Recurrent Multimodal Interaction for Referring Image Segmentation (Supplementary Material)

Chenxi Liu[1] Zhe Lin[2] Xiaohui Shen[2] Jimei Yang[2] Xin Lu[2] Alan Yuille[1]
Johns Hopkins University[1] Adobe Research[2]
{cxliu, alan.yuille}@jhu.edu {zlin, xshen, jimyang, xinl}@adobe.com

In section 1 of this supplementary material, we provide the full table of segmentation performance on the four datasets. In addition to the IOU reported in the main paper, we also report Precision@X ($X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$) which means the percentage of images with IOU higher than X. This is consistent with previous work [1, 2] to allow for comparison. We show that the improvement of our RMI model over the baseline is solid.

In section 2 we provide more visualization of intermediate segmentation results. We show from the progression how our RMI model can better memorize long-term information instead of being distracted by later words.

In section 3 we provide more visualization of final segmentation results. We show that our RMI model can usually generate better segmentation masks, and also DenseCRF [3] offers more visually pleasing results through refinement.

References

[1] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, pages 108–124. Springer, 2016. 1, 3

[2] R. Hu, M. Rohrbach, S. Venugopalan, and T. Darrell. Utilizing large scale vision and text datasets for image segmentation from referring expressions. *CoRR*, abs/1608.08305, 2016. 1

[3] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 1

1. Full Table of Segmentation Performance

Table 1.1: Comparison of segmentation performance on Google-Ref.

Model	Set	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IOU
[2]	val	15.25	8.37	3.75	1.29	0.06	28.14
R+LSTM	val	17.57	10.37	4.52	**1.26**	0.05	28.60
R+RMI	val	**20.32**	**11.62**	**5.02**	1.13	**0.09**	**32.06**
R+LSTM+DCRF	val	20.50	14.16	7.86	3.01	0.38	28.94
R+RMI+DCRF	val	**23.87**	**16.12**	**9.04**	**3.03**	**0.42**	**32.85**
D+LSTM	val	25.66	18.23	**10.82**	4.17	0.64	33.08
D+RMI	val	**26.19**	**18.46**	10.68	**4.28**	**0.73**	**34.40**
D+LSTM+DCRF	val	26.71	20.50	13.69	**7.11**	1.32	33.11
D+RMI+DCRF	val	**27.77**	**21.06**	**13.92**	6.83	**1.43**	**34.52**

Table 1.2: Comparison of segmentation performance on UNC.

Model	Set	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IOU
R+LSTM	val	30.98	18.65	8.23	2.12	**0.10**	38.74
R+RMI	val	**32.12**	**19.49**	**8.60**	**2.12**	0.03	**39.74**
R+LSTM+DCRF	val	34.83	24.12	13.97	5.24	**0.59**	39.88
R+RMI+DCRF	val	**36.38**	**25.36**	**14.53**	**5.58**	0.56	**41.17**
D+LSTM	val	39.93	28.19	18.04	**7.45**	**0.91**	43.27
D+RMI	val	**41.27**	**29.71**	**18.41**	7.37	0.76	**44.33**
D+LSTM+DCRF	val	41.42	31.86	22.15	11.90	**2.36**	43.97
D+RMI+DCRF	val	**42.99**	**33.24**	**22.75**	**12.11**	2.23	**45.18**
R+LSTM	testA	30.49	18.63	**8.71**	1.96	**0.09**	39.18
R+RMI	testA	**31.36**	**18.95**	8.34	**2.02**	0.04	**39.99**
R+LSTM+DCRF	testA	34.45	24.62	14.39	5.43	**0.50**	40.44
R+RMI+DCRF	testA	**35.92**	**24.94**	**14.50**	**5.73**	0.44	**41.35**
D+LSTM	testA	39.77	29.61	18.83	8.01	0.88	43.60
D+RMI	testA	**40.68**	**30.14**	**18.99**	**8.03**	**0.88**	**44.74**
D+LSTM+DCRF	testA	41.52	32.61	23.63	12.59	2.19	44.25
D+RMI+DCRF	testA	**42.99**	**33.59**	**23.69**	**12.94**	**2.44**	**45.69**
R+LSTM	testB	32.54	20.80	10.21	2.12	**0.16**	39.01
R+RMI	testB	**35.00**	**22.30**	**10.40**	**2.67**	0.12	**40.44**
R+LSTM+DCRF	testB	35.53	25.28	15.76	6.08	**0.65**	40.07
R+RMI+DCRF	testB	**38.31**	**27.34**	**16.96**	**7.01**	0.55	**41.87**
D+LSTM	testB	40.00	29.03	17.41	7.56	**0.92**	43.31
D+RMI	testB	**42.75**	**30.40**	**18.19**	**7.83**	0.86	**44.63**
D+LSTM+DCRF	testB	41.86	31.66	21.55	11.25	2.30	44.07
D+RMI+DCRF	testB	**44.99**	**34.21**	**22.69**	**11.84**	**2.65**	**45.57**

Table 1.3: Comparison of segmentation performance on UNC+.

Model	Set	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IOU
R+LSTM	val	13.16	6.80	2.40	0.51	0.02	26.25
R+RMI	val	**14.49**	**7.57**	**2.56**	**0.59**	**0.06**	**27.85**
R+LSTM+DCRF	val	16.17	9.76	4.82	1.57	0.11	26.29
R+RMI+DCRF	val	**17.88**	**11.12**	**5.34**	**1.62**	**0.15**	**28.26**
D+LSTM	val	16.58	10.64	5.30	1.78	0.17	28.42
D+RMI	val	**18.39**	**11.50**	**5.86**	**1.85**	**0.20**	**29.91**
D+LSTM+DCRF	val	18.36	12.88	7.85	3.37	0.62	28.07
D+RMI+DCRF	val	**20.52**	**14.02**	**8.46**	**3.77**	**0.62**	**29.86**
R+LSTM	testA	14.01	6.97	2.41	0.40	**0.02**	26.95
R+RMI	testA	**15.33**	**7.75**	**2.50**	**0.40**	0.00	**28.69**
R+LSTM+DCRF	testA	17.22	10.76	5.31	**1.82**	0.10	27.03
R+RMI+DCRF	testA	**19.21**	**12.10**	**5.80**	1.73	**0.10**	**29.16**
D+LSTM	testA	17.31	11.11	5.78	1.59	0.19	28.57
D+RMI	testA	**18.76**	**11.67**	**6.08**	**1.78**	**0.26**	**30.37**
D+LSTM+DCRF	testA	18.97	13.46	8.61	3.63	**0.51**	28.29
D+RMI+DCRF	testA	**21.22**	**14.43**	**8.99**	**3.91**	0.49	**30.48**
R+LSTM	testB	13.15	6.91	2.95	0.51	**0.14**	24.57
R+RMI	testB	**14.15**	**7.92**	**3.46**	**0.82**	0.08	**26.65**
R+LSTM+DCRF	testB	14.99	9.47	5.22	1.76	0.20	24.44
R+RMI+DCRF	testB	**16.69**	**10.62**	**5.89**	**2.25**	**0.27**	**26.86**
D+LSTM	testB	17.26	10.88	6.05	2.25	0.27	27.70
D+RMI	testB	**19.08**	**12.11**	**6.44**	**2.70**	**0.31**	**29.43**
D+LSTM+DCRF	testB	18.41	13.17	8.14	3.95	0.59	27.44
D+RMI+DCRF	testB	**20.78**	**14.56**	**8.80**	**4.58**	**0.80**	**29.50**

Table 1.4: Comparison of segmentation performance on ReferItGame.

