Relation-Aware Graph Attention Network for Visual Question Answering: Supplementary Material

Linjie Li, Zhe Gan, Yu Cheng, Jingjing Liu
Microsoft Dynamics 365 AI Research
{lindsey.li, zhe.gan, yu.cheng, jingjl}@microsoft.com

1. Definition of Explicit Relations

In this section, we explain in detail how the edge labels (*i.e.*, relation classes) are obtained for the spatial graph and the semantic graph, respectively.

Spatial Relation Spatial relation is defined as the relative geometric position of object j against object i, denoted as $spa_{i,j} = <i\text{-}p\text{-}j>$. The predicate p is the class label of the spatial relation, out of 11 pre-defined categories. We built a rule-based classifier depending on the distance and angle between the center points of object i and object j. Denote the center point of the two regions as $c_i = (c_x^i, c_y^i)$ and $c_j = (c_x^j, c_y^j)$, respectively. $diag$ is the length of the image diagonal. $IoU(i,j)$ is the intersection over union between region i and j. $dist(c_i, c_j)$ denotes the distance between two points c_i and c_j. If $IoU(i,j) < 0.5$ and $dist(c_i, c_j) > 0.5 \cdot diag$, we assume that the interactions between the two regions are too weak to form a relation. Otherwise, we classify $spa_{i,j}$ into 11 different categories through a rule-based classifier:

- If i is inside of j, class 1;
- If j is inside of i, class 2;
- If $IoU(i,j) \geq 0.5$, class 3;
- If $IoU(i,j) < 0.5$ and $dist(c_i, c_j) \leq 0.5 \cdot diag$, class $\lceil \arctan(\frac{c_y^j - c_y^i}{c_x^j - c_x^i})/(\pi/4) \rceil + 3$.

Semantic Relation Given two object regions i and j, we train a classification model to determine which predicate p forms a semantic relation $<i\text{-}p\text{-}j>$ between these two regions. The classification model takes in three inputs: feature vector of the subject region $\boldsymbol{v}_i$, feature vector of the object region $\boldsymbol{v}_j$, and region-level feature vector $\boldsymbol{v}_{i,j}$ of the union bounding box containing both i and j. The output softmax probability is over 15 semantic relation classes, including a `no-relation` class. The predicted semantic relation predicate p is obtained through the following equation:

$$p = \begin{cases} \texttt{no-relation}, & \text{if } Pr(\texttt{no-relation}) \geq 0.5\,; \\ \arg\max_{x \in P} Pr(x), & \text{else}\,; \end{cases}$$

where P denotes the set of 15 semantic relation predicates.

Examples of semantic and spatial relations are shown in Table 1.

Type	#Classes	Examples
Spatial	11	above, around, below, inside, cover, etc.
Semantic	15	wearing, holding, sitting on, standing on, riding, eating, hanging from, carrying, etc.

Table 1. Examples of spatial and semantic relations.

2. Visualization of Ablation Experiments

We show more examples of the attention maps learned by different ablated instances in Figure 1. These examples further support the contribution of each individual design of our model. Specifically, the visualization results demonstrate that both the question-adaptive mechanism and the graph-attention mechanism benefit our model performance by forcing the model to focus on more relevant regions and sharpening the learned attention map, respectively. Note that the fourth column is our complete model, where both the question-adaptive mechanism and the graph-attention mechanism are employed.

