Supplementary Materials

1 Implementation Details

Encoder In this work, we apply ResNet-152 [He *et al.*, 2016] as our visual encoder model pre-trained using the ImageNet dataset [Deng *et al.*, 2009]. In comparison with other CNN models, ResNet-152 has shown more effective results on different image-caption datasets [Chen *et al.*, 2017]. We specifically use its Res5c layer to extract the spatial features of an image. The layer gives us $7 \times 7 \times 2048$ feature map converted to 49×2048 representing 49 semantic-based regions with 2048 dimensions.

Vocabulary Our vocabulary has 9703 words, coming form both the MSCOCO and SentiCap datasets, for all our models. Each word is embedded into a 300 dimensional vector.

Generator and Discriminator The size of the hidden state and the memory cell of our LSTM is set to 512. For the caption generator, we use the Adam function [Kingma and Ba, 2014] for optimization and set the learning rate to 0.0001. We set the the size of our mini-batches to 64. To optimize the caption discriminator, we use the RMSprop solver [Tieleman and Hinton, 2012] and clip the weights to $[-0.01, 0.01]$. The mini-batches are fixed to 80 for the discriminator. We apply Monte Carlo search 5 times (Eq (2)). We set λ_1 and λ_2 to 1.0 and 0.1 in Eq (3) and (8), respectively. During the adversarial training, we alternate between Eq (8) and Eq (10) to optimize the generator and the discriminator, respectively. We particularly operate a single gradient decent phase on the generator (g steps) and 3 gradient phases (d steps) on the discriminator every time. The models are trained for 20 epochs to converge. The METEOR metric is used to select the model with the best performance on the validation sets of positive and negative datasets of SentiCap because it has a close correlation with human judgments and is less computationally expensive than SPICE which requires dependency parsing [Anderson *et al.*, 2016].