Model	Set	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	IOU
[1]	test	34.02	26.71	19.32	11.63	3.92	48.03
R+LSTM	test	38.17	29.73	20.91	**11.99**	**3.81**	54.01
R+RMI	test	**39.44**	**30.50**	**21.20**	11.95	3.64	**54.55**
R+LSTM+DCRF	test	40.56	33.40	25.27	16.08	**6.09**	55.90
R+RMI+DCRF	test	**42.21**	**34.53**	**25.72**	**16.09**	6.05	**56.61**
D+LSTM	test	43.86	35.75	26.65	16.75	**6.47**	56.83
D+RMI	test	**44.33**	**36.13**	**27.20**	**16.99**	6.43	**57.34**
D+LSTM+DCRF	test	45.26	38.37	30.20	20.41	**8.56**	58.20
D+RMI+DCRF	test	**46.08**	**38.90**	**30.77**	**20.62**	8.54	**58.73**

2. Visualization of Intermediate Segmentation

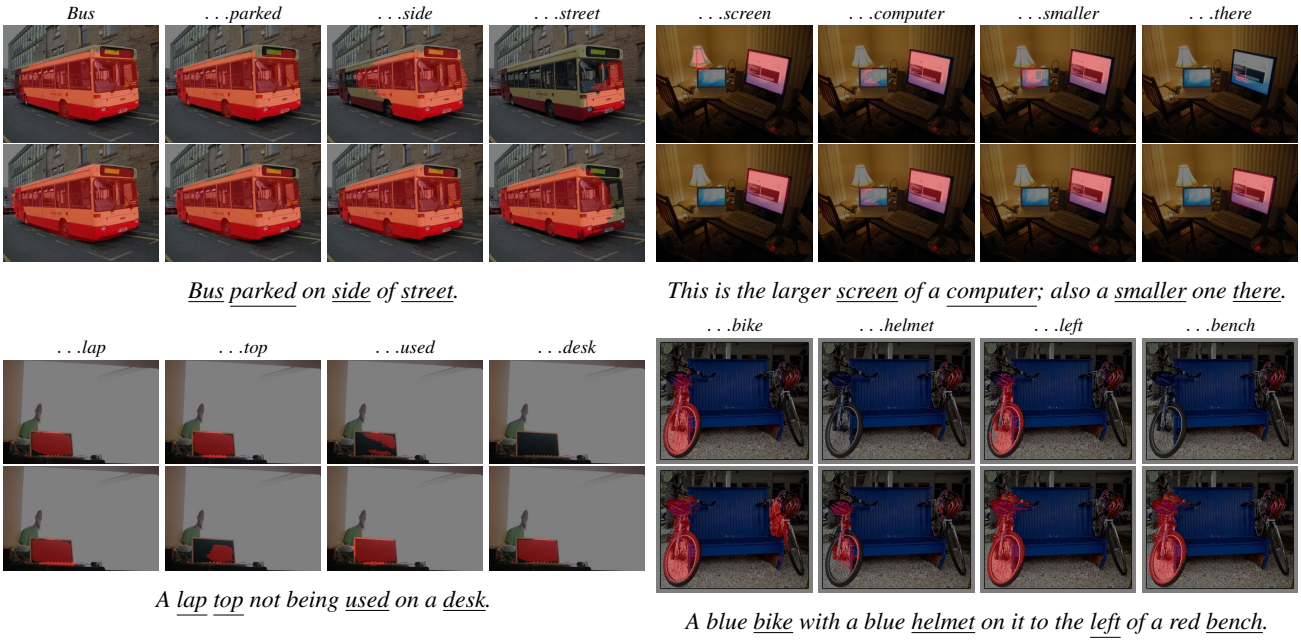


Figure 2.1: Comparison of D+LSTM+DCRF (first row) and D+RMI+DCRF (second row) on Google-Ref. Each column shows segmentation result until after reading the underlined word.

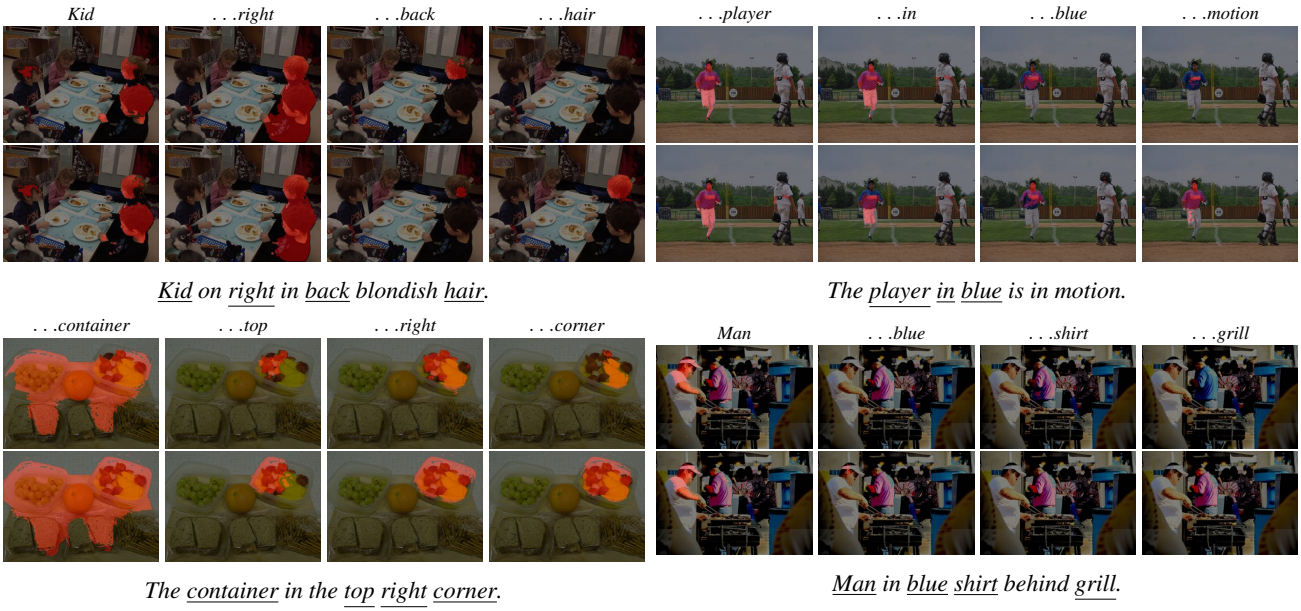


Figure 2.2: Comparison of D+LSTM+DCRF (first row) and D+RMI+DCRF (second row) on UNC. Each column shows segmentation result until after reading the underlined word.

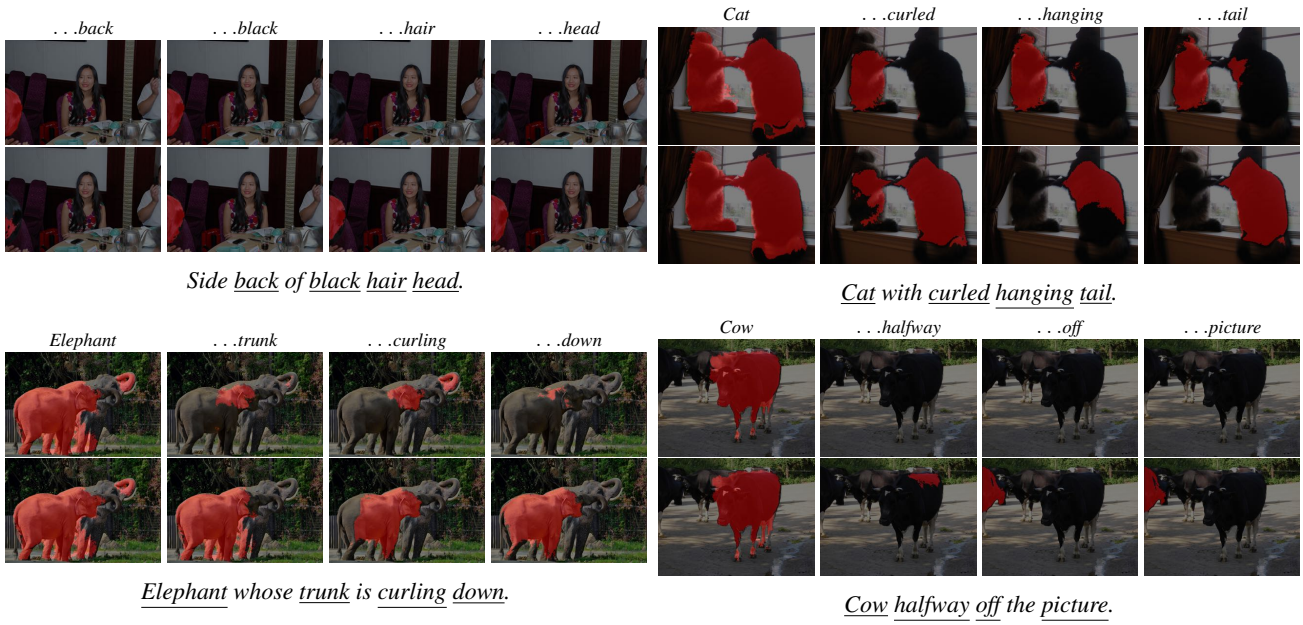


Figure 2.3: Comparison of D+LSTM+DCRF (first row) and D+RMI+DCRF (second row) on UNC+. Each column shows segmentation result until after reading the underlined word.

