AiR: Attention with Reasoning Capability (Supplementary Materials)

Shi Chen*[0000−0002−3749−4767], Ming Jiang*[0000−0001−6439−5476], Jinhui Yang[0000−0001−8322−1121], and Qi Zhao[0000−0003−3054−8934]

University of Minnesota, Minneapolis MN 55455, USA
{chen4595,mjiang,yang7004,qzhao}@umn.edu

The supplementary materials consist of results and details of the proposed Attention with Reasoning capability (AiR) framework:

1. We complement the results presented in the main paper with analyses on different baselines, including Multi-modal Factorized Bilinear (MFB) [10] and Multi-Modal Tucker Fusion (MUTAN) [2] (Section 1.1).
2. We present ablation studies on different attention supervision strategies and hyperparameters of the proposed AiR-M method (Section 1.2 and Section 1.3).
3. We present additional qualitative comparisons between the AiR-M method and various attention supervision methods (Section 1.4).
4. We present details about the decomposition of the reasoning process (Section 2.1).
5. We report details of the proposed AiR-M method with applications to different types of attention mechanisms (Section 2.2).

We also provide a **supplementary video** to illustrate the spatiotemporal dynamics of both model and human attentions throughout the reasoning process. It demonstrates the effectiveness of the proposed attention supervision method (AiR-M) in improving both the attention accuracy and task performance. In addition, the video also highlights the significant spatiotemporal discrepancy between human attentions with correct and incorrect answers.

1 Supplementary Results

1.1 Analyses on MFB and MUTAN Models

Due to the page limit of the main paper, here we provide supplementary analyses on two additional baselines, *i.e.,* MFB [10] and MUTAN [2], to demonstrate the generality of the analyses for different baselines. The experimental procedures are consistent with those in Section 4.1 of main paper.

Overall, the results on MFB [10] and MUTAN [2] agree with our observations reported in the main paper (on UpDown [1]): (1) Compared with computational models, humans have stronger reliance on attention and attend more accurately

* Equal contributions.

Table 1: Quantitative evaluation of AiR-E scores and task performances of the MFB [10] baseline. Bold numbers indicate the best attention performance.

	Attention	and	compare	filter	or	query	relate	select	verify
AiR-E	H-Tot	2.197	2.669	2.810	2.429	3.951	3.516	2.913	3.629
	H-Cor	2.258	2.717	2.925	2.529	4.169	3.581	2.954	3.580
	H-Inc	1.542	1.856	1.763	1.363	2.032	2.380	1.980	2.512
	MFB-O-Soft	**1.841**	1.055	**1.294**	**2.295**	**3.799**	**1.779**	1.372	**2.563**
	MFB-O-Trans	1.787	**1.446**	1.054	1.957	3.740	1.730	**1.405**	2.386
	MFB-S-Soft	0.217	-0.044	0.176	0.477	0.746	0.341	0.195	0.005
	MFB-S-Trans	0.438	0.367	0.524	0.702	0.765	0.640	0.468	0.652
Accuracy	H-Tot	0.700	0.625	0.668	0.732	0.633	0.672	0.670	0.707
	MFB-O-Soft	0.626	0.593	0.549	**0.834**	**0.389**	**0.467**	**0.533**	0.645
	MFB-O-Trans	**0.631**	**0.598**	**0.550**	0.833	0.388	0.466	**0.533**	**0.649**
	MFB-S-Soft	0.595	0.581	0.508	0.805	0.316	0.408	0.481	0.614
	MFB-S-Trans	0.598	0.581	0.506	0.802	0.313	0.406	0.479	0.615

Table 2: Pearson's correlation coefficients between attention accuracy (AiR-E) and task performances of the MFB [10] baseline. Bold numbers indicate significant positive correlations ($p<0.05$).

Attention	and	compare	filter	or	query	relate	select	verify
H-Tot	0.205	**0.329**	0.051	0.176	**0.282**	**0.210**	**0.134**	**0.270**
MFB-O-Soft	0.103	-0.096	**0.131**	**0.244**	**0.370**	0.045	0.041	**0.182**
MFB-O-Trans	**0.225**	0.050	**0.121**	**0.197**	**0.370**	0.038	0.043	**0.256**
MFB-S-Soft	0.027	0.064	-0.084	0.015	**0.219**	-0.013	-0.028	0.084
MFB-S-Trans	-0.013	0.077	-0.033	**0.238**	**0.165**	0.037	0.001	-0.019

Table 3: Quantitative evaluation of AiR-E scores and task performances of the MUTAN [2] baseline. Bold numbers indicate the best attention performance.

