A Appendices

A.1 Implementation Details.

Following [Karpathy and Fei-Fei, 2015], we convert all the captions to lowercases and remove its non-alphabet characters. We also discard the tokens with frequency less than 5 in the training dataset, resulting in a vocabulary size of 8,791. Both image encoders F_{G_θ} and F_{D_ϕ} in the generator and discriminator are implemented using *ResNet* [He *et al.*, 2016] with 152 layers, separately (we reimplementG-GANby usingResNet-152net-work as image encoders). The image activations in the *pool5* layer are extracted, yielding 2048-dimensional image features. Noise vector z with 100-dimensions is sampled from a uniform distribution. All the image features are projected to 512 dimensions by fully connected layers. The text-decoder in the generator and the text-encoder in the discriminator are all implemented using LSTMs with 512 hidden nodes. We use the last hidden activations from the text-encoder as text feature, which shares the same dimension with the projected image feature.

Before adversarial training, the caption generator G_θ is pretrained by the standard MLE method [Vinyals *et al.*, 2015] [Karpathy and Fei-Fei, 2015] for 20 epochs, and the cr-discriminator is pretrained according to Equation 4 for 10 epochs. During the experiment, we found the generator pretraining is necessary, otherwise it will encounter mode collapse problem and generate nonsense captions. On the other hand, pretraining discriminator helps more stable training later. In the adversarial learning stage, two sub-networks are trained jointly, in which every one generator iteration is followed by 5 discriminator iterations. We set the learning rate to 0.0005 and the batch size to 64. The rollout number K is empirically set to 16, and γ is set to 10. During testing, the generated captions are sampled based on policy and the one with the best cr-score is chosen for evaluation.

A.2 Diversity visualization

As distinctive image contents are described by specific words, caption diversity across images could be visualized by diverse word usages. For this purpose, we can inspect the word usage frequency at each position t of the generated captions:

$$p(w_t) = \mathbb{E}_{w_{t-1}} p(w_t|w_{t-1})p(w_{t-1}), \;\; t \in (1, T)$$

where w_0 only takes one word which is the "START" token and thus $p(w_0 = \text{"START"}) = 1$, and the image notation I is ignored for simplification. The above probability can be estimated by the Markov Chain method over the generated vocabulary distribution (Equation 2). However, as the word space is too large, it is practically difficult to calculate the frequency distribution for all the words. Instead, we use a sampling method to approximate the frequency of the observed words in the generated captions:

$$p(w_t) \simeq count(w_t)/count(\bar{w}_t) \;\; t \in (1, T)$$

where $count(\bar{w}_t)$ denotes the total count of all the observed words at position t. Ideally, diverse word usage implies that each word in the vocabulary is used less repeatedly in caption generation across different images, leading to lower $p(w_t)$ for each word.

To visualize the word usage frequency, we sampled 300 images from the test set and visualize the statistics in Figure 7. Here we chose 300 but not more images for the sake of visualization clarity. As can be seen, the MLE-generated captions usually pick up fewer content words such as "sitting", "riding", and "standing", regardless of different and distinctive image contents. Meanwhile, its corresponding * regions are much narrower than those in other models, meaning that it rarely uses other words in the vocabulary. In contrast, although our CAL model also uses the same function words (e.g. "a", "the", "of", "is", etc.), most of the used content words are less identical and contribute to much wider * regions. We also find that our CAL uses more adjectives and adverbs in generated captions.

By comparison, we can conclude that the word frequency distribution of our CAL is more akin to that of Human. This demonstrates that our CAL model has more diverse word usages than the baseline G-GAN, resulting in more distinctive captions across images.

A.3 Failure analysis

For some images with complex content, we find the CAL-generated captions are imprecise or defective in describing their contents. One possible reason is that if a complicate image does not have a focused topic, its ground truths are normally divergent for different aspects of the image. As a result, during comparative adversarial learning, the caption generator can not simultaneously capture all the opposed details from ground truths. Additionally, to generate more descriptive and diverse captions for images, our framework bears risks that involving some incorrect details. Figure 8 shows some failure examples from our method. We will consider these problems in our future study.

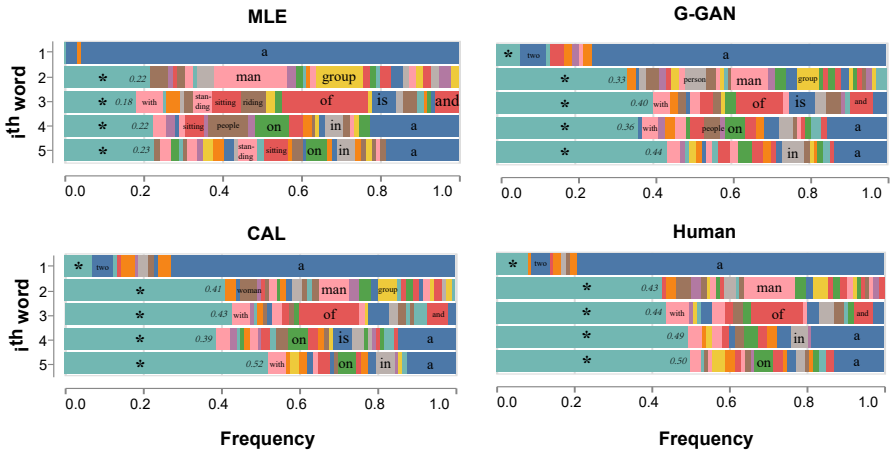

Figure 7: Visualization of word diversity produced by different models. For each subgraph, the i^{th} row represents the i^{th} word's frequency distribution in all the generated captions. Different colors denote different words, and the width of each region is proportional to the frequency of the corresponding word. We only plot the first five words for easy readability. We mark the words of high frequency in the figure. The words of low frequency ($< 0.5\%$) are merged into *the others* (denoted as *) category. Larger proportion of * means more chances of using diverse words (or long tail words) in the vocabulary. The decimal in each * region denotes its proportion value for easy comparison.

a police officer directing traffic at a bus stop

a display store of various old speckled benches

Figure 8: Failure examples from the proposed network.