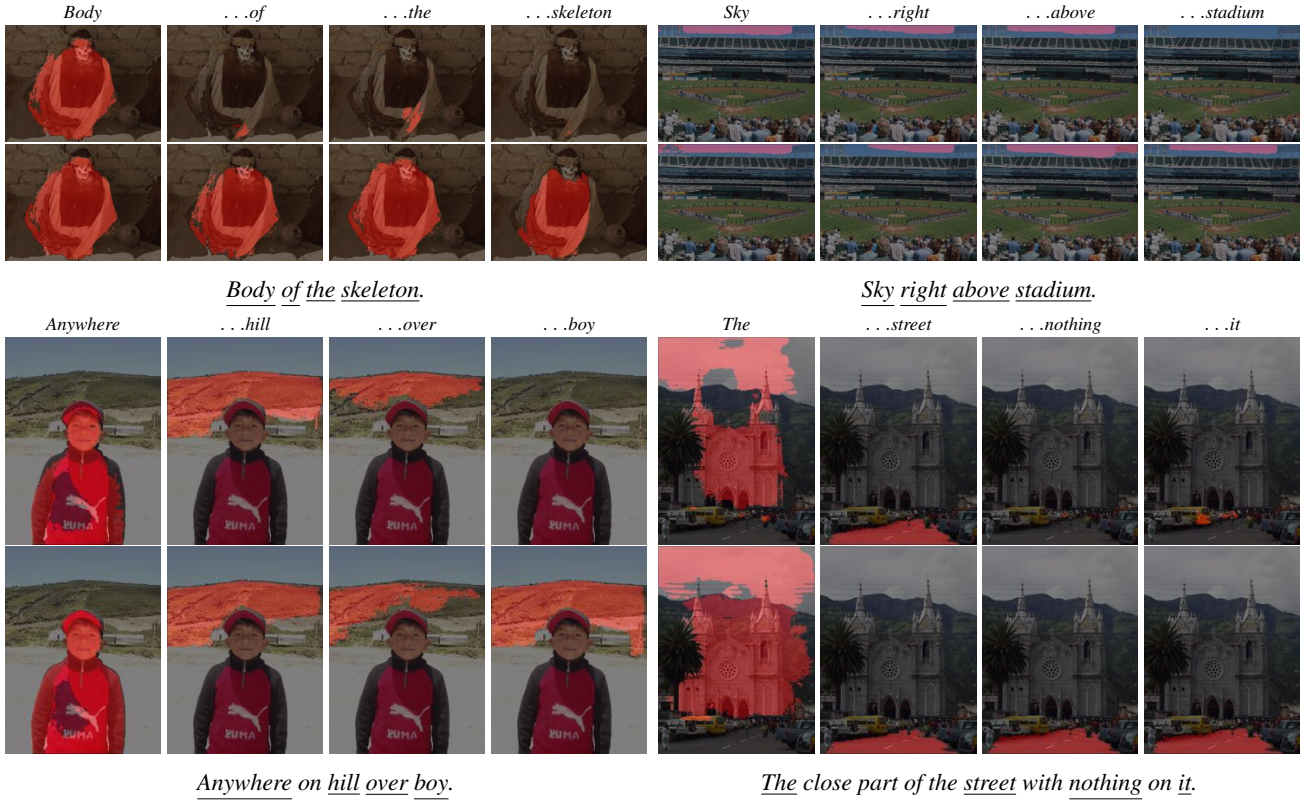
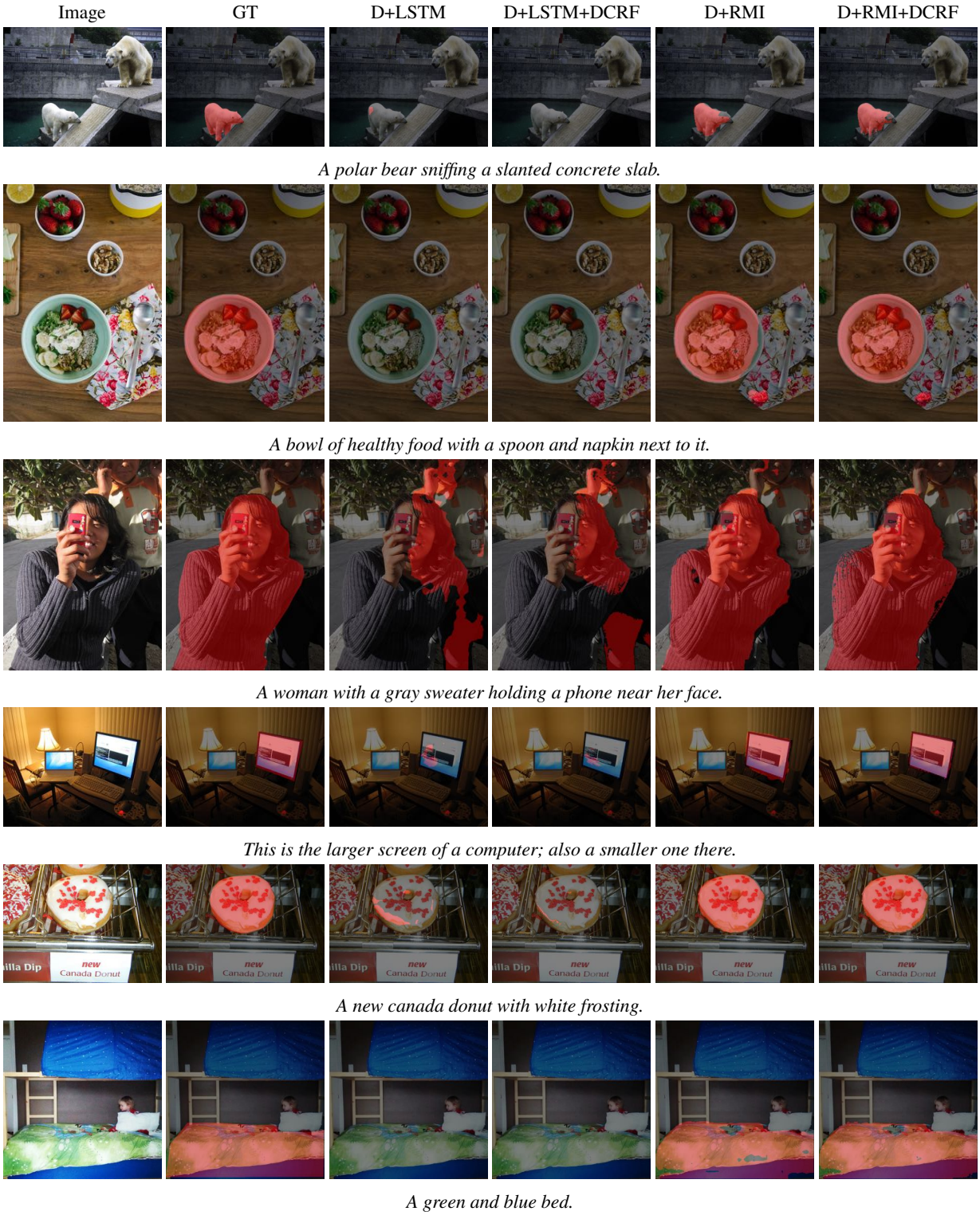


Figure 2.4: Comparison of D+LSTM+DCRF (first row) and D+RMI+DCRF (second row) on ReferItGame. Each column shows segmentation result until after reading the underlined word.