	Attention	and	compare	filter	or	query	relate	select	verify
AiR-E	H-Tot	2.197	2.669	2.810	2.429	3.951	3.516	2.913	3.629
	H-Cor	2.258	2.717	2.925	2.529	4.169	3.581	2.954	3.580
	H-Inc	1.542	1.856	1.763	1.363	2.032	2.380	1.980	2.512
	MUTAN-O-Soft	**2.051**	**1.490**	**1.676**	**2.644**	**3.683**	**2.096**	**1.695**	**2.762**
	MUTAN-O-Trans	0.973	0.851	1.137	1.655	2.559	1.609	1.130	1.974
	MUTAN-S-Soft	0.124	0.098	0.253	0.347	0.761	0.359	0.243	0.172
	MUTAN-S-Trans	0.290	0.074	0.208	0.182	0.778	0.399	0.253	0.155
Accuracy	H-Tot	0.700	0.625	0.668	0.732	0.633	0.672	0.670	0.707
	MUTAN-O-Soft	**0.602**	**0.593**	**0.541**	**0.787**	**0.385**	**0.457**	**0.521**	**0.645**
	MUTAN-O-Trans	0.582	0.588	0.529	0.771	0.374	0.443	0.507	0.635
	MUTAN-S-Soft	0.568	0.583	0.502	0.765	0.320	0.404	0.473	0.616
	MUTAN-S-Trans	0.576	0.583	0.501	0.763	0.319	0.402	0.473	0.615

Table 4: Pearson's correlation coefficients between attention accuracy (AiR-E) and task performance of the MUTAN [2] baseline. Bold numbers indicate significant positive correlations ($p<0.05$).

Attention	and	compare	filter	or	query	relate	select	verify
H-Tot	0.205	**0.329**	0.051	0.176	**0.282**	**0.210**	**0.134**	**0.270**
MUTAN-O-Soft	0.142	-0.027	**0.130**	**0.205**	**0.369**	0.034	0.018	**0.229**
MUTAN-O-Trans	**0.174**	0.027	0.015	0.128	**0.284**	**0.076**	0.002	0.122
MUTAN-S-Soft	**0.228**	0.062	-0.160	0.059	**0.121**	-0.090	-0.058	0.116
MUTAN-S-Trans	-0.021	0.065	-0.208	-0.034	**0.105**	-0.064	-0.118	0.074

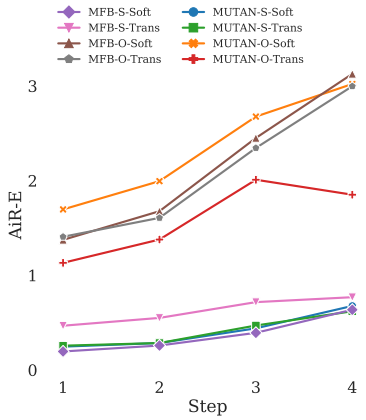

Fig. 1: Spatiotemporal accuracy of attention throughout the reasoning process.

during visual reasoning (see Table 1 and Table 3), which agrees with Table 2 of the main paper; (2) The task performance is jointly influenced by both attention accuracy and reasoning operations (see Table 2 and Table 4), which agrees with Table 3 of the main paper; (3) Existing machine attentions tend to focus more on the ROIs closely related to the task outcome but are less related to intermediate reasoning steps (see Fig. 1), which agrees with Fig. 5a of the main paper. These agreements confirm that our observations in the main paper are general and consistent across different baselines.

1.2 Ablation Study of Attention Supervision Strategies

To evaluate the effectiveness of different components in the proposed AiR-M attention supervision method, we conduct an ablation study on the UpDown [1] baseline with different supervision strategies:

1. Joint supervision of attention, reasoning operation, and task performance, but in a single-glimpse manner, denoted as $L_{\boldsymbol{\alpha}} + L_{\boldsymbol{r}}$. Specifically, the model only predicts a single attention map by incorporating the last hidden state of the GRU for operation prediction, and attention supervision is accomplished with ground truth attention aggregated across all steps.

