1. Visual Question Answering Model

Section 1.1 extends section 4 in the main paper, detailing the VQA answering model of PJ-X by providing the detailed formulas we omitted for brevity. The VQA model that we use throughout the experiments is based on the state-of-the-art MCB model [2], but trains and evaluates faster (reduction of \( \sim 30\% \)). In this section, we outline the model in detail.

### 1.1. Model Details

For spatial image features \( f^I(I, n, m) \) from the last convolutional layer of ResNet-152, question \( Q \), 2-layer LSTM \( f^Q(Q) \) we compute:

\[
\bar{f}^I(I, n, m, Q) = (W_1 f^I(I, n, m) + b_1) \odot f^Q(Q)
\]

\[
f^I(I, Q) = L2(signed\_sqrt(f^I(I, Q)))
\]

\[
\bar{\alpha}_{n,m}^{pointA} = f^{pointA}(I, n, m, Q)
\]

\[
= W_3 \rho(W_2 \bar{f}^I(I, Q) + b_2) + b_3
\]

with ReLU \( \rho(x) = max(x, 0) \). This process gives us a \( N \times M \) attention map \( \bar{\alpha}_{n,m} \). We apply softmax to produce a normalized soft attention map for predicting the answer:

\[
\alpha_{n,m}^{pointA} = \frac{\exp(\bar{\alpha}_{n,m}^{pointA})}{\sum_{i=1}^{N} \sum_{j=1}^{M} \exp(\bar{\alpha}_{i,j}^{pointA})}
\]

\[
\tilde{f}^y(I, Q) = (\sum_{x=1}^{N} \sum_{y=1}^{M} \alpha_{n,m}^{pointA} f^I(I, n, m)) \odot f^Q(Q)
\]

\[
f^y(I, Q) = W_4 \tilde{f}^y(I, Q) + b_4
\]

\[
p(y|I, Q) = Softmax(f^y(I, Q))
\]

\[
\hat{y} = \arg\max_{y \in Y} p(y|I, Q)
\]

### 1.2. Results

As shown in Table 1, our VQA model leads to a considerable improvement in performance, especially in VQA v2. Our model shows an overall 0.51% improvement on VQA v1 and 2.57% on VQA v2, where the most significant gain is achieved in the “Number” category. In addition, our model is more efficient than MCB [2] in that the training and inference are \( \sim 30\% \) faster.
Table 1: OpenEnded results on VQA v1 [1] and VQA v2 [3] datasets. The models are trained on train set, validated on validation set, and reported on the test-dev accuracies. The columns indicate the accuracies of the model for each different question type. Our model achieves higher accuracy than MCB [2] while being faster at train and test time.

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA v1</th>
<th></th>
<th>VQA v2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>Our VQA model</td>
<td>63.01</td>
<td>82.01</td>
<td>35.47</td>
<td>52.99</td>
</tr>
</tbody>
</table>

References