3. Visualization of Final Segmentation



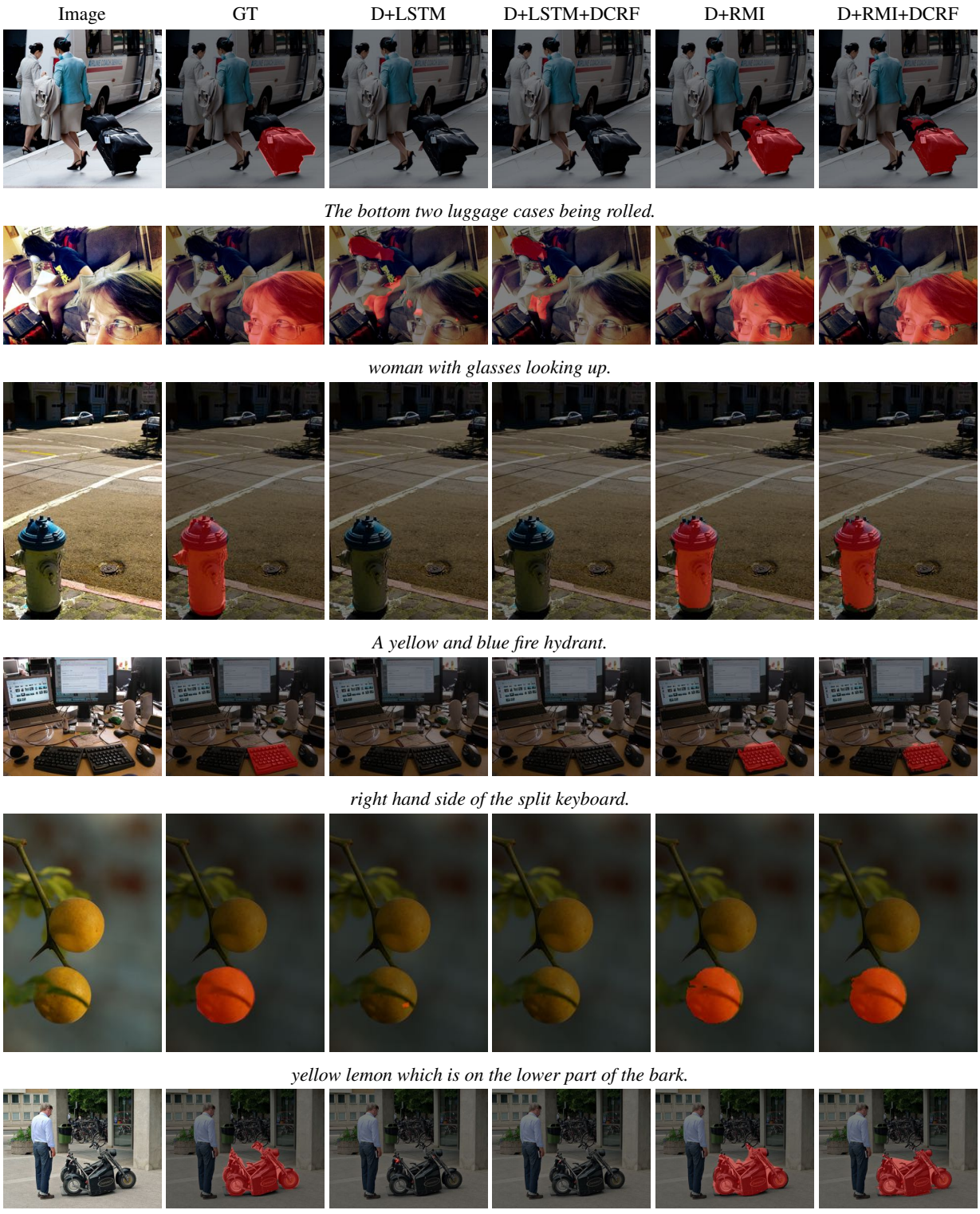


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
The bottom two luggage cases being rolled.
woman with glasses looking up.
A yellow and blue fire hydrant.
right hand side of the split keyboard.
yellow lemon which is on the lower part of the bark.
a weird black bike that sticks to the ground and there are tires for it.

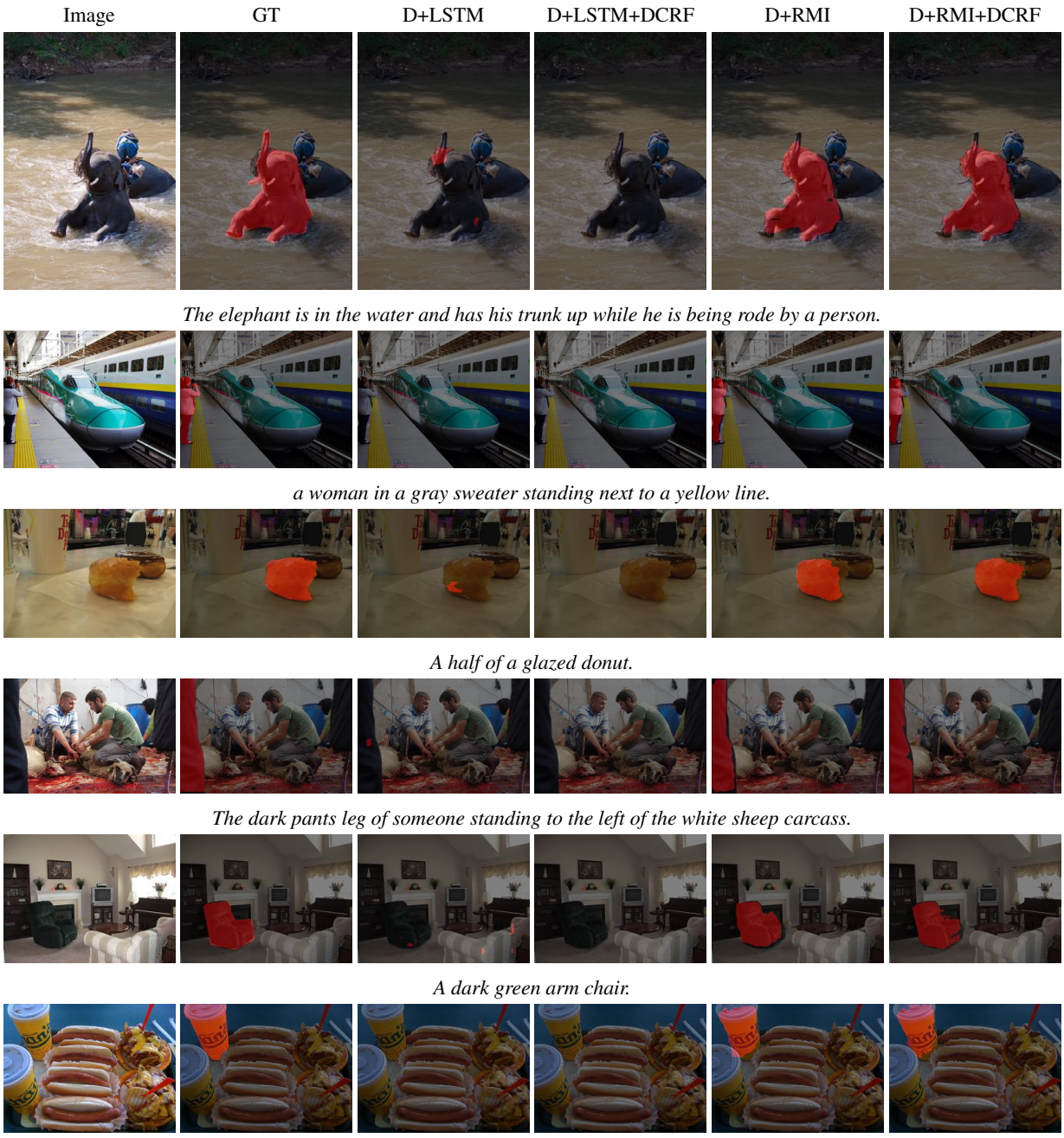


Figure 3.1: Qualitative results of referring image segmentation on Google-Ref.