2 Additional Generated Captions

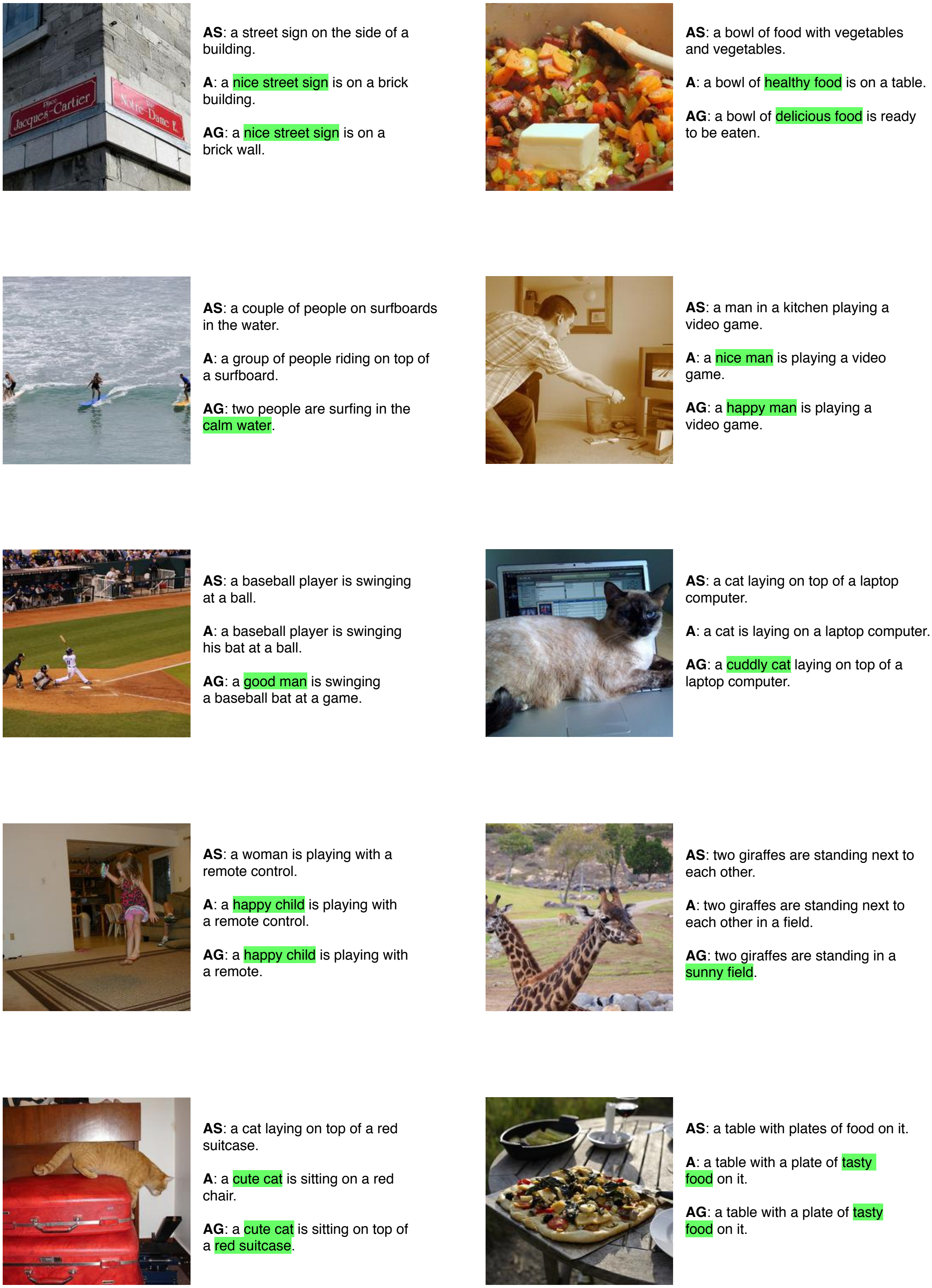
In Figure 1, we compare additional positive samples on the SentiCap positive dataset. The samples show that ATTEND-GAN generates positive captions which are correlated with the visual content and include a wider collection of positive adjectives. For example, for the second image in row 1, ATTEND-GAN generates “a bowl of delicious food is ready to be eaten” and for the first image in row 2, it generates “two people are surfing in the calm water”. For the image, other models cannot even generate positive captions. Similarly, ATTEND-GAN can generate positive image captions in the first and second images in row 3. Sometimes, both ATTEND-GAN and ATTEND-GAN$_{-A}$ generate a similar caption such as the second image in the last row. In Figure 2, our models are compared using additional negative captions on the SentiCap negative dataset. ATTEND-GAN achieves more diverse and correlated negative captions compared to other models. For example, for the second image in row 1, it generates “cold snow” which appropriately describes the image. However, other models cannot generate sentiment-bearing captions for the image. For the first image in row 3, ATTEND-GAN generates “a plate of disgusting food is served on a table” which is more descriptive than “a plate of disgusting food is on a table” generated by ATTEND-GAN$_{-A}$. For the second image in row 3, ATTEND-GAN generates “a group of stupid people standing around a lot of luggage”. It properly uses “a lot of luggage” to describe the image. However, ATTEND-GAN$_{-A}$ generates “in a store” which is not compatible with the image. Here, ATTEND-GAN performs better than ATTEND-GAN$_{-SA}$ generating “table” improperly. Both ATTEND-GAN and ATTEND-GAN$_{-A}$ generate similar captions for some images including the two images in the last row.