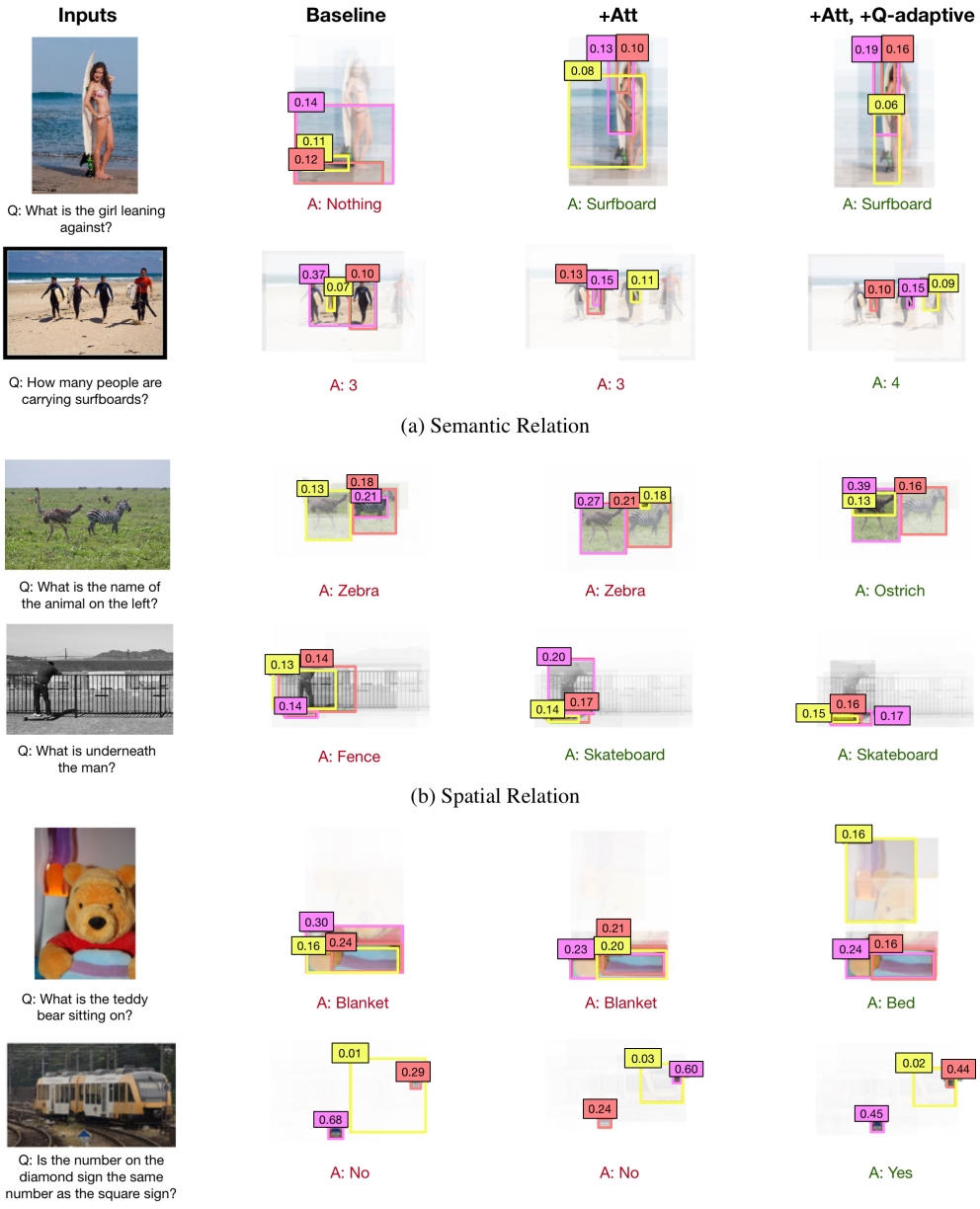
