ReSTR: Convolution-free Referring Image Segmentation Using Transformers
—— *Supplementary Material* ——

Namyup Kim[1] Dongwon Kim[1] Cuiling Lan[2] Wenjun Zeng[3] Suha Kwak[1]
[1]POSTECH [2]Microsoft Research Asia [3]EIT Institute for Advanced Study
http://cvlab.postech.ac.kr/research/restr/

This supplementary material presents experimental results omitted from the main paper due to the space limit. Sec. A analyzes performance according to language expression length compared with previous methods. In Sec. B, we investigate the sensitivity of our model to hyperparameters in terms of performance. Finally, Sec. C describes more qualitative results of our method on the Gref dataset.

A. Impact of the length of language expression

We present the detail analysis of performance according to language expression length in Fig. A1. Following [3], each test set on the dataset is split into four groups in terms of language expression length (*i.e.* sentence length), and each group is roughly equal size. Our method outperforms most previous methods except on the 1-5 length group of the Gref dataset, where the gap is marginally 1.2%p. Furthermore, our method has less performance degradation from the shortest to the longest sentence length group on four datasets than ACM [1], which is the most recent work. Although ACM is proposed to capture long-range dependencies, it still seems to struggle to understand the complex interaction between words of long language expressions. Therefore, the performance improvement of ACM mostly comes from its performance on the short sentence length groups. However, ReSTR shows the improvement of performance on most groups, which suggests that our model captures better long-range interactions of the language expression than the previous work.

B. Sensitivity to hyperparameters

We investigate the effect of the two hyperparameters, the loss balancing weights λ and the thresholding value τ, to generate patch-level labels. The results of our analysis are summarized in Fig. A2, in which we examine IoU of ReSTR by varying the values of the hyperparameters $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1\}$ and $\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. The results suggest that when λ is between 0.05 and 0.5, the performance of ReSTR is high and stable, thus insensitive to the hyperparameter setting. Note that the hyperparameter setting of ReSTR reported in the main paper (the underlined performance in Fig. A2) is not the best, although it outperforms all existing methods, as we do not tune the hyperparameters to optimize the test performance.

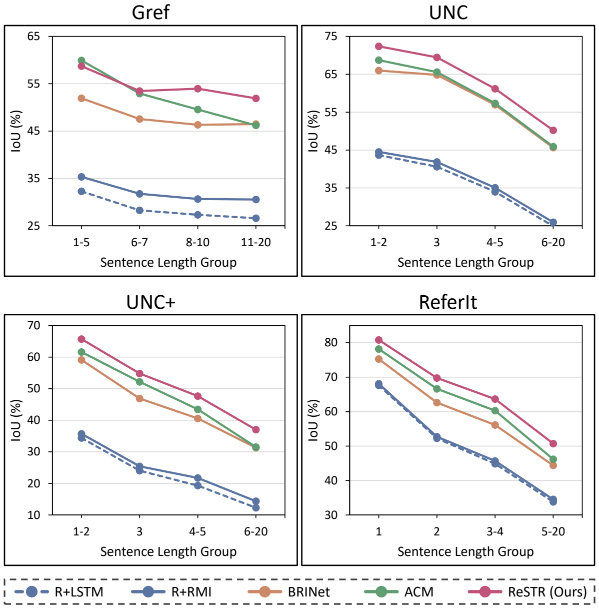


Figure A1. Performance in IoU(%) versus sentence length group on four datasets.

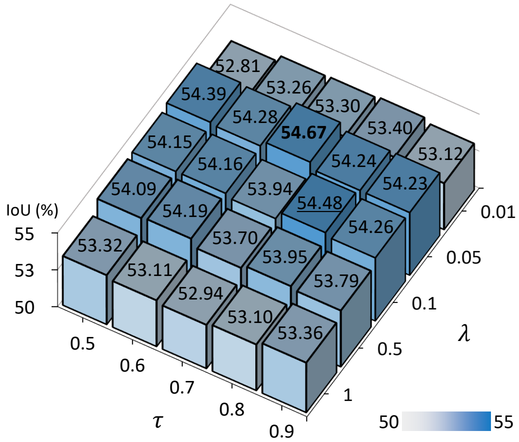


Figure A2. Performance in IoU versus λ and τ on the Gref *val* set.