2. Progressive supervision only on the reasoning operation and task performance, denoted as AiR-M (w/o $L_{\boldsymbol{\alpha}}$).
3. Progressive supervision only on attention and task performance, denoted as AiR-M (w/o $L_{\boldsymbol{r}}$).

Table 5: Experimental results of AiR-M under different supervision strategies. All reported results are on the GQA [6] test-dev set. Bold numbers indicate the best performance.

Method	Performance on GQA test-dev
w/o supervision	51.31
PAAN [8]	48.03
HAN [9]	49.96
ASM [11]	52.96
$L_{\boldsymbol{\alpha}} + L_{\boldsymbol{r}}$	52.84
AiR-M (w/o $L_{\boldsymbol{\alpha}}$)	50.01
AiR-M (w/o $L_{\boldsymbol{r}}$)	50.33
AiR-M	**53.46**

As shown in Table 5, with the single-glimpse supervision of both the attention and reasoning operations (*i.e.,* $L_{\boldsymbol{\alpha}} + L_{\boldsymbol{r}}$), the task performance increases marginally from 51.31 to 52.96, which is comparable with the ASM methods [11]. With independent progressive supervision on either the reasoning operation (AiR-M (w/o $L_{\boldsymbol{\alpha}}$)) or the attention (AiR-M (w/o $L_{\boldsymbol{r}}$)), however, the task performances decrease by about 1%. In comparison, with the joint progressive supervision of attention and reasoning operations (*i.e.,* AiR-M), the task performance is improved by more than 2%. The performance change indicates that it is essential to jointly supervise attention and reasoning, so that the model can develop sufficient understanding of the complex interactions between them.

1.3 Ablation Study of Hyperparameters

The objective function of the proposed AiR-M attention supervision method consists of three loss terms:

$$L = L_{ans} + \theta \sum_{t} L_{\boldsymbol{\alpha_t}} + \phi \sum_{t} L_{\boldsymbol{r}_t} \tag{1}$$

where L_{ans}, $L_{\boldsymbol{\alpha_t}}$ and $L_{\boldsymbol{r}_t}$ are the loss terms on answer, attention and reasoning operations, respectively. The three loss terms are linearly combined with two hyperparameters, *i.e.* θ and ϕ. For θ, we follow the dynamic hyperparameter proposed in [11], and define it as $\theta = 0.5(1 + \cos(\pi \cdot Iter/C))$, where $Iter$ is the current iteration and C denotes the maximal number of training iterations ($C = 300$k in our experiments). In terms of ϕ, we conduct experiments on the

balanced validation set of GQA with different ϕ values from 0.01 to 10. Table 6 reports the results using UpDown [1] as the baseline.

Table 6: Experimental results of models trained with different settings of the hyperparameter (*i.e.,* ϕ for objective related to operation prediction). All reported results are on the GQA [6] balanced validation set. Bold numbers indicate the best performance.

ϕ	Performance on GQA Validation
0.01	62.02
0.1	62.11
0.5	**62.57**
1	61.93
10	61.39

According to the results, the proposed method is relatively robust against various settings of ϕ, and we empirically find that $\phi = 0.5$ provides the best validation accuracy. However, setting an much larger weight to the operation objective (*e.g.,* $\phi = 10$) tends to hamper the learning of attention and results in a considerable performance drop.

1.4 Qualitative Results of AiR-M Attention Supervision

To further support our observations in the main paper (see Fig. 6 of the main paper), we present additional qualitative examples of the AiR-M method, in comparison with the UpDown baseline and state-of-the-art attention supervision methods. As shown in Fig. 2, our method can accurately guide machine attention to focus on the ROIs related to the final answers (rows 1-7) as well as intermediate reasoning steps (rows 8-9).