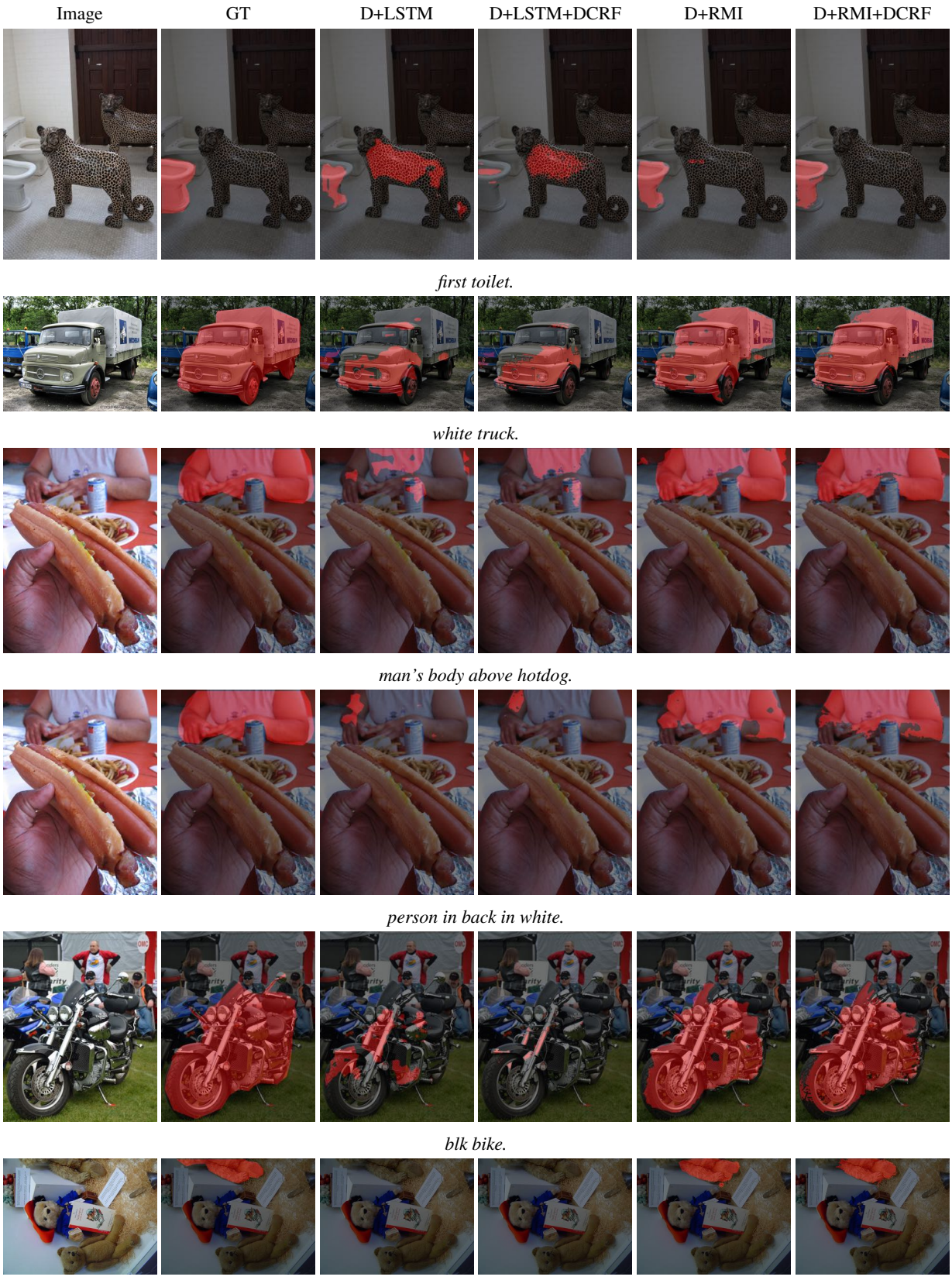


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
first toilet.
white truck.
man's body above hotdog.
person in back in white.
blk bike.
bear arm top edge of photo.

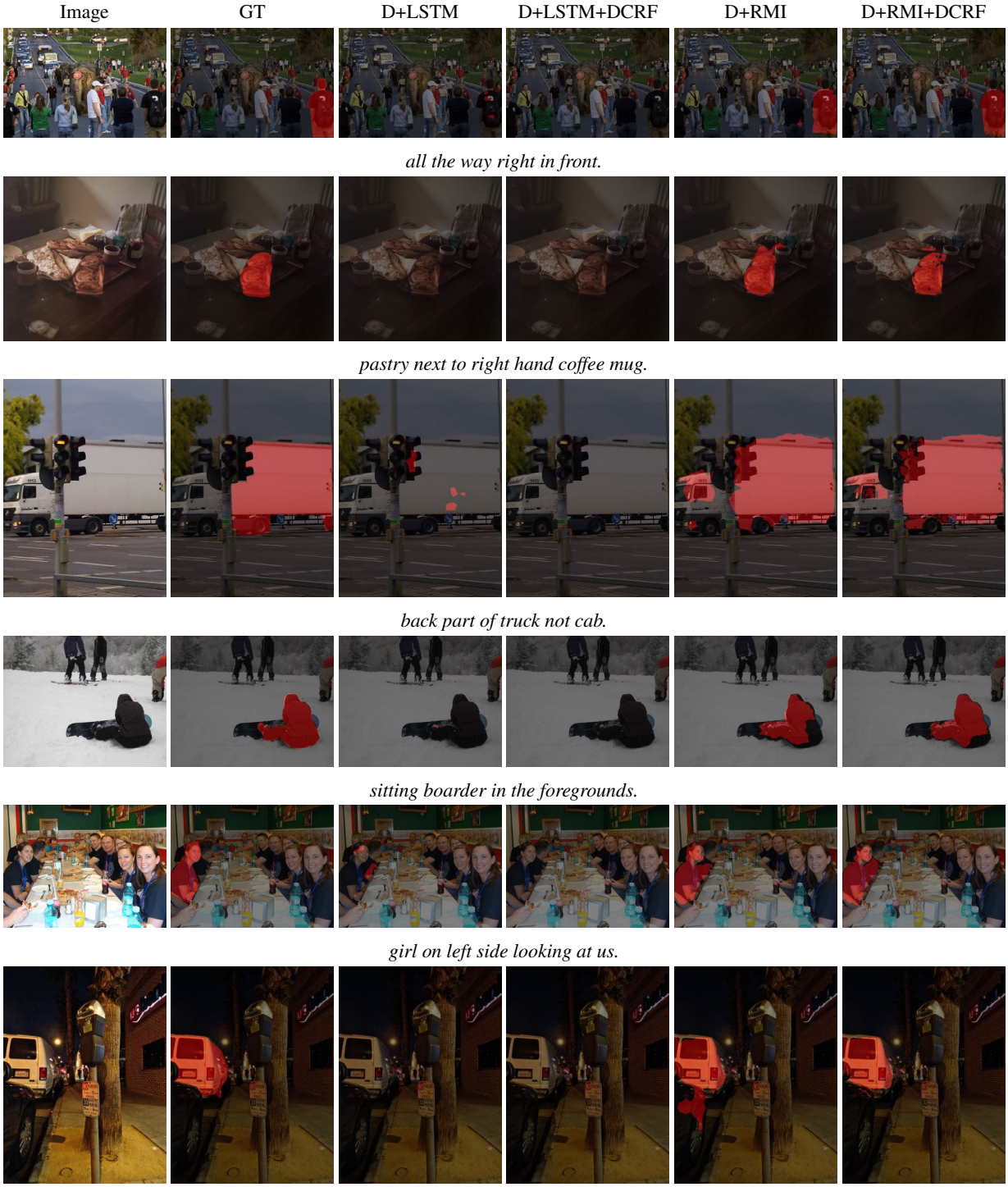


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
all the way right in front.
pastry next to right hand coffee mug.
back part of truck not cab.
sitting boarder in the foregrounds.
girl on left side looking at us.
van.

