In this supplementary material, we first present the spearman correlation to show the legitimacy of our attention correctness metric, then provide more visualization in the generated caption experiment.

The Spearman Correlation

Table 1: The spearman correlation of potential correctness metrics on the Flickr30k test set with implicit attention and ground truth caption. Attention correctness correlates well with the other metrics.

<table>
<thead>
<tr>
<th></th>
<th>AC</th>
<th>L1</th>
<th>L2</th>
<th>KL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>1.0000</td>
<td>0.9890</td>
<td>0.9624</td>
<td>0.9838</td>
</tr>
<tr>
<td>L1</td>
<td>1.0000</td>
<td>0.9822</td>
<td>0.9958</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>1.0000</td>
<td>0.9827</td>
<td>0.9827</td>
<td></td>
</tr>
<tr>
<td>KL</td>
<td></td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Attention Map Visualization

We visualize the attention maps of both the implicit attention model and our supervised attention model on the Flickr30k test set. As mentioned in the paper, 909 noun phrases are aligned for the implicit model and 901 for the supervised model. 635 of these alignments are common for both, and 595 of them have corresponding bounding boxes. Here we present a subset due to space.

For every figure, the original image is on the left, the implicit attention result is in the middle, and the supervised attention result is on the right. The red box marks the ground truth attention region, as annotated in the Flickr30k Entities dataset. The attention correctness score for this phrase is in the parenthesis.

Figure 1: people (0.242). people (0.477).

Figure 2: people (0.257). people (0.457).

Figure 3: people (0.637). people (0.822).

Figure 4: jeans (0.386). jeans (0.487).

Figure 5: jeans (0.320). jeans (0.320).

Figure 6: people (0.279). people (0.453).

Figure 7: people (0.564). people (0.602).

Copyright © 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
Figure 56: A woman (0.905). A woman (0.861).

Figure 57: A man (0.605). A man (0.635).

Figure 58: A man (0.977). A man (0.968).

Figure 59: a bulldozer (0.756). a bulldozer (0.874).

Figure 60: A man (0.261). A man (0.635).

Figure 61: the water (0.933). the water (0.964).

Figure 62: A man (0.460). A man (0.575).

Figure 63: Two dogs (0.224). Two dogs (0.438).

Figure 64: a baby (0.400). a baby (0.562).

Figure 65: a woman (0.461). a woman (0.555).

Figure 66: A man (0.311). A man (0.570).

Figure 67: A man (0.031). A man (0.054).

Figure 68: A man (0.068). A person (0.199).

Figure 69: a bench (0.646). a bench (0.748).

Figure 70: A woman (0.950). a woman (0.972).

Figure 71: the water (0.995). the water (0.997).
Figure 88: *A little girl* (0.847). *A little girl* (0.868).

Figure 89: *A young boy* (0.231). *A young boy* (0.434).

Figure 90: *a blue shirt* (0.411). *a blue shirt* (0.450).

Figure 91: *a blue shirt* (0.579). *a blue shirt* (0.706).

Figure 92: *A brown dog* (0.985). *A brown dog* (0.955).

Figure 93: *a black shirt* (0.308). *a black shirt* (0.401).

Figure 94: *A little girl* (0.932). *A little girl* (0.925).

Figure 95: *A brown dog* (0.338). *A dog* (0.403).
Figure 96: *a blue shirt* (0.559). *a blue shirt* (0.661).

Figure 97: *A black dog* (0.708). *A black dog* (0.760).

Figure 98: *a red bench* (0.500). *a red bench* (0.601).

Figure 99: *a white shirt* (0.352). *a white shirt* (0.496).

Figure 100: *A black dog* (0.613). *A black dog* (0.566).

Figure 101: *a pink dress* (0.416). *a pink dress* (0.638).

Figure 102: *a blue shirt* (0.314). *a blue shirt* (0.471).

Figure 103: *a white shirt* (0.842). *a white shirt* (0.874).
Figure 104: A group of people (0.799). A group of people (0.766).

Figure 105: A group of people (0.735). A group of people (0.952).

Figure 106: A group of children (0.924). A group of children (0.922).

Figure 107: A group of people (0.842). A group of people (0.882).

Figure 108: A group of people (0.723). A crowd (0.828).

Figure 109: A group of people (0.717). A group of people (0.840).

Figure 110: A group of people (0.356). A group of people (0.396).

Figure 111: A group of people (0.756). A group of people (0.760).
Figure 112: A group of people (0.259). A group of people (0.444).

Figure 113: A group of people (0.634). A group of people (0.792).

Figure 114: A group of people (0.568). A group of people (0.634).

Figure 115: A group of people (0.665). A group of people (0.842).

Figure 116: A group of people (0.826). A group of people (0.840).

Figure 117: A group of people (0.869). A group of people (0.905).

Figure 118: A group of people (0.207). A group of people (0.265).

Figure 119: A group of people (0.323). A group of people (0.442).