AiR: Attention with Reasoning Capability

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

Abstract. While attention has been an increasingly popular component in deep neural networks to both interpret and boost performance of models, little work has examined how attention progresses to accomplish a task and whether it is reasonable. In this work, we propose an Attention with Reasoning capability (AiR) framework that uses attention to understand and improve the process leading to task outcomes. We first define an evaluation metric based on a sequence of atomic reasoning operations, enabling quantitative measurement of attention that considers the reasoning process. We then collect human eye-tracking and answer correctness data, and analyze various machine and human attentions on their reasoning capability and how they impact task performance. Furthermore, we propose a supervision method to jointly and progressively optimize attention, reasoning, and task performance so that models learn to look at regions of interests by following a reasoning process. We demonstrate the effectiveness of the proposed framework in analyzing and modeling attention with better reasoning capability and task performance. The code and data are available at https://github.com/szzexpoi/AiR.

Keywords: Attention, Reasoning, Eye-Tracking Dataset

1 Introduction

Recent progress in deep neural networks (DNNs) has resulted in models with significant performance gains in many tasks. Attention, as an information selection mechanism, has been widely used in various DNN models, to improve their ability of localizing important parts of the inputs, as well as task performances. It also enables fine-grained analysis and understanding of the black-box DNN models, by highlighting important information in their decision-making. Recent studies explored different machine attentions and showed varied degrees of agreement on where human consider important in various vision tasks, such as captioning [15, 34] and visual question answering (VQA) [7].

Similar to humans who look and reason actively and iteratively to perform a visual task, attention and reasoning are two intertwined mechanisms underlying the decision-making process. As shown in Fig. 1, answering the question requires humans or machines to make a sequence of decisions based on the relevant regions

* Equal contributions.

of interest (ROIs) (*i.e.,* to sequentially look for the jeans, the girl wearing the jeans, and the bag to the left of the girl). Guiding attention to explicitly look for these objects following the reasoning process has the potential to improve both interpretability and performance of a computer vision model.

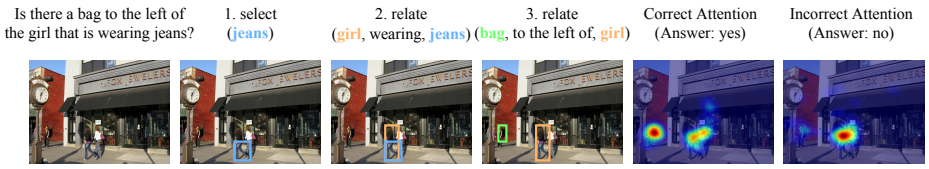

Fig. 1: Attention is an essential mechanism that affects task performances in visual question answering. Eye fixation maps of humans suggest that people who answer correctly look at the most relevant ROIs in the reasoning process (*i.e.,* jeans, girl, and bag), while incorrect answers are caused by misdirected attention

To understand the roles of visual attention in the visual reasoning context, and leverage it for model development, we propose an integrated Attention with Reasoning capability (AiR) framework. It represents the visual reasoning process as a sequence of atomic operations each with specific ROIs, defines a metric and proposes a supervision method that enables the quantitative evaluation and guidance of attentions based on the intermediate steps of the visual reasoning process. A new eye-tracking dataset is collected to support the understanding of human visual attention during the visual reasoning process, and is also used as a baseline for studying machine attention. This framework is a useful toolkit for research in visual attention and its interaction with visual reasoning.

Our work has three distinctions from previous attention evaluation [7, 18, 19, 27] and supervision [29, 30, 45] methods: (1) We go beyond the existing evaluation methods that are either qualitative or focused only on the alignment with outputs, and propose a measure that encodes the progressive attention and reasoning defined by a set of atomic operations. (2) We emphasize the tight correlation between attention, reasoning, and task performance, conducting fine-grained analyses of the proposed method with various types of attention, and incorporating attention with the reasoning process to enhance model interpretability and performance. (3) Our new dataset with human eye movements and answer correctness enables more accurate evaluation and diagnosis of attention.

To summarize, the proposed framework makes the following contributions:

1. A new quantitative evaluation metric (AiR-E) to measure attention in the reasoning context, based on a set of constructed atomic reasoning operations.
2. A supervision method (AiR-M) to progressively optimize attention throughout the entire reasoning process.
3. An eye-tracking dataset (AiR-D) featuring high-quality attention and reasoning labels as well as ground truth answer correctness.