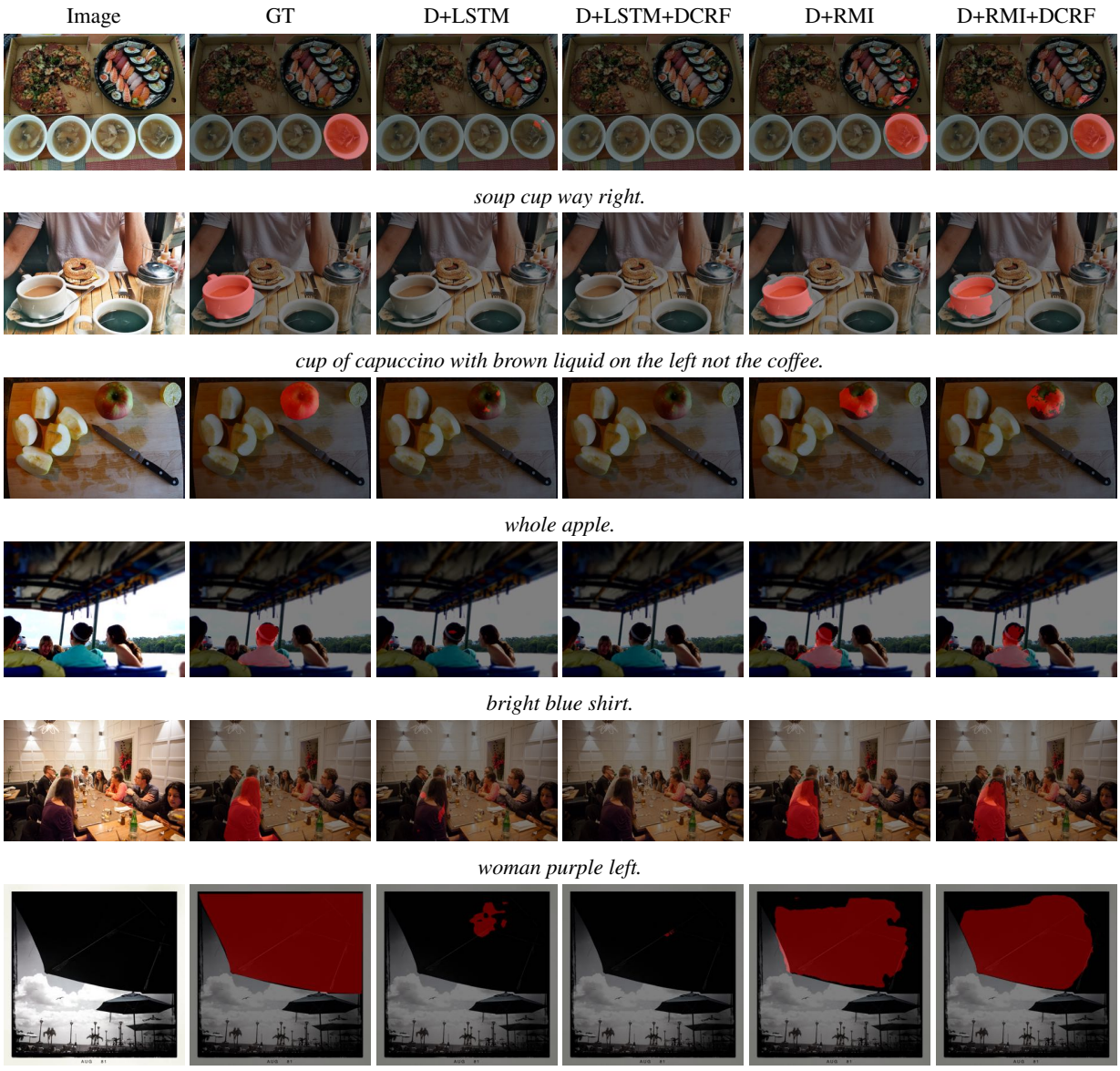


Figure 3.2: Qualitative results of referring image segmentation on UNC.

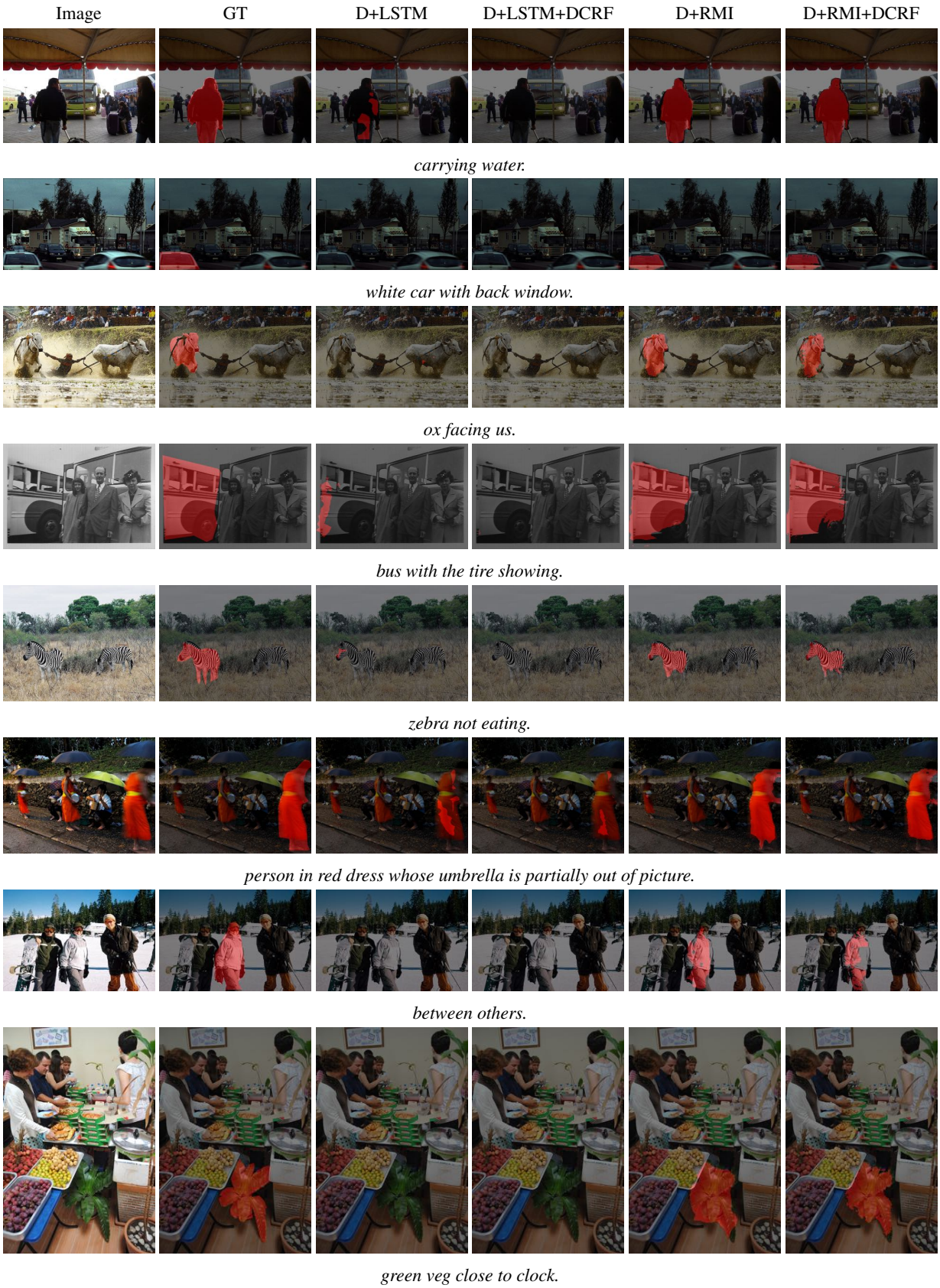


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
carrying water.
white car with back window.
ox facing us.
bus with the tire showing.
zebra not eating.
person in red dress whose umbrella is partially out of picture.
between others.
green veg close to clock.

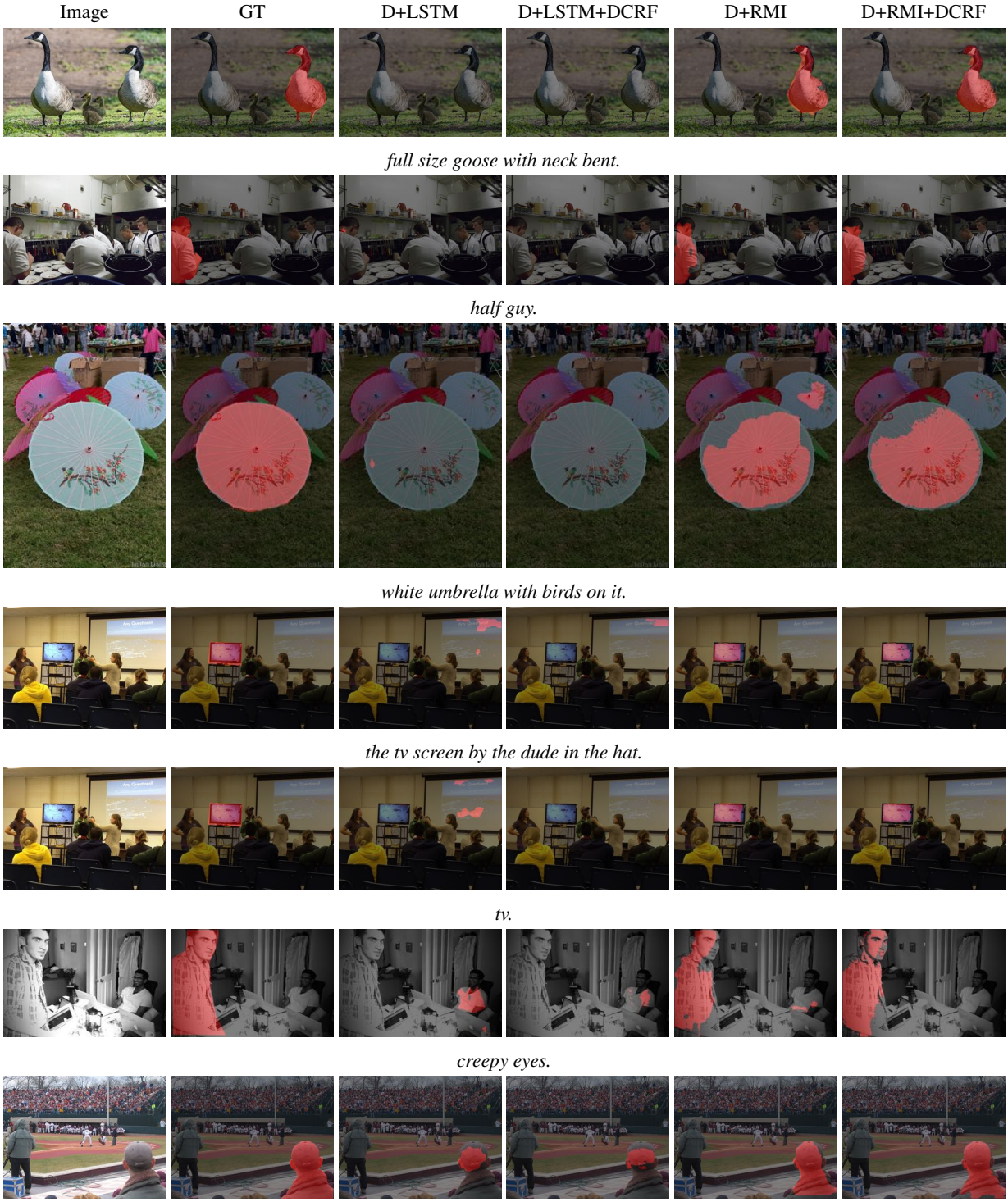


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
full size goose with neck bent.
half guy.
white umbrella with birds on it.
the tv screen by the dude in the hat.
tv.
creepy eyes.
guy with tan hat and fur collar.

