R1: Cut down on some sections (3.2.1, 3.2.2 and 3.2.5) to spare space for the qualitative examples.

We will revise our paper according to the suggestion in the final version.

Table 1: Experiments on MS-COCO and Flicker30k datasets using single-head attention. (Row Steps shows the min./max./avg. attention time steps of each model.)

<table>
<thead>
<tr>
<th>Model</th>
<th>MS-COCO</th>
<th>Flicker30k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>C</td>
</tr>
<tr>
<td>Base</td>
<td>1/1/1</td>
<td>27.8</td>
</tr>
<tr>
<td>Recurrent</td>
<td>2/2/2</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>4/4/4</td>
<td>27.8</td>
</tr>
<tr>
<td>Adaptive</td>
<td>0/4/2.4</td>
<td>28.0</td>
</tr>
<tr>
<td></td>
<td>1/4/2.8</td>
<td>27.9</td>
</tr>
<tr>
<td></td>
<td>2/4/3.2</td>
<td>27.8</td>
</tr>
</tbody>
</table>

R2: Apply AAT on traditional single-head instead of multi-head attention to show that AAT helps.

We added experiments on MS-COCO and Flicker30k using single-head attention, Table 1. As can be seen, adaptive attention model with (0/4/2.4) yields best results, which show that AAT also helps single-head attention.

R2: The base attention model performs better than up-down and GCN-LSTM.

The reason lies in that the base attention model adopts a different structure (LSTM in Section 3.1) and different experimental settings (batch size, learning rate and schedule sampling rate in Section 4.1).

R2: Provide more analysis to find the reason for improvement from recurrent attention model to adaptive attention model.

The reason for adaptive attention model (AAT) improves from recurrent is that AAT helps to decide how many attention steps (from zero to multiple, adaptively) to take before outputting a word, while the number of attention steps is fixed for recurrent. Fixing attention steps introduces redundant or even misleading information since not all words require visual clues [14]. In addition, our experimental results showed that increasing the number of min. attention steps for adaptive attention model (1/4/2.8 and 2/4/3.2) degrades the performances, in Table 1.

R2: How much does the attention change over multiple attention steps for each word position?

It changed very much as shown in Fig. 1 in the appendix. For each word, the attention changes: a) towards more accurate objects than previous steps; b) for objects which have connections with each other to obtain a better overview.

R2: How does the attention time steps vary with word position?

The numbers of attention time steps at the beginning of the sentence or phrases (e.g. “on the side” and “at a ball”) are larger than those at other positions.

R2: Does this number change significantly after self-critical training?

It doesn’t change significantly after self-critical training but requires relatively less attention steps.

R2: Is it the case that self-critical training is necessary to fully utilize the potential of AAT?

Yes. We experimentally found that self-critical training significantly boosted the performance (Table 1 in the paper).

R2: Why words at early decoding steps have little access to image information?

Because the decoder incorporates little information about the image at early steps.

R2: Are the ablations in Table 1 done on the same split as Table 2 (in the main paper)?

Yes, all the experiments in this paper are done on the ‘Karpathy’ splits.

R3: Add flicker results and report STD.

We experimented on Flicker-30k and reported results as well as STD in Table 1. STD on COCO dataset will be added to the final version.

R3: N(t) in eq. 14 is non-differentiable.

N(t) doesn’t contribute for the gradients, and it solely indicates the number of attention steps.