4. Extensive analyses of various machine and human attention with respect to reasoning capability and task performance. Multiple factors of machine attention have been examined and discussed. Experiments show the importance of progressive supervision on both attention and task performance.

2 Related Works

This paper is most closely related to prior studies on the evaluation of attention in visual question answering (VQA) [7, 18, 19, 27]. In particular, the pioneering work by Das *et al.* [7] is the only one that collected human attention data on a VQA dataset and compared them with machine attention, showing considerable discrepancies in the attention maps. Our proposed study highlights several distinctions from related works: (1) Instead of only considering one-step attention and its alignment with a single ground-truth map, we propose to integrate attention with progressive reasoning that involves a sequence of operations each related to different objects. (2) While most VQA studies assume human answers to be accurate, it is not always the case [39]. We collect ground truth correctness labels to examine the effects of attention and reasoning on task performance. (3) The only available dataset [7], with post-hoc attention annotation collected on blurry images using a "bubble-like" paradigm and crowdsourcing, may not accurately reflect the actual attention of the task performers [33]. Our work addresses these limitations by using on-site eye tracking data and QA annotations collected from the same participants. (4) Das *et al.* [7] compared only spatial attention with human attention. Since recent studies [18, 27] suggest that attention based on object proposals are more semantically meaningful, we conduct the first quantitative and principled evaluation of object-based attentions.

This paper also presents a progressive supervision approach for attention, which is related to the recent efforts on improving attention accuracy with explicit supervision. Several studies use different sources of attention ground truth, such as human attention [30], adversarial learning [29] and objects mined from textual descriptions [45], to explicitly supervise the learning of attentions. Similar to the evaluation studies introduced above, these attention supervision studies only consider attention as a single-output mechanism and ignores the progressive nature of the attention process or whether it is reasonable or not. As a result, they fall short of acquiring sufficient information from intermediate steps. Our work addresses these challenges with joint prediction of the reasoning operations and the desired attentions along the entire decision-making process.

Our work is also related to a collection of datasets for eye-tracking and visual reasoning. Eye tracking data is collected to study passive exploration [1, 5, 10, 22, 25, 37] as well as task-guided attention [1, 9, 25]. Despite the less accurate and post-hoc mouse-clicking approximation [7], there has been no eye-tracking data recorded from human participants performing the VQA tasks. To facilitate the analysis of human attention in VQA tasks, we construct the first dataset of eye-tracking data collected from humans performing the VQA tasks. A number of visual reasoning datasets [3, 13, 18, 20, 32, 44] are collected in the form of VQA.

Some are annotated with human-generated questions and answers [3, 32], while others are developed with synthetic scenes and rule-based templates to remove the subjectiveness of human answers and language biases [13, 18, 20, 44]. The one most closely related to this work is GQA [18], which offers naturalistic images annotated with scene graphs and synthetic question-answer pairs. With balanced questions and answers, it reduces the language bias without compromising generality. Their data efforts benefit the development of various visual reasoning models [2, 8, 11, 16, 24, 28, 36, 40–43, 31]. In this work, we use a selection of GQA data and annotations in the development of the proposed framework.

3 Method

Real-life vision tasks require looking and reasoning interactively. This section presents a principled framework to study attention in the reasoning context. It consists of three novel components: (1) a quantitative measure to evaluate attention accuracy in the reasoning context, (2) a progressive supervision method for models to learn where to look throughout the reasoning process, and (3) an eye-tracking dataset featuring human eye-tracking and answer correctness data.

3.1 Attention with Reasoning Capability

To model attention as a process and examine its reasoning capability, we describe reasoning as a sequence of atomic operations. Following the sequence, an intelligent agent progressively attends to the key ROIs at each step and reasons what to do next until eventually making a final decision. A successful decision-making relies on accurate attention for various reasoning operations, so that the most important information is not filtered out but passed along to the final step.