C. More qualitative results

In Fig. A3 and Fig. A4, qualitative results of ReSTR on the Gref dataset are presented. Pixel-level predictions and the results post-processed with DenseCRF [2] are provided together. The results show that ReSTR successfully segments masks of the target entities described in various language expressions. For example, ReSTR predicts accurate masks for language expressions about non-human objects (Fig. A3), partially appeared objects (rows 1-3 in Fig. A4), and occluded objects (rows 5-7 in Fig. A4). Moreover, the qualitative results of the pixel-level prediction show that the segmentation decoder of ReSTR produces the fine-grained prediction as well as removes false positives in the patch-level prediction. As shown in Fig. A5, we also present more qualitative results of ReSTR according to varying language expressions for each image. The results show that ReSTR can comprehend various types of objects (rows 1-2 in Fig. A5), a sense of locality (rows 3-4 in Fig. A5), and fine-grained details (row 5 in Fig. A5).

References

[1] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2011.

[3] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

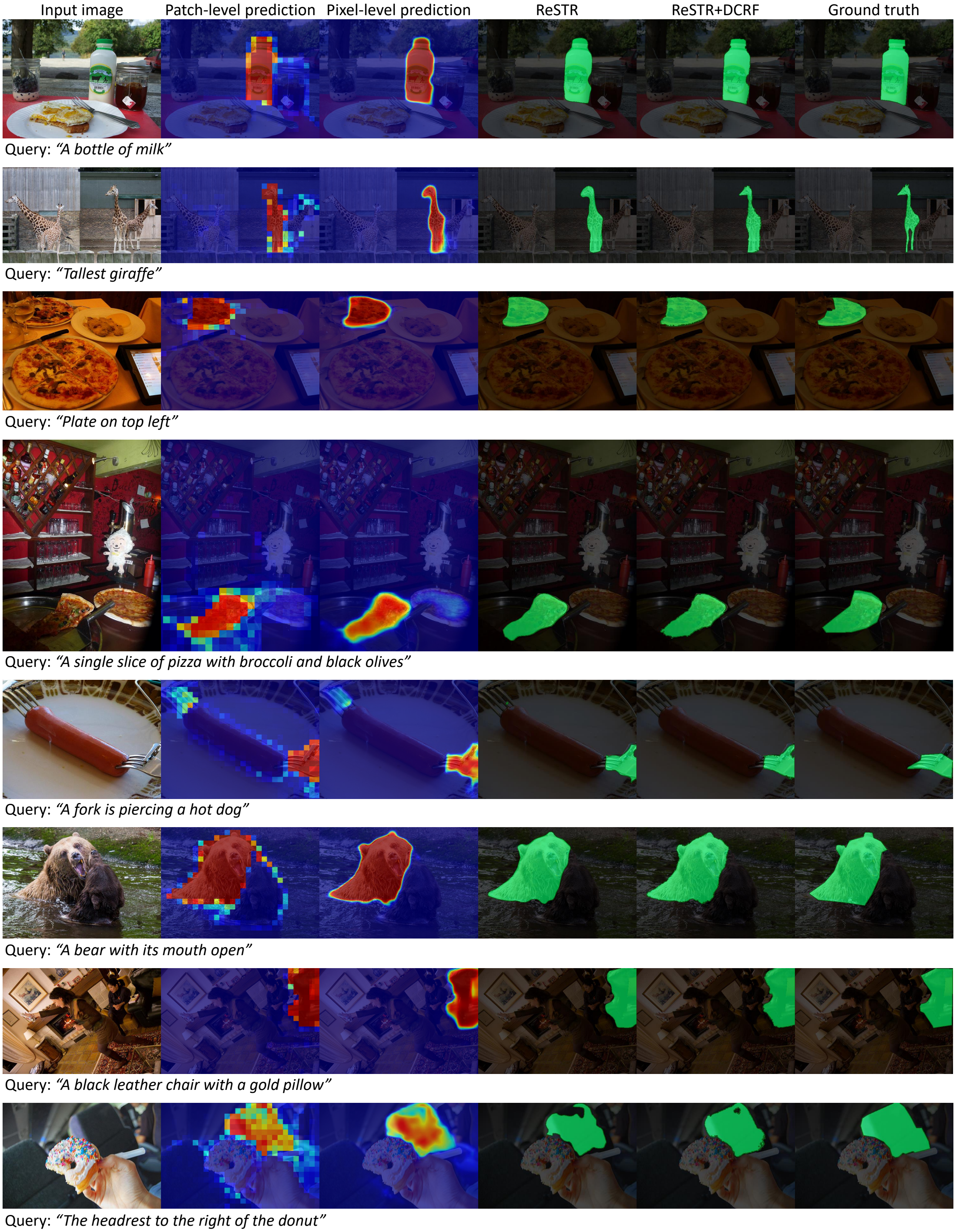
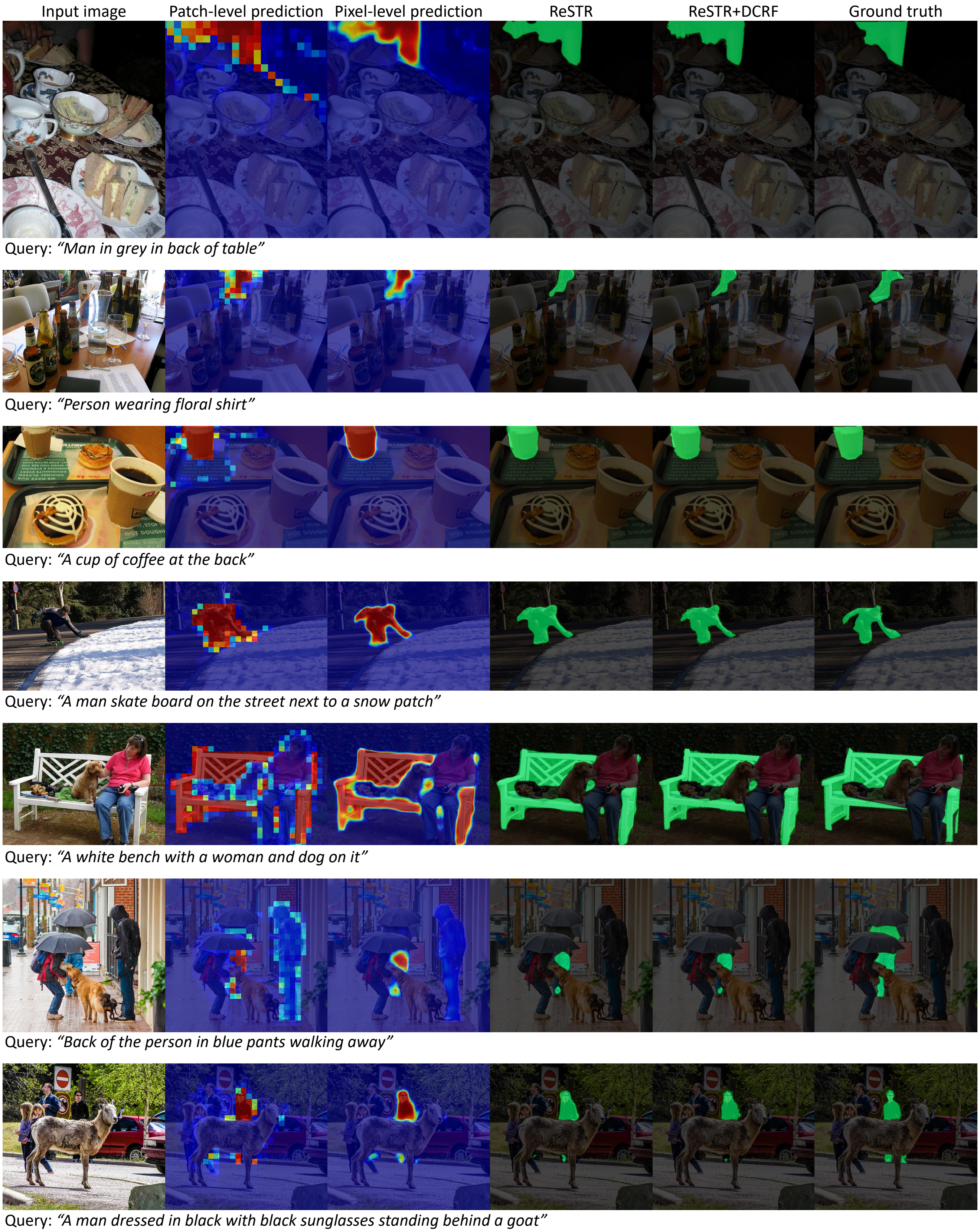

