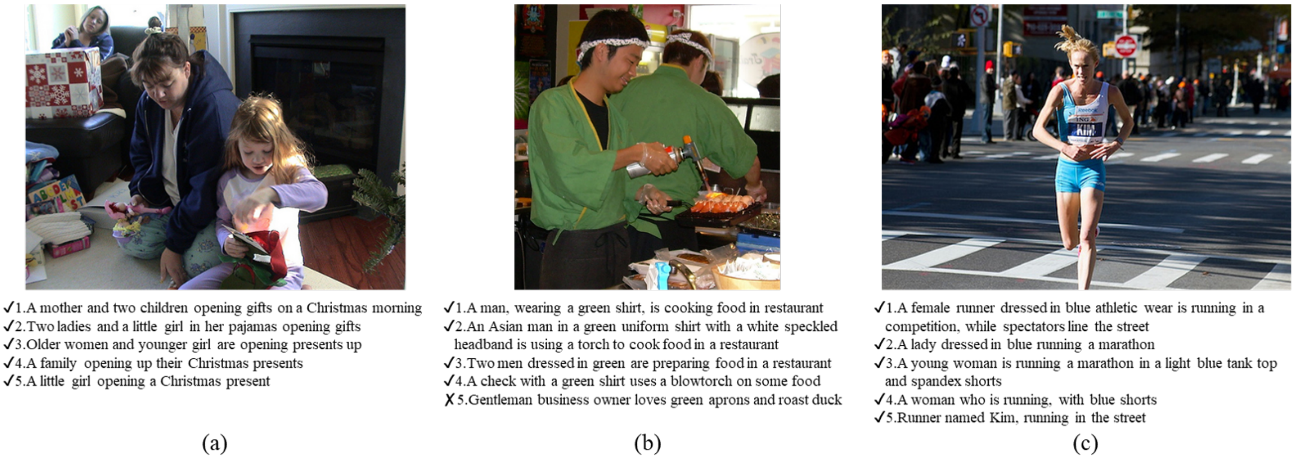

✓1.A mother and two children opening gifts on a Christmas morning
✓2.Two ladies and a little girl in her pajamas opening gifts
✓3.Older women and younger girl are opening presents up
✓4.A family opening up their Christmas presents
✓5.A little girl opening a Christmas present

(a)

✓1.A man, wearing a green shirt, is cooking food in restaurant
✓2.An Asian man in a green uniform shirt with a white speckled headband is using a torch to cook food in a restaurant
✓3.Two men dressed in green are preparing food in a restaurant
✓4.A check with a green shirt uses a blowtorch on some food
✗5.Gentleman business owner loves green aprons and roast duck

(b)

✓1.A female runner dressed in blue athletic wear is running in a competition, while spectators line the street
✓2.A lady dressed in blue running a marathon
✓3.A young woman is running a marathon in a light blue tank top and spandex shorts
✓4.A woman who is running, with blue shorts
✓5.Runner named Kim, running in the street

(c)

Figure 1: Visualization of text retrieval result on Flickr30K. The top 5 ranked texts are shown for each image query.

A. Text Retrieval Case Study

We also visualize the text retrieval results using our proposed GraDual Figure 1. We demonstrate the same text retrieval samples as in SCAN [1]. Same to SCAN, our GraDual retrieves texts (a)1-(a)5 and (b)1-(b)4 all correctly for Figure 1(a) and (b) but with slightly different order due to the variance of representation and alignment. For Figure 1(c), SCAN fails to retrieve the (c)5 correctly whereas our GraDual does it correctly after integrating the cross-modal contextual information into our initial modality representation via GraDual.

References

[1] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.