To represent the reasoning process and obtain the corresponding ROIs, we define a vocabulary of atomic operations emphasing the role of attention. These operations are grounded on the 127 types of operations of GQA [18] that completely represent all the questions. As described in Table 1, some operations require attention to a specific object (*query*, *verify*); some require attention to objects of the same category (*select*), attribute (*filter*), or relationship (*relate*);

Table 1: Semantic operations of the reasoning process

Operation	Semantic
Select	Searching for objects from a specific category.
Filter	Determining the targeted objects by looking for a specific attribute.
Query	Retrieving the value of a specific attribute from the ROIs.
Verify	Examining the targeted objects and checking if they have a given attribute.
Compare	Comparing the values of an attribute between multiple objects.
Relate	Connecting different objects through their relationships.
And/Or	Serving as basic logical operations that combine the results of the previous operation(s).

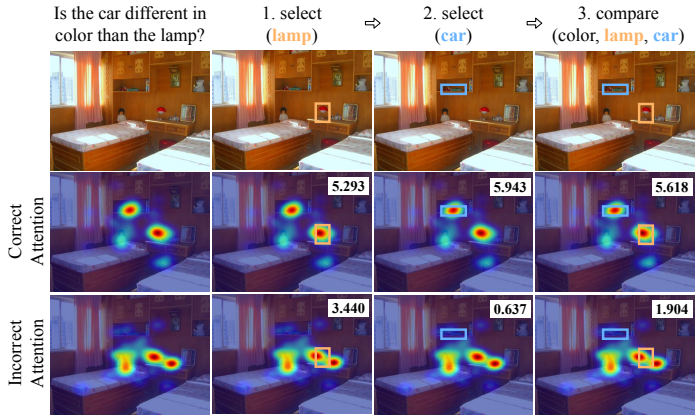

Fig. 2: AiR-E scores of Correct and Incorrect human attention maps, measuring their alignments with the bounding boxes of the ROIs

and others require attention to any (*or*) or all (*and*, *compare*) ROIs from the previous operations. The ROIs of each operation are jointly determined by the type of operation and the scene information (*i.e.,* object categories, attributes and relationships). Given the operation sequence and annotated scene information, we can traverse the reasoning process, starting with all objects in the scene, and sequentially apply the operations to obtain the ROIs at each step. Details of this method are described in the supplementary materials.

3.2 Measuring Attention Accuracy with ROIs

Decomposing the reasoning process into a sequence of operations allows us to evaluate the quality of attention (machine and human attentions) according to its alignment with the ROIs at each operation. Attention can be represented as a 2D probability map where values indicate the importance of the corresponding input pixels. To quantitatively evaluate attention accuracy in the reasoning context, we propose the AiR-E metric that measures the alignment of the attention maps with ROIs relevant to reasoning. As shown in Fig. 2, for humans, a better attention map leading to the correct answer has higher AiR-E scores, while the incorrect attention with lower scores fails to focus on the most important object (*i.e.,* car). It suggests potential correlation between the AiR-E and the task performance. The specific definition of AiR-E is introduced as follows:

Inspired by the Normalized Scanpath Saliency [6] (NSS), given an attention map $A(x)$ where each value represents the importance of a pixel x, we first standardize the attention map into $A^*(x) = (A(x) - \mu)/\sigma$, where μ and σ are the mean and standard deviation of the attention values in $A(x)$, respectively. For each ROI, we compute AiR-E as the average of $A^*(x)$ inside its bounding box B: $\text{AiR-E}(B) = \sum_{x \in B} A^*(x)/|B|$. Finally, we aggregate the AiR-E of all ROIs for each reasoning step:

1. For operations with one set of ROIs (*i.e., select*, *query*, *verify*, and *filter*), as well as *or* that requires attention to one of multiple sets of ROIs, an accurate attention map should align well with at least one ROI. Therefore, the aggregated AiR-E score is the maximum AiR-E of all ROIs.
2. For those with multiple sets of ROIs (*i.e., relate*, *compare*, *and*), we compute the aggregated AiR-E for each set, and take the mean across all sets.

3.3 Reasoning-aware Attention Supervision

For models to learn where to look along the reasoning process, we propose a reasoning-aware attention supervision method (AiR-M) to guide models to progressively look at relevant places following each reasoning operation. Different from previous attention supervision methods [29, 30, 45], the AiR-M method considers the attention throughout the reasoning process and jointly supervises the prediction of reasoning operations and ROIs across the sequence of multiple reasoning steps. Integrating attention with reasoning allows models to accurately capture ROIs along the entire reasoning process for deriving the correct answers.

