Supplementary Material: Generating Easy-to-Understand Referring Expressions for Target Identifications

Mikihiro Tanaka[1], Takayuki Itamochi[2], Kenichi Narioka[2], Ikuro Sato[3],
Yoshitaka Ushiku[1] and Tatsuya Harada[1,4]
[1]The University of Tokyo, [2]DENSO CORPORATION, [3]DENSO IT Laboratory, Inc., [4]RIKEN

1. Our task from the viewpoint of Gricean Maxims

Gricean Maxims [3], which is advocated as a collaborative principle for effective conversation between a speaker and a listener, has often been discussed in referring expression generation [6, 2, 5]. Gricean Maxims has four aspects: quality, quantity, relation and manner. When the targets are salient like on RefCOCO, the evaluation of the comprehension accuracy is enough to satisfy Gricean Maxims. However, when the composition of the image becomes complex like Fig. 1 in the main paper, the comprehension time which relates to quantity and manner is also needed for sentence evaluations to satisfy Gricean Maxims.

2. The novelty of our "context"

"Context" in our paper refers to the visual context of the target, such as nearby objects or features and also the context during generation of sentences where context here refers to previously generated words in the sentence. The visual context of the target allows us to identify its global location whilst also distinguishing from other targets. We back propagated the loss for generating expressions uniquely referring to the target back through to global, local and sentinel attention in Fig. 2 in the main paper. Our model can generate sentences by selecting important information for identification from inside and outside the target bounding box. "Context" plays an important role especially in such cases where the target is less salient or the target is hard to refer to by just mentioning its attributes.

Existing referring expression generation research [5] also uses "context." This research aims to distinguish the target from others however does not attempt to inform us of the location and does not utilize the relationship of the target to nearby objects or features which are not the same class as the target in their "context."

	RefCOCO	RefCOCO+	RefCOCOg	[1] (Video)	RefGTA
# of images	19,994	19,992	25,799	4,818	28,750
# of created instances	50,000	49,856	49,822	29,901	78,272
# of referring expressions	169,806	166,403	95,010	30,320	213,175

Table A. Statistics of annotations on existing datasets and our dataset (RefGTA). RefGTA contains more images, instances and referring expressions than the other datasets.

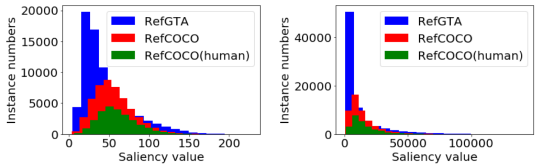


Figure A. Targets' saliency of RefGTA, RefCOCO and RefCOCO (human) calculated as Fig. 5 in our main paper. As the saliency becomes higher, the ratio of human instances becomes larger in RefCOCO.

3. Dataset comparison

3.1. Size

The comparison of the size in existing datasets (RefCOCO, refCOCO+ [5], RefCOCOg [4] and [1]) and RefGTA is shown in Table A. RefGTA contains more images, instances and referring expressions than existing datasets. The number of instances and the number of referring expressions are almost the same in [1] because the purpose of [1] is comprehension and does not need multiple sentences for automatic evaluation like generation.

3.2. The reason why RefGTA tends to have more targets with lower saliency than RefCOCO

As shown in Fig. A, even if we limit targets to humans in RefCOCO, there are few instances with low saliency. Images captured automatically are different from images taken by a person as they do not have subjects and tend to have miscellaneous information.

Figure B. Word clouds (from left to right: RefCOCO, RefCOCO+, RefCOCOg, RefGTA)

	baseline	Our SR	Our SR + rank loss
# of vocabularies	134	172	176

Table B. The number of vocabulary in the sentences generated by each method in the test set on RefGTA.

	RefCOCO (test)	RefCOCO+ (test)	RefCOCOg (val)	RefGTA (test)
baseline	2.68	2.51	6.50	9.46
Our SR	2.93	2.63	7.27	10.18
Our SR + rank loss	-	-	-	9.82
Ground-Truth	3.71	3.58	8.48	10.04

Table C. The average lengths of generated and ground-truth sentences. Our SR generated longer sentences than the baseline method in all datasets.