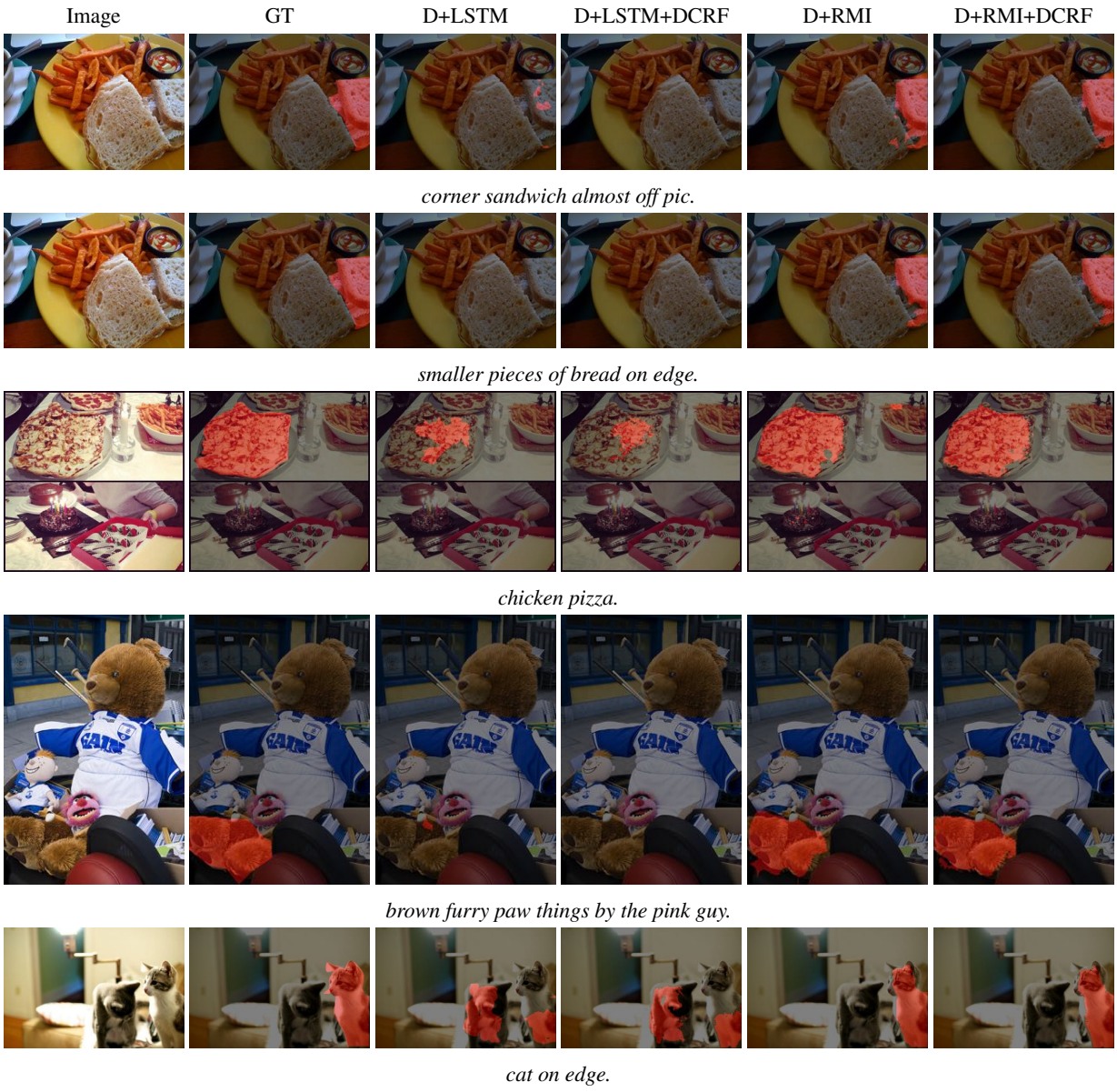


Figure 3.3: Qualitative results of referring image segmentation on UNC+.

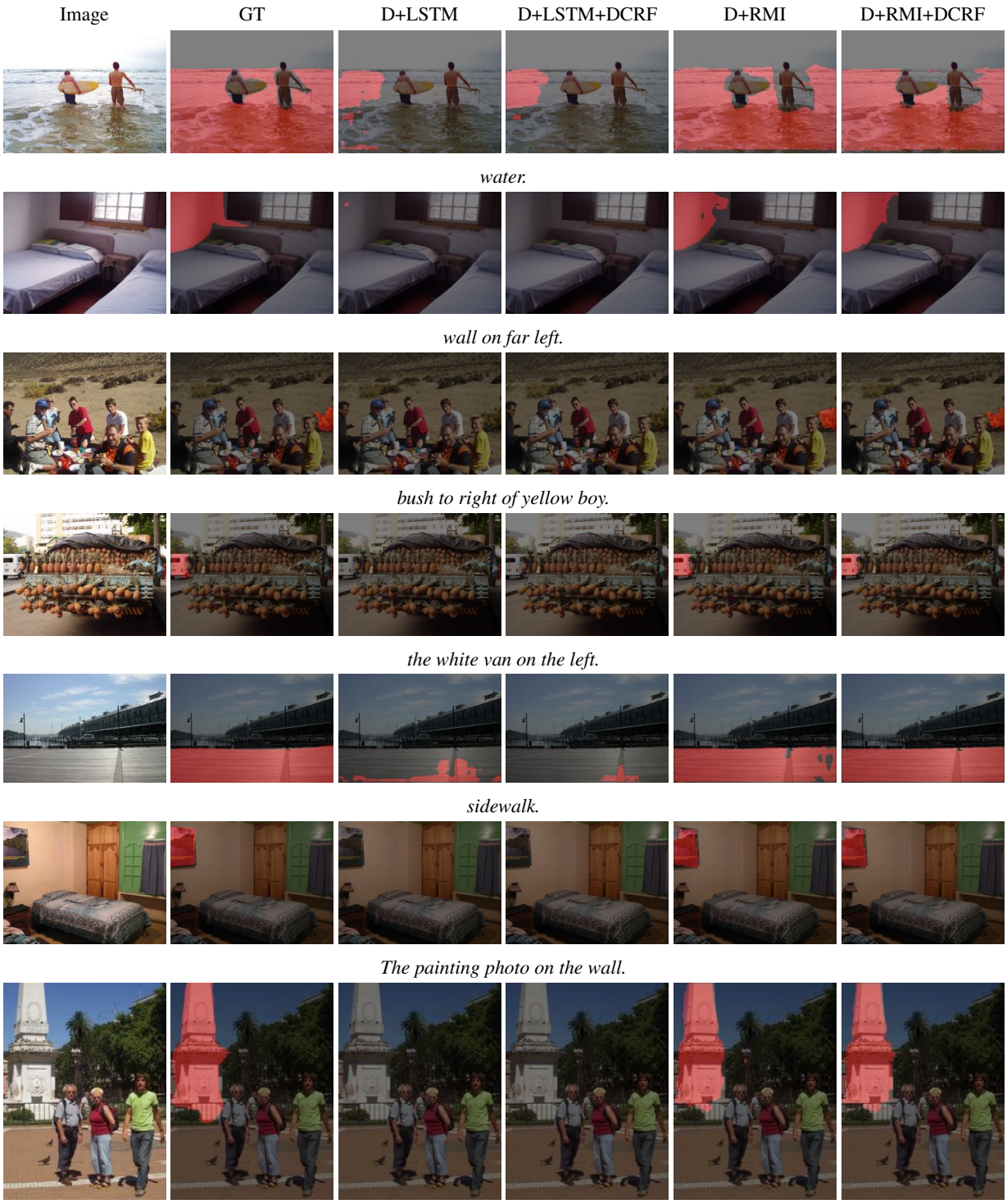


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
water.
wall on far left.
bush to right of yellow boy.
the white van on the left.
sidewalk.
The painting photo on the wall.
the white pillar thing.