The proposed method has two major distinctions: (1) integrating attention progressively throughout the entire reasoning process and (2) joint supervision on attention, reasoning operations and answer correctness. Specifically, following the reasoning decomposition discussed in Section 3.1, at the t-th reasoning step, the proposed method predicts the reasoning operation $\boldsymbol{r}_t$, and generates an attention map $\boldsymbol{\alpha}_t$ to predict the ROIs. With the joint prediction, models learn desirable attentions for capturing the ROIs throughout the reasoning process and deriving the answer. The predicted operations and the attentions are supervised together with the prediction of answers:

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

where θ and ϕ are hyperparameters. We use the standard cross-entropy loss L_{ans} and $L_{\boldsymbol{r}_t}$ to supervise the answer and operation prediction, and a Kullback–Leibler divergence loss $L_{\boldsymbol{\alpha}_t}$ to supervise the attention prediction. We aggregate the loss for operation and attention predictions over all reasoning steps.

The proposed AiR-M supervision method is general, and can be applied to various models with attention mechanisms. In the supplementary materials, we illustrate the implementation details for integrating AiR-M with different state-of-the-art models used in our experiments.

3.4 Evaluation Benchmark and Human Attention Baseline

Previous attention data collected under passive image viewing [21], approximations with post-hoc mouse clicks [7], or visually grounded answers [19] may not accurately or completely reflect human attention in the reasoning process. They also do not explicitly verify the correctness of human answers. To demonstrate the effectiveness of the proposed evaluation metric and supervision method, and

to provide a benchmark for attention evaluation, we construct the first eye-tracking dataset for VQA. It, for the first time, enables the step-by-step comparison of how humans and machines allocate attention during visual reasoning.

Specifically, we (1) select images and questions that require humans to actively look and reason; (2) remove ambiguous or ill-formed questions and verify the ground truth answer to be correct and unique; (3) collect eye-tracking data and answers from the same human participants, and evaluate their correctness with the ground-truth answers.

Images and questions. Our images and questions are selected from the balanced validation set of GQA [18]. Since the questions of the GQA dataset are automatically generated from a number of templates based on scene graphs [26], the quality of these automatically generated questions may not be sufficiently high. Some questions may be too trivial or too ambiguous. Therefore, we perform automated and manual screenings to control the quality of the questions. First, to avoid trivial questions, all images and questions are first screened with these criteria: (1) image resolution is at least 320×320 pixels; (2) image scene graph consists of at least 16 relationships; (3) total area of question-related objects does not exceed 4% of the image. Next, one of the authors manually selects 987 images and 1,422 questions to ensure that the ground-truth answers are accurate and unique. The selected questions are non-trivial and free of ambiguity, which require paying close attention to the scene and actively searching for the answer.

Eye-tracking experiment. The eye-tracking data are collected from 20 paid participants, including 16 males and 4 females from age 18 to 38. They are asked to wear a Vive Pro Eye headset with an integrated eye-tracker to answer questions from images presented in a customized Unity interface. The questions are randomly grouped into 18 blocks, each shown in a 20-minute session. The eye-tracker is calibrated at the beginning of each session. During each trial, a question is first presented, and the participant is given unlimited time to read and understand it. The participant presses a controller button to start viewing the image. The image is presented in the center for 3 seconds. The image is scaled such that both the height and width occupy 30 degrees of visual angle (DVA). After that, the question is shown again and the participant is instructed to provide an answer. The answer is then recorded by the experimenter. The participant presses another button to proceed to the next trial.

Human attention maps and performances. Eye fixations are extracted from the raw data using the Cluster Fix algorithm [23], and a fixation map is computed for each question by aggregating the fixations from all participants. The fixation maps are scaled into 256×256 pixels, smoothed using a Gaussian kernel ($\sigma = 9$ pixels, ≈ 1 DVA) and normalized to the range of [0,1]. The overall accuracy of human answers is $77.64 \pm 24.55\%$ (M±SD). A total of 479 questions have consistently correct answers, and 934 have both correct and incorrect answers. The histogram of human answer accuracy is shown in Fig. 3a. We further separate the fixations into two groups based on answer correctness and compute a fixation map for each group. Correct and incorrect answers have comparable numbers of fixations per trial (10.12 *vs.* 10.27), while the numbers of fixations