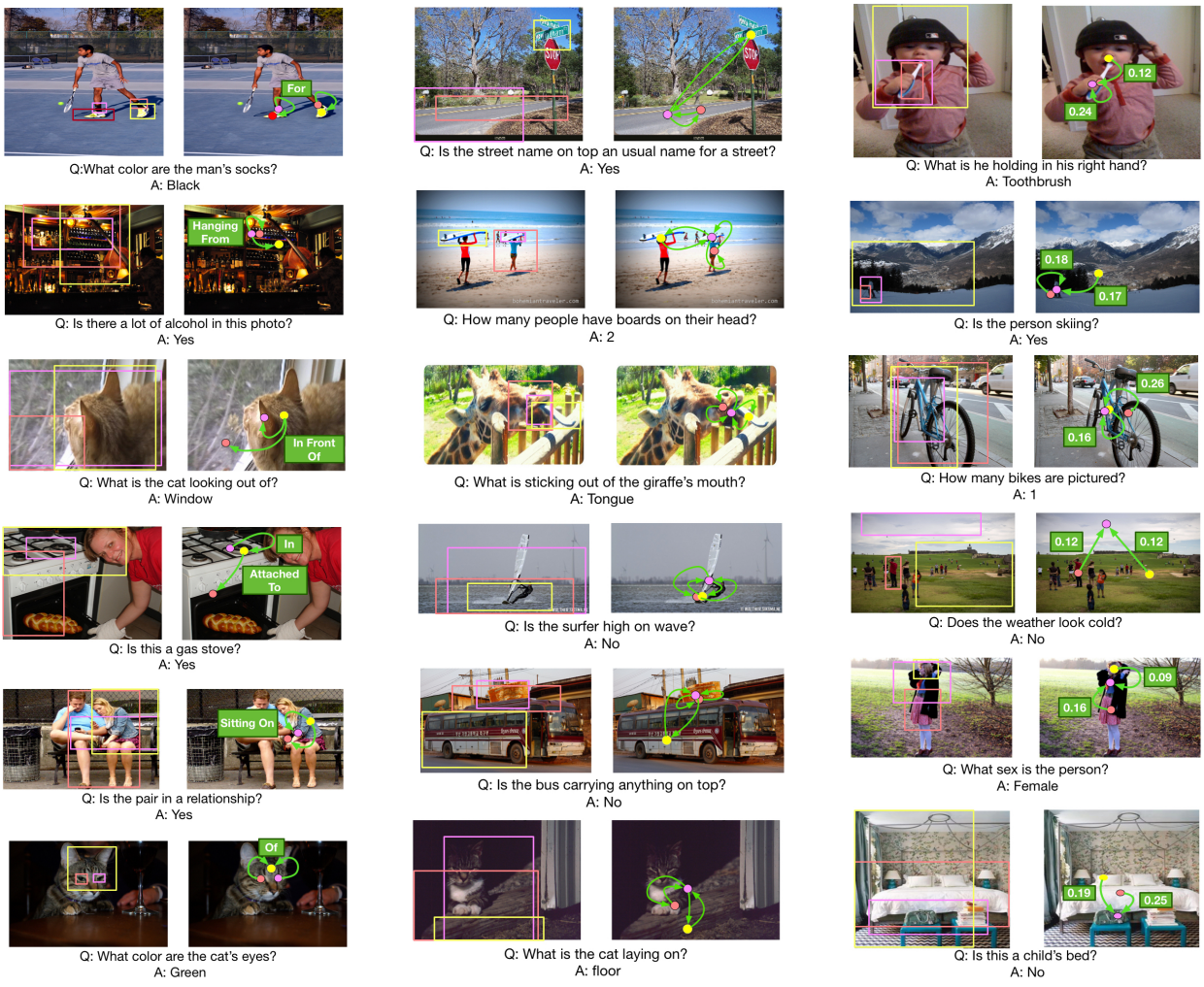


