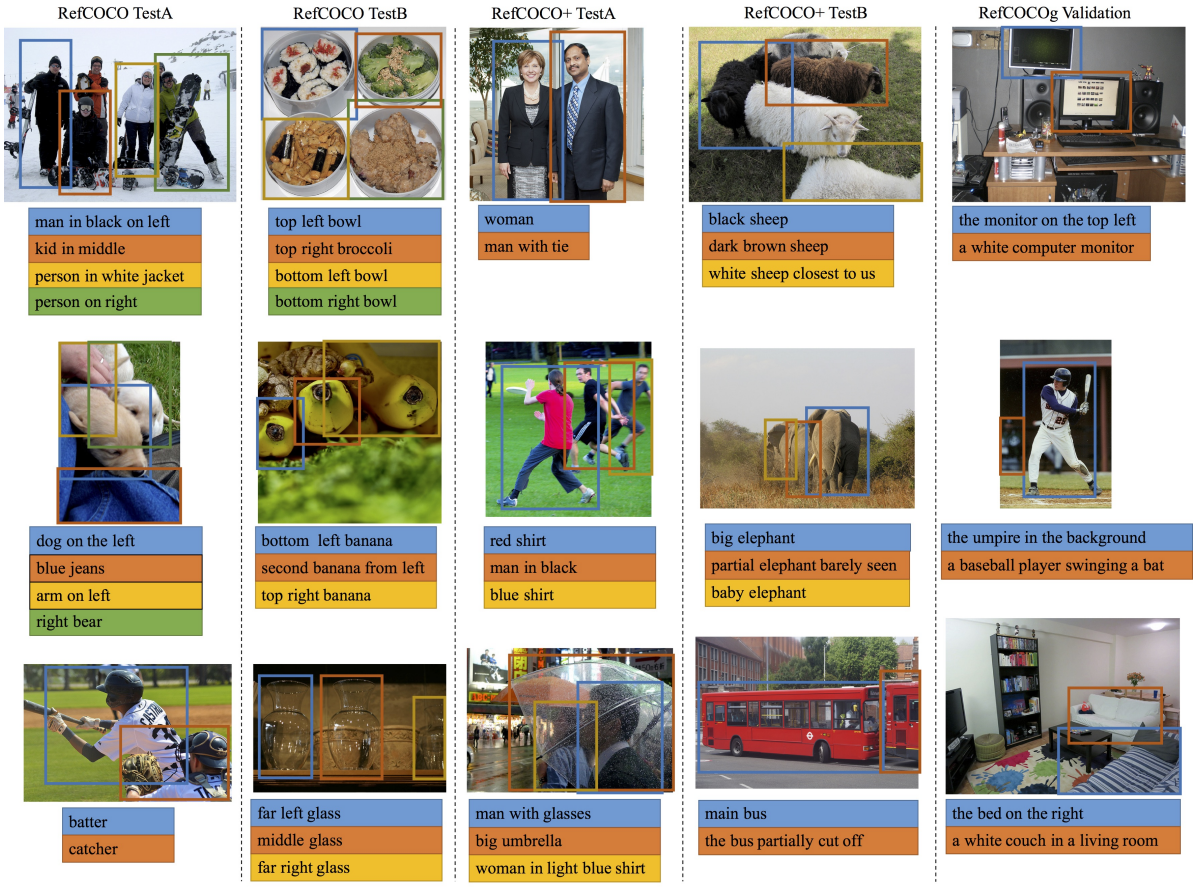

Figure 4. Joint generation examples using speaker+listener+reinforcer+MMI+rerank. Each expression shows the generated expression for one of the depicted objects (color coded to indicate correspondence)