2 Supplementary Method

2.1 Decomposition of Reasoning Process

As introduced in the main paper, we decompose the reasoning process into a sequence of atomic operations with regions of interest (ROIs) to conduct fine-grained evaluation of attentions for each step of the sequence. In this section, we describe this decomposition method in details:

Deriving the Atomic Operations. We derive the atomic operations (see Table 1 of the main paper) by characterizing and abstracting the complex operations in the functional programs of the GQA datset [6]. Specifically, we define each reasoning operation as a triplet, *i.e.*, <operation, attribute, category> and categorize the original operations in the program based on their semantic meanings: (1) For the original operations that exactly align with our definitions, we

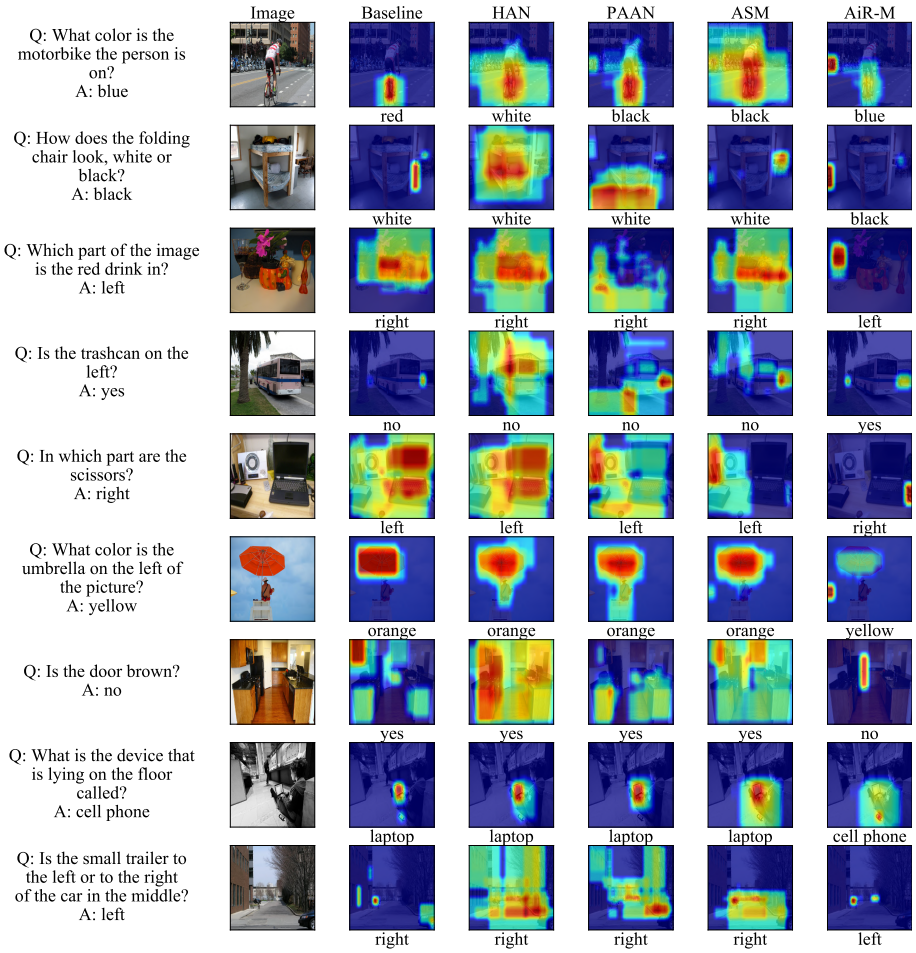

Fig. 2: Qualitative comparison between attention supervision methods, where Baseline refers to UpDown [1]. For each row, from left to right are the questions and the correct answers, input images, and attention maps learned by different methods. The predicted answers associated with the attentions are shown below their respective attention maps.

directly convert them into our triplet representation, for example, from "filter size table" to <filter, large/small, table>; (2) If the original operations do not have an exact match, we convert them into our defined operations with similar semantic meanings. For example, we convert "different color object A and object B" to <compare, color, category A and category B>. Most of the original GQA operations can be effectively converted into such triplet representations without loss of information. The triplets allow us to efficiently traverse the reasoning process by investigating the semantics of operations and their corresponding ROIs.

Determining the ROIs of Each Operation. The selection of ROIs depends on the semantics of the operations:

- **Select**: The ROIs belong to a specific category of objects. We query all objects in the scene graph and select those with the same category as defined in the triplet.
- **Query, Verify**: The ROIs are defined in a similar way as the "select" operation. The difference is that they are selected from the ROIs of the previous step, instead of the entire scene graph.
- **Filter**: The ROIs are a subset of the previous step's ROIs with the same attribute as defined in the triplet.
- **Compare, And, Or**: These operations are based on multiple groups of objects. Therefore, the ROIs are the combination of all the ROIs of the related previous reasoning steps.
- **Relate**: The ROIs are a combination of two groups of objects: the ROIs of the previous reasoning step and a specific category of objects from the scene graph.