3.3. Word distribution

The word clouds on RefCOCO, RefCOCO+, RefCOCOg and RefGTA are shown in Fig. B.

4. Detailed results

4.1. Vocabulary in the generated sentences

The number of vocabulary in the sentences generated by each method on RefGTA is shown in Table B. Both of the sentences generated by our methods contain more vocabularies to represent the targets' surroundings (such as "beach", "bridge", "bus", "palm", "stairs", "store" , "pillar", "plant" , "railing" and "truck") than the sentences generated by the baseline method.

4.2. Generated sentence length

The mean lengths of generated and ground-truth sentences (i.e., the number of words in a description) are shown in Table C.

Our SR generated longer sentences than the baseline method. Considering that our method improved the automatic evaluation metrics as shown in Table 5 and Table 6 in the main paper, this indicates that our SR has the ability to describe in more detail than the baseline method. However, our SR generated shorter sentences than ground-truth in existing datasets, indicating that there is still a need for improvement in the capability of describing the targets.

On the other hand, the comprehension difficulty of GT varies on RefGTA unlike RefCOCO/+/g. Our SR generated sentences as long as ground-truth in RefGTA because

	short GT	middle GT	long GT
baseline	9.26	9.48	9.62
Our SR	9.87	10.28	10.39
Our SR+rank loss	9.54	9.90	10.02
GT	7.21	9.81	13.17

Table D. The average sentence lengths when instances are divided equally into three parts by calculating the average lengths of GT.

RefGTA is a large-scale dataset limiting targets to humans. This enabled our model to focus on easy-to-understand referring expression generation, and our SR+rank loss learned the concise sentences that are relatively easy to be comprehended.

The mean lengths of generated sentences on RefGTA grouped by GT lengths are also shown in Table D. Our SR+rank loss generated longer sentences than the baseline while shorter than our SR in each group.

4.3. Qualitative results

In this paper, we showed qualitative results on existing datasets (RefCOCO, RefCOCO+ and RefCOCOg) and our dataset (RefGTA). Here we show more results in Fig. C, Fig. D, Fig. E and Fig. F.

References

[1] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object Referring in Videos with Language and Human Gaze. In *CVPR*, 2018.

[2] Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *EMNLP*, 2010.

[3] Herbert Paul Grice. Logic and conversation. In *P. Cole, J.L. Morgan, editors, Syntax and Semantics: Vol. 3: Speech Acts, pp. 4158. Academic Press, San Diego, CA*, 1975.

[4] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *CVPR*, 2016.

[5] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *ECCV*, 2016.

[6] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. In *CogSci*, 1995.

TestA

Ground Truth:
- White shirt.
- Boy in white shirt.

baseline:
Middle guy.
Our SR:
Man in white shirt.

Ground Truth:
- Man on the far left.
- Left man most visible.

baseline:
Man on left.
Our SR:
Guy in white on left.

TestB

Ground Truth:
- Clear glass between cow mugs.
- Made of glass on right of large plate.

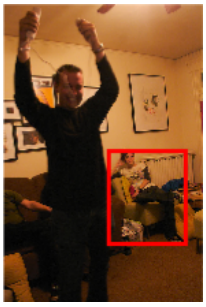
baseline:
Top right glass.
Our SR:
Glass behind plate in the middle of the picture.

Ground Truth:
- Suitcase with a bag on top.
- Far right suitcase.

baseline:
Right luggage.
Our SR:
Black suitcase on right.

Figure C. Generation example on RefCOCO by different methods with two ground-truth captions.

TestA



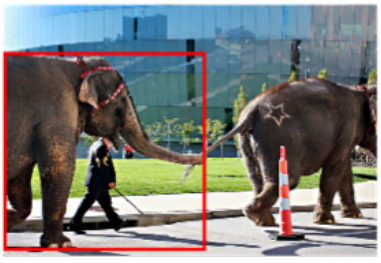
Ground Truth:
- Man sitting in yellow couch.
- Person in white.

baseline:
White shirt.
Our SR:
Woman sitting on couch.

Ground Truth:
- Red patterned suitcase.
- Brown backpack.