References

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016.

[Chen *et al.*, 2017] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.



AS: a street sign on the side of a building.

A: a nice street sign is on a brick building.

AG: a nice street sign is on a brick wall.

AS: a bowl of food with vegetables and vegetables.

A: a bowl of healthy food is on a table.

AG: a bowl of delicious food is ready to be eaten.

AS: a couple of people on surfboards in the water.

A: a group of people riding on top of a surfboard.

AG: two people are surfing in the calm water.

AS: a man in a kitchen playing a video game.

A: a nice man is playing a video game.

AG: a happy man is playing a video game.

AS: a baseball player is swinging at a ball.

A: a baseball player is swinging his bat at a ball.

AG: a good man is swinging a baseball bat at a game.

AS: a cat laying on top of a laptop computer.

A: a cat is laying on a laptop computer.

AG: a cuddly cat laying on top of a laptop computer.

AS: a woman is playing with a remote control.

A: a happy child is playing with a remote control.

AG: a happy child is playing with a remote.

AS: two giraffes are standing next to each other.

A: two giraffes are standing next to each other in a field.

AG: two giraffes are standing in a sunny field.

AS: a cat laying on top of a red suitcase.

A: a cute cat is sitting on a red chair.

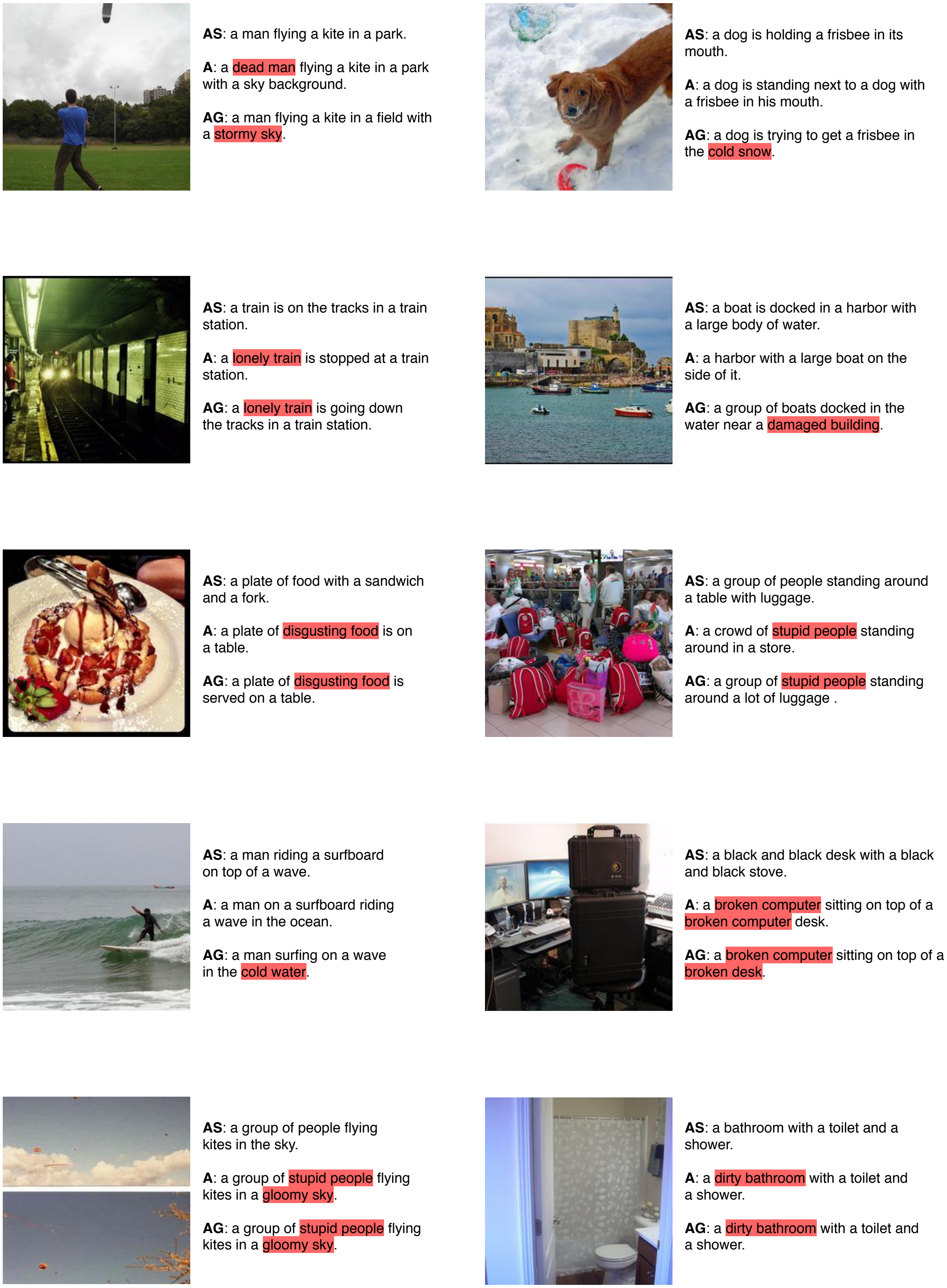
AG: a cute cat is sitting on top of a red suitcase.

AS: a table with plates of food on it.

A: a table with a plate of tasty food on it.

AG: a table with a plate of tasty food on it.

Figure 1: Additional generated captions for the SentiCap positive section (AS for ATTEND-GAN$_{-SA}$, A for ATTEND-GAN$_{-A}$ and AG for ATTEND-GAN).



AS: a man flying a kite in a park.

A: a dead man flying a kite in a park with a sky background.

AG: a man flying a kite in a field with a stormy sky.

AS: a dog is holding a frisbee in its mouth.

A: a dog is standing next to a dog with a frisbee in his mouth.

AG: a dog is trying to get a frisbee in the cold snow.

AS: a train is on the tracks in a train station.

A: a lonely train is stopped at a train station.

AG: a lonely train is going down the tracks in a train station.

AS: a boat is docked in a harbor with a large body of water.

A: a harbor with a large boat on the side of it.

AG: a group of boats docked in the water near a damaged building.

AS: a plate of food with a sandwich and a fork.

A: a plate of disgusting food is on a table.

AG: a plate of disgusting food is served on a table.

AS: a group of people standing around a table with luggage.

A: a crowd of stupid people standing around in a store.

AG: a group of stupid people standing around a lot of luggage .

AS: a man riding a surfboard on top of a wave.

A: a man on a surfboard riding a wave in the ocean.

AG: a man surfing on a wave in the cold water.

AS: a black and black desk with a black and black stove.

A: a broken computer sitting on top of a broken computer desk.

AG: a broken computer sitting on top of a broken desk.

AS: a group of people flying kites in the sky.

A: a group of stupid people flying kites in a gloomy sky.

AG: a group of stupid people flying kites in a gloomy sky.

AS: a bathroom with a toilet and a shower.

A: a dirty bathroom with a toilet and a shower.

AG: a dirty bathroom with a toilet and a shower.

Figure 2: Additional generated captions for the SentiCap negative section (AS for ATTEND-GAN$_{-SA}$, A for ATTEND-GAN$_{-A}$ and AG for ATTEND-GAN).

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.