Some questions in GQA [6], *e.g.,* "Is there a red bottle on top of the table" with answer "no", refer to non-existing objects. In such cases, we select the k most frequently co-existent objects as the ROIs. Specifically, based on the GQA scene graphs, we first compute the frequency of co-existence between different object categories on the training set. Next, given a particular reasoning operation referring to an non-existing object, the top-k ($k = 20$) co-existing objects in the scene are selected as the corresponding ROIs.

2.2 Attention Supervision Method (AiR-M)

To demonstrate the effectiveness of the proposed AiR-M method for attention supervision, we apply it to three state-of-the-art VQA models, including Up-Down [1] and MUTAN [2] with standard visual attention, and BAN [7] with co-attention between vision and language. In this section, we present the implementation details of applying our method on different baseline models.

Application of AiR-M on UpDown [1] and MUTAN [2]. Many reasoning models, including UpDown and MUTAN, adopt the standard visual attention mechanism that computes attention weights for different units of visual input

(*i.e.* region proposals or spatial locations). To apply AiR-M on UpDown and MUTAN, we substitute their original attention mechanisms with the proposed one that jointly predicts the attention maps and the corresponding reasoning operations. Fig. 3 shows the high-level architecture of a model (*e.g.,* UpDown[1] and MUTAN [2]) with its visual attention replaced with our proposed AiR-M.

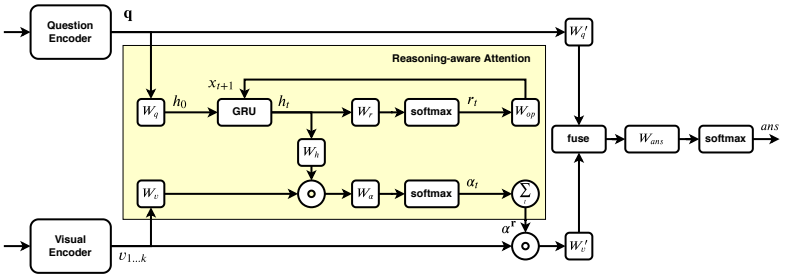

Fig. 3: High-level architecture of the proposed AiR-M method on models with standard visual attention.

Specifically, the AiR-M takes q and V as the inputs, and uses a Gated Recurrent Unit [3] (GRU) to sequentially predict the operations $\boldsymbol{r}_t$ and the desired attention weights $\boldsymbol{\alpha}_t$ at the t-th step. The attentions are aggregated into a final attention vector $\boldsymbol{\alpha}^r$ to dynamically prioritize the visual features.

At the beginning of the reasoning process, the hidden state of GRU $\boldsymbol{h}_0$ with the question features $\boldsymbol{q}$ is defined as:

$$\boldsymbol{h}_0 = \boldsymbol{W}_q \boldsymbol{q}, \tag{2}$$

where $\boldsymbol{W}_q$ represents trainable weights. We update the hidden state $\boldsymbol{h}_t$, and simultaneously predict the reasoning operation $\boldsymbol{r}_t$ and attention $\boldsymbol{\alpha}_t$:

$$\boldsymbol{r}_t = \text{softmax}(\boldsymbol{W}_r \boldsymbol{h}_t), \tag{3}$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\boldsymbol{W}_\alpha(\boldsymbol{W}_v \boldsymbol{v} \circ \boldsymbol{W}_h \boldsymbol{h}_t)) \tag{4}$$

where $\boldsymbol{W}_r, \boldsymbol{W}_\alpha, \boldsymbol{W}_h$ are all trainable weights, and $\circ$ is the Hadamard product. The next step input $\boldsymbol{x}_{t+1}$ is computed with the predicted operation:

$$\boldsymbol{x}_{t+1} = \boldsymbol{W}_{op} \boldsymbol{r}_t \tag{5}$$

where $\boldsymbol{W}_{op}$ represents the weights of an embedding layer. By iterating over the whole sequence of reasoning steps, we compute the aggregated reasoning-aware attention

$$\boldsymbol{\alpha}^r = \sum_t \boldsymbol{\alpha}_t / T \tag{6}$$

that takes into account all the intermediate attention weights along the reasoning process, where T is the total number of reasoning steps. With the supervision

from the ROIs for different reasoning steps, the model is able to adaptively aggregate attention over time to perform complex visual reasoning.