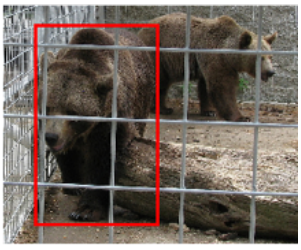
baseline:
Red chair.
Our SR:
Brown bag next to red bag.

TestB



Ground Truth:
- Trailing elephant.
- Trunk holding tail.

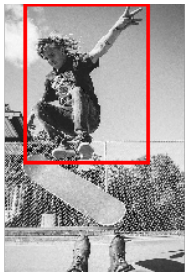
baseline:
Big elephant.
Our SR:
Elephant with trunk on it.



Ground Truth:
- Bear walking to us.
- Bear closest.

baseline:
Darker bear.
Our SR:
Bear looking at us.

Figure D. Generation example on RefCOCO+ by different methods with two ground-truth captions.

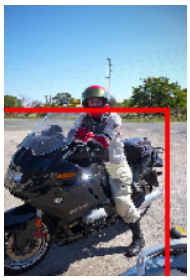


Ground Truth:
- A person in the air with his arm out, the arm has tattoos, with a corner of a skateboard in it.
- Shaggy haired man with tattoo on forearm in mid air doing a skateboard trick.

baseline:
A man doing a trick.
Our SR:
A woman doing a trick on a skateboard.

Ground Truth:
- A red bag that is hanging on the arm of a woman.

baseline:
The bottom of a woman in a red shirt.
Our SR:
A bag being held by a man in a purple shirt.



Ground Truth:
- The motorcycle that the woman is riding.
- The motorcycle that is fully shown.

baseline:
The motorcycle on the left.
Our SR:
A motorcycle with a man sitting on it.

Ground Truth:
- The zebra in the middle.
- A zebra with its back under the head of another zebra.

baseline:
The zebra on the left.
Our SR:
A zebra in the middle of two other zebras.

Ground Truth:
- A man in a plaid shirt sitting down.
- Man sitting on couch in living room.

baseline:
A man in a black shirt.
Our SR:
A person in a blue shirt sitting on a couch.

Ground Truth:
- A gate between three people in the stands and the playing field.

baseline:
The bench behind the fence.
Our SR:
The back of a bench behind a baseball game.

Ground Truth:
- A rich brown colored cow with an ear tag standing near some very small trees.
- A cow looking into the camera.

baseline:
The cow on the left.
Our SR:
A brown cow standing in front of another cow.

Ground Truth:
- The brown and white horse.
- The horse that isn't covered.

baseline:
The horse on the left.
Our SR:
A brown and white horse.

Figure E. Generation example on RefCOCOg by different methods with one or two ground-truth captions.

Ground Truth:
- A bald man in grey suit standing near a store glass window.
- Man in dark clothing standing near store front.

baseline:
A man in a dark suit talking on the phone.
Our SR:
A man in a blue suit standing next to a building.
Our SR + rank loss:
A man in a blue suit standing next to a building.

Ground Truth:
- Man in purple shirt with grey pants standing at the red curb.
- A woman wearing a purple shirt preparing to cross the street.

baseline:
A woman in a purple shirt.
Our SR:
A woman in a purple shirt waiting to cross the street.
Our SR + rank loss:
A woman in a purple shirt waiting to cross the street.

Ground Truth:
- A man wearing a dark shirt and shorts standing on a chair.
- A man wearing shorts and gray top standing on a chair.

baseline:
A man in a striped shirt sitting on the balcony.
Our SR:
A man in a black shirt and shorts standing next to a truck.
Our SR + rank loss:
A man in a black shirt and shorts standing next to a truck.

Ground Truth:
- A man on the phone.
- A man talking on his phone standing next to a man.

baseline:
A man in a white shirt talking on a cell phone.
Our SR:
A man in a black shirt talking on the phone.
Our SR + rank loss:
A man in a white shirt behind a car.

Ground Truth:
- A woman wearing purple standing.
- A person in purple top.

baseline:
A man in a purple shirt standing on the sidewalk.
Our SR:
A man in a purple jacket standing on the corner.
Our SR + rank loss:
A person in purple jacket.

Figure F. Generation example on RefGTA by different methods with two ground-truth captions.