Figure 1. Comparison of attention maps learned by 3 ablated instances of our model with different relations: (a) Semantic Relation, (b) Spatial Relation and (c) Implicit Relation. The top-3 attended regions with their corresponding attention weights are shown in the second, third and fourth column for the baseline model (BUTD), baseline with the graph-attention mechanism (+Att) and our complete model (+Att, +Q-adaptive), respectively. Correct answers are colored with *green* and wrong answers are colored with *red*.

3. Visualization of Relations

In addition to the examples of relations we have shown in Section 4.5 of the main text, we visualize each type of relations with more examples from VQA 2.0 dataset in Figure 2. For explicit relations, we show the top-3 attended regions and the learned relations between these regions. Spatial relations are indicated with bi-directional arrows between regions and each semantic relation is indicated with a single directional arrow and its corresponding relation label. For implicit relations, we show the top-1 attended region i and two corresponding regions that contribute to the top-2 attention weights to region i. The attention weights from those two regions to region i are annotated in the green boxes. From the examples, we can see that each type of relations contributes to the better alignment between image regions and questions.



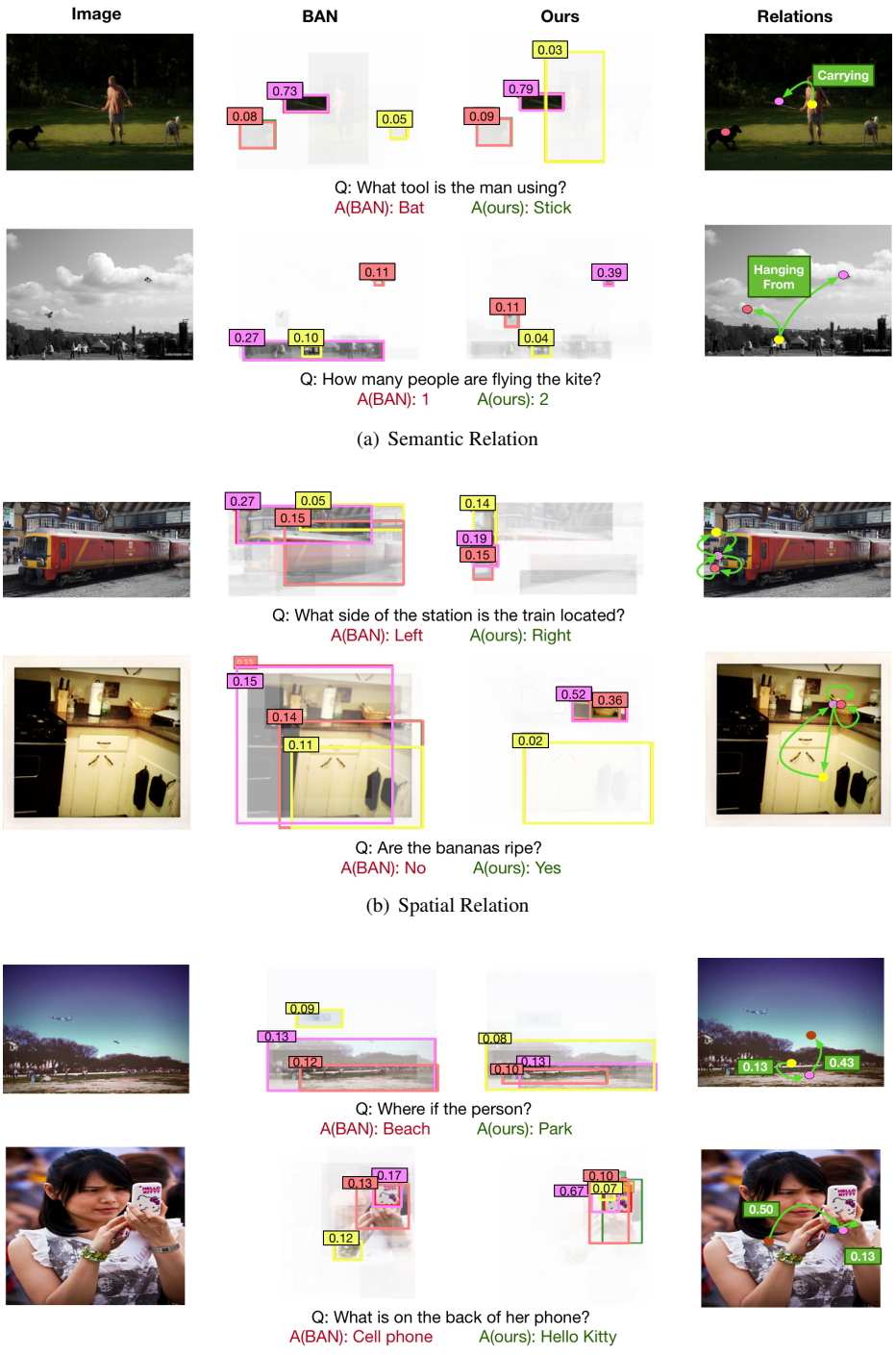
(a) Semantic Relation (b) Spatial Relation (c) Implicit Relation

Figure 2. Visualization of different types of visual object relations: (a) Semantic Relation, (b) Spatial Relation and (c) Implicit Relation. The three bounding boxes shown in each image are the top-3 attended regions. The green arrows indicate relations from object to subject. Labels and numbers in green boxes are class labels for semantic relations and attention weights for implicit relations.

4. Visualization of Attention (BAN vs. ReGAT)

In this section, we compare the attention maps learned by ReGAT with those learned by the BAN model in Figure 3. Similar to the visualization of relations, we show the top-3 attended regions for BAN in the second column and ReGAT in the third column, respectively. Figure 3 also provides visualization on different types of visual object relations for each example in the fourth column. Relations are visualized in the same way as in Figure 2. The comparisons shown in these examples

further demonstrate that each type of relations contributes to the better alignment between image regions and questions. Specifically, our model results in sharper attention maps and focuses on more relevant regions compared to the BAN model.



(a) Semantic Relation

(b) Spatial Relation

(c) Implicit Relation

Figure 3. Comparison of attention maps learned by BAN and our model: (a) Semantic Relation, (b) Spatial Relation and (c) Implicit Relation. The top-3 attended regions with their corresponding attention weights are shown in the second and third column for BAN and our model, respectively. Relations are also visualized for each example in the fourth column. Correct answers are colored with *green* and wrong answers are colored with *red*.