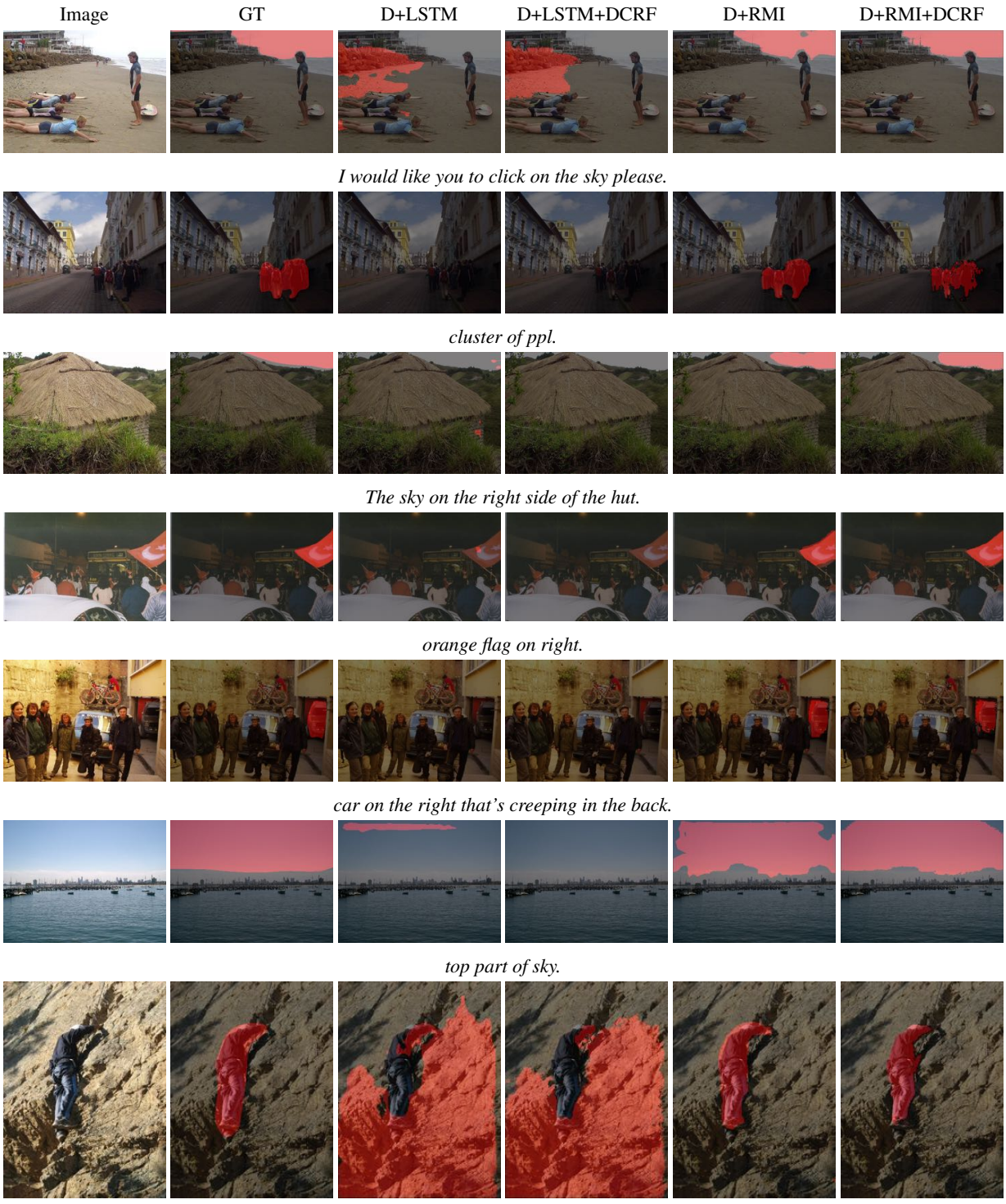


Image
GT
D+LSTM
D+LSTM+DCRF
D+RMI
D+RMI+DCRF
I would like you to click on the sky please.
cluster of ppl.
The sky on the right side of the hut.
orange flag on right.
car on the right that's creeping in the back.
top part of sky.
climber.

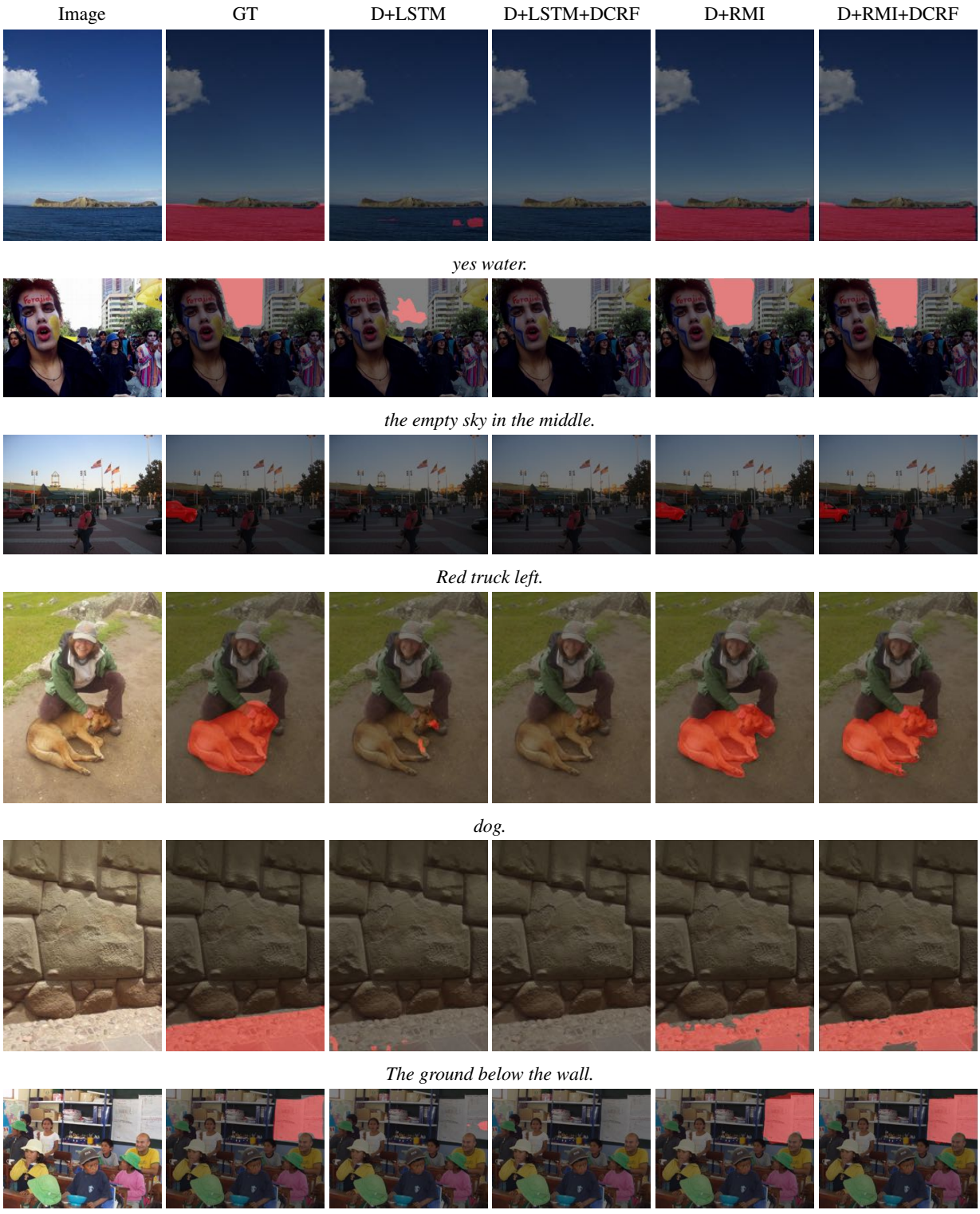


Figure 3.4: Qualitative results of referring image segmentation on ReferItGame.