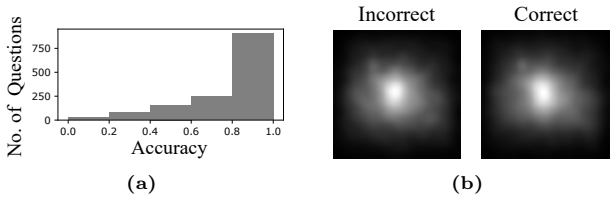

Fig. 3: Distributions of answer accuracy and eye fixations of humans. (a) Histogram of human answer accuracy (b) Center biases of the correct and incorrect attention

for the correct answers have a lower standard deviation across trials (0.99 *vs.* 1.54). Fig. 3b shows the prior distributions of the two groups of fixations, and their high similarity (Pearson's $r = 0.997$) suggests that the answer correctness is independent of center bias. The correct and incorrect fixation maps are considered as two human attention baselines to compare with machine attentions, and also play a role in validating the effectiveness of the proposed AiR-E metric. More illustration is provided in the supplementary video.

4 Experiments and Analyses

In this section, we conduct experiments and analyze various attention mechanisms of humans and machines. Our experiments aim to shed light on the following questions that have yet to be answered:

1. Do machines or humans look at places relevant to the reasoning process? How does the attention process influence task performances? (Section 4.1)
2. How does attention accuracy evolve over time, and what about its correlation with the reasoning process? (Section 4.2)
3. Does guiding models to look at places progressively following the reasoning process help? (Section 4.3)

4.1 Do Machines or Humans Look at Places Important to Reasoning? How Does Attention Influence Task Performances?

First, we measure attention accuracy throughout the reasoning process with the proposed AiR-E metric. Answer correctness is also compared, and its correlation with the attention accuracy reveals the joint influence of attention and reasoning operations to task performance. With these experiments, we observe that humans attend more accurately than machines, and the correlation between attention accuracy and task performance is dependent on the reasoning operations.

We evaluate four types of attentions that are commonly used in VQA models, including spatial soft attention (S-Soft), spatial Transformer attention (S-Trans), object-based soft attention (O-Soft), and object-based Transformer attention (O-Trans). Spatial and object-based attentions differ in terms of their inputs (*i.e.,* image features or regional features), while soft and Transformer attention

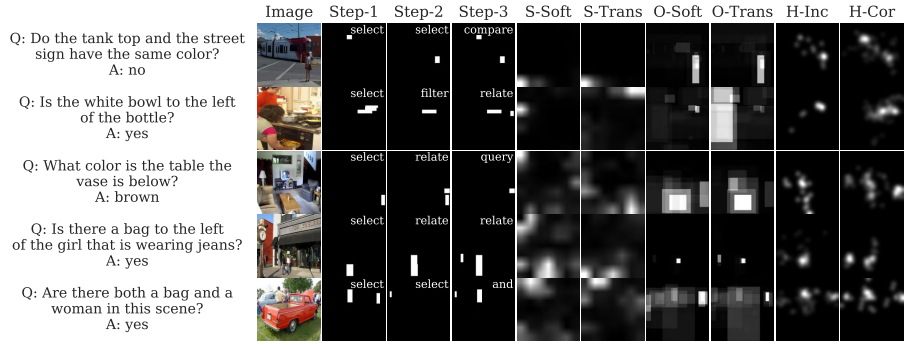

Fig. 4: Example question-answer pairs (column 1), images (column 2), ROIs at each reasoning step (columns 3-5), and attention maps (columns 6-11)

methods differ in terms of the computational methods of attention (*i.e.,* with convolutional layers or matrix multiplication). We use spatial features extracted from ResNet-101 [14] and object-based features from [2] as the two types of inputs, and follow the implementations of [2] and [12] for the soft attention [38] and Transformer attention [35] computation, respectively. We integrate the aforementioned attentions with different state-of-the-art VQA models as backbones. Our observations are general and consistent across various backbones. In the following sections, we use the results on UpDown [2] for illustration (results for the other backbones are provided in the supplementary materials). For human attentions, we denote the fixation maps associated with correct and incorrect answers as H-Cor and H-Inc, and the aggregated fixation map regardless of correctness is denoted as H-Tot. Fig. 4 presents examples of ROIs for different reasoning operations and the compared attention maps.