Input image
Patch-level prediction
Pixel-level prediction
ReSTR
ReSTR+DCRF
Ground truth
Query: *“A bottle of milk”*
Query: *“Tallest giraffe”*
Query: *“Plate on top left”*
Query: *“A single slice of pizza with broccoli and black olives”*
Query: *“A fork is piercing a hot dog”*
Query: *“A bear with its mouth open”*
Query: *“A black leather chair with a gold pillow”*
Query: *“The headrest to the right of the donut”*

Figure A3. Qualitative results of ReSTR on the Gref *val* set.



Query: *“Man in grey in back of table”*

Query: *“Person wearing floral shirt”*

Query: *“A cup of coffee at the back”*

Query: *“A man skate board on the street next to a snow patch”*

Query: *“A white bench with a woman and dog on it”*

Query: *“Back of the person in blue pants walking away”*

Query: *“A man dressed in black with black sunglasses standing behind a goat”*

Figure A4. Qualitative results of ReSTR on the Gref *val* set.

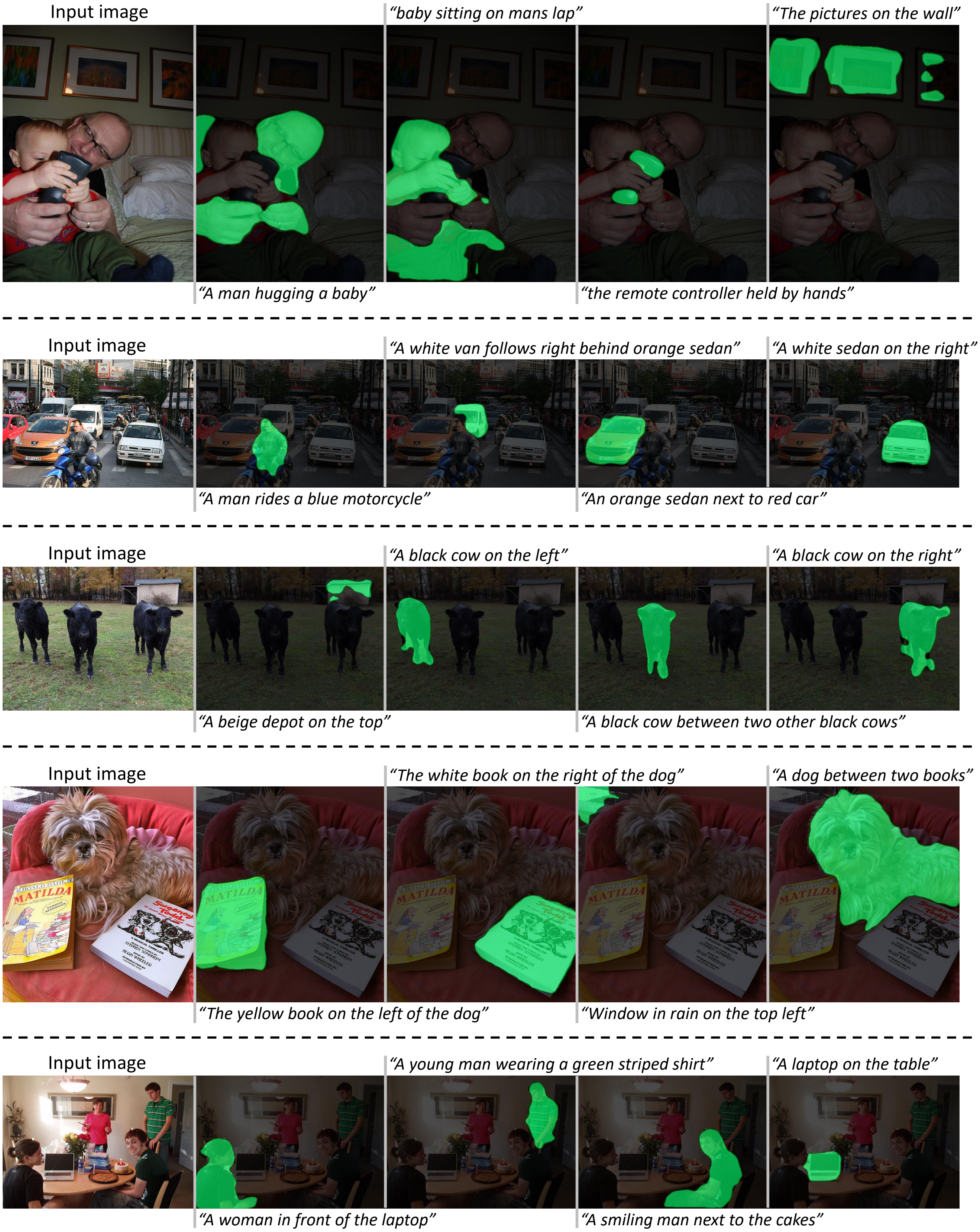

Input image
“baby sitting on mans lap”
“The pictures on the wall”
“A man hugging a baby”
“the remote controller held by hands”
Input image
“A white van follows right behind orange sedan”
“A white sedan on the right”
“A man rides a blue motorcycle”
“An orange sedan next to red car”
Input image
“A black cow on the left”
“A black cow on the right”
“A beige depot on the top”
“A black cow between two other black cows”
Input image
“The white book on the right of the dog”
“A dog between two books”
“The yellow book on the left of the dog”
“Window in rain on the top left”
Input image
“A young man wearing a green striped shirt”
“A laptop on the table”
“A woman in front of the laptop”
“A smiling man next to the cakes”

Figure A5. Qualitative results of ReSTR according to different language expression queries for each image on the Gref *val* set.