Finally, following the original multi-modal fusion methods of UpDown [1] and MUTAN [2], we use $\boldsymbol{\alpha}^r$ to determine the contributions of visual features via a dynamic weighting scheme, and the final answer is predicted based on the question and the attended visual features:

$$ans = \text{softmax}\left(\boldsymbol{W}_{ans}\ \text{fuse}(\boldsymbol{W}_{v'}\sum_i \boldsymbol{\alpha}_i^r \boldsymbol{v}_i, \boldsymbol{W}_{q'}\boldsymbol{q})\right) \tag{7}$$

where i denotes the index of visual features (*e.g.,* region proposals or spatial locations), and the fuse$(\cdot)$ operator represents multi-modal fusion used in the baseline models (*e.g.,* low-rank bilinear in UpDown [1] and Tucker decomposition in MUTAN [2]).

Application of AiR-M on BAN [7]. To demonstrate the generality of our method, we further apply the AiR-M on a multi-glimpse co-attention model, *i.e.,* BAN [7]. Previous attention supervision methods [8, 9, 11] typically consider attention as a single-output mechanism, and have difficulty generalizing to multi-glimpse co-attention with multiple attention maps measuring the correlation between vision and language. Differently, our AiR-M method decomposes the reasoning process into a set of reasoning steps that naturally align with the multi-glimpse structure. Therefore, AiR-M with attention and reasoning supervision can guide the models to capture various ROIs with multi-glimpse attention.

Specifically, instead of using a fixed number of glimpses, we dynamically determine the number of glimpses based on the reasoning process. Following the same process as above for visual attention, we jointly compute the reasoning operation $\boldsymbol{r}_t$ and corresponding attention $\boldsymbol{\alpha}_t$ for reasoning step t. The attention $\boldsymbol{\alpha}_t$ is applied to the visual features before computing the co-attention via $\boldsymbol{v}'_t = \sum \boldsymbol{\alpha}_t \boldsymbol{v}$.

Derivation of Ground-truth Attention Weights. The ground-truth attention weights used in the training objective for attention prediction (*i.e.,* $L_{\boldsymbol{\alpha_t}}$ in Equation 1) are derived from the GQA annotations. Specifically, we first extract the ROIs for each operation, and then compute the Intersection of Union (IoU) between each ROI and each input region proposal [1]. The attention weight for each input region proposal is defined as the sum of its IoUs with all ROIs. Finally, the ground-truth attention weights of all input region proposals are normalized with their sum.

Other Implementation Details. We train all of the models following the original settings proposed in the corresponding papers [1, 2, 7]. Please refer to the original papers for further details. Since the original settings are designed for the VQA [4] dataset, we make two modifications to accommodate the differences

between GQA [6] and VQA [4]: (1) We use batch size 150 for UpDown [1] and MUTAN [2], which tends to provide better results than the original settings for all models; (2) Instead of using $G = 8$ glimpses in BAN [7] which leads to a severe overfitting, we follow [6] (*i.e.,* application of another multi-glimpse model MAC [5]) and use $G = 4$.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: cvpr (2018)
2. Ben-Younes, H., Cadène, R., Thome, N., Cord, M.: Mutan: Multimodal tucker fusion for visual question answering. ICCV (2017)
3. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)
4. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: CVPR (2017)
5. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning (2018)
6. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
7. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear Attention Networks. In: NeurIPS. pp. 1571–1581 (2018)
8. Patro, B.N., Anupriy, Namboodiri, V.P.: Explanation vs attention: A two-player game to obtain attention for vqa. In: AAAI (2020)
9. Qiao, T., Dong, J., Xu, D.: Exploring human-like attention supervision in visual question answering. In: AAAI (2018)
10. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: ICCV (2017)
11. Zhang, Y., Niebles, J.C., Soto, A.: Interpretable visual question answering by visual grounding from attention supervision mining. In: WACV. pp. 349–357 (2019)