Attention accuracy and task performance of humans and models. Table 2 quantitatively compares the AiR-E scores and VQA task performance across humans and models with different types of attentions. The task performance for models is the classification score of the correct answer, while the task performance for humans is the proportion of correct answers. Three clear gaps can be observed from the table: (1) Humans who answer correctly have significantly higher AiR-E scores than those who answer incorrectly. (2) Humans consistently outperform models in both attention and task performance. (3) Object-based attentions attend much more accurately than spatial attentions. The low AiR-E of spatial attentions confirms the previous conclusion drawn from the VQA-HAT dataset [7]. By constraining the visual inputs to a set of semantically meaningful objects, object-based attention typically increases the probabilities of attending to the correct ROIs. Between the two object-based attentions, the soft attention slightly outperforms its Transformer counterpart. Since the Transformer attentions explicitly learn the inter-object relationships, they perform better for logical operations (*i.e., and*, *or*). However, due to the complexity of the scenes and fewer parameters used [35], they do not perform as well as soft

Table 2: Quantitative evaluation of AiR-E scores and task 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
	O-Soft	1.334	**1.204**	**1.518**	1.857	**3.241**	**2.243**	**1.586**	2.091
	O-Trans	**1.579**	1.046	1.202	**1.910**	3.041	1.839	1.324	**2.228**
	S-Soft	-0.001	-0.110	0.251	0.413	0.725	0.305	0.145	0.136
	S-Trans	0.060	-0.172	0.243	0.343	0.718	0.370	0.173	0.101
Accuracy	H-Tot	0.700	0.625	0.668	0.732	0.633	0.672	0.670	0.707
	O-Soft	0.604	**0.547**	0.603	0.809	**0.287**	0.483	0.548	0.605
	O-Trans	**0.606**	0.536	**0.608**	**0.832**	0.282	**0.487**	**0.550**	0.592
	S-Soft	0.592	0.520	0.558	0.814	0.203	0.427	0.511	0.544
	S-Trans	0.597	0.525	0.557	0.811	0.211	0.435	0.517	**0.607**

Table 3: Pearson's r between attention accuracy (AiR-E) and task performance. 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**
O-Soft	0.167	**0.217**	-0.022	0.059	**0.331**	0.058	0.003	0.121
O-Trans	0.168	**0.205**	0.090	0.174	**0.298**	0.041	**0.063**	-0.027
S-Soft	0.177	**0.237**	-0.084	0.082	-0.017	-0.170	-0.084	0.066
S-Trans	0.171	**0.210**	-0.152	0.086	-0.024	-0.139	-0.100	**0.270**

attention. The ranks of different attentions are consistent with the intuition and literature, suggesting the effectiveness of the proposed AiR-E metric.

Attention accuracy and task performance among different reasoning operations. Comparing the different operations, Table 2 shows that *query* is the most challenging operation for models. Even with the highest attention accuracy among all operations, the task performance is the lowest. This is probably due to the inferior recognition capability of models compared with humans. To humans, 'compare' is the most challenging in terms of task performance, largely because it often appears in complex questions that require close attention to multiple objects and thus take longer processing time. Since models can process multiple input objects in parallel, their performance is not highly influenced by the number of objects to look at.

Correlation between attention accuracy and task performance. The similar rankings of AiR-E and task performance suggest a correlation between attention accuracy and task performance. To further investigate this correlation on a sample basis, for each attention and operation, we compute the Pearson's r between the attention accuracy and task performance across different questions.

As shown in Table 3, human attention accuracy and task performance are correlated for most of the operations (up to $r = 0.329$). The correlation is higher than most of the compared machine attentions, suggesting that humans' task

performance is more consistent with their attention quality. In contrast, though commonly referred as an interface for interpreting models' decisions [7, 19, 27], spatial attention maps do not reflect the decision-making process of models. They typically have very low and even negative correlations (*e.g., relate*, *select*). By limiting the visual inputs to foreground objects, object-based attentions achieve higher attention-answer correlations.

The differences of correlations between operations are also significant. For the questions requiring focused attention to answer (*i.e.,* with *query* and *compare* operations), the correlations are relatively higher than the others.

4.2 How Does Attention Accuracy Evolve Throughout the Reasoning Process?

To complement our previous analysis on the spatial allocation of attentions,we move forward to analyze the spatiotemporal alignment of attentions. Specifically, we analyze the AiR-E scores according to the chronological order of reasoning operations. We show in Fig. 5a that the AiR-E scores peak at the 3^{rd} or 4^{th} steps, suggesting that human and machine attentions focus more on the ROIs closely related to the final task outcome, instead of the earlier steps. In the rest of this section, we focus our analysis on the spatiotemporal alignment between multiple attention maps and the ROIs at different reasoning steps. In particular, we study the change of human attention over time, and compare it with multi-glimpse machine attentions. Our analysis reveals the significant spatiotemporal discrepancy between human and machine attentions.

Do human attentions follow the reasoning process? First, to analyze the spatiotemporal deployment of human attention in visual reasoning, we group the fixations into three temporal bins (0-1s, 1-2s and 2-3s), and compute AiR-E scores for each fixation map and reasoning step (see Fig. 5b-c). Humans start exploration (0-1s) with relatively low attention accuracy. After the initial exploration, human attention shows improved accuracy across all reasoning steps (1-2s), and particularly focuses on the early-step ROIs. In the final steps (2-3s), depending on the correctness of answers, human attention either shifts to the ROIs at later stages (correct), or becomes less accurate with lowered AiR-E scores (incorrect). Such observations suggest high spatiotemporal alignments between human attention and the sequence of reasoning operations.

Do machine attentions follow the reasoning process? Similarly, we evaluate multi-glimpse machine attentions. We compare the stacked attention from SAN [40], compositional attention from MAC [17] and the multi-head attention [11, 43], which all adopt the object-based attention. Fig. 5d-f shows that multi-glimpse attentions do not evolve with the reasoning process. Stacked attention's first glimpse already attends to the ROIs at the 4^{th} step, and the other glimpses contribute little to the attention accuracy. Compositional attention and multi-head attention consistently align the best with the ROIs at the 3^{rd} or 4^{th} step, and ignore those at the early steps.

The spatiotemporal correlations indicate that following the correct order of reasoning operations is important for humans to attend and answer correctly.

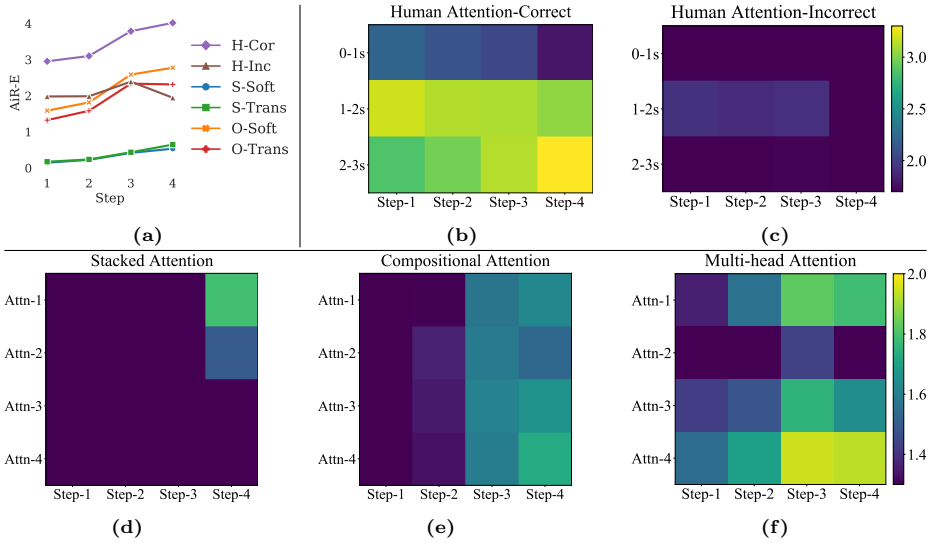

Fig. 5: Spatiotemporal accuracy of attention throughout the reasoning process. (a) shows the AiR-E of different reasoning steps for human aggregated attentions and single-glimpse machine attentions, (b)-(c) AiR-E scores for decomposed human attentions with correct and incorrect answers, (d)-(f) AiR-E for multi-glimpse machine attentions. For heat maps shown in (b)-(f), the x-axis denotes different reasoning steps while the y-axis corresponds to the indices of attention maps

In contrast, models tend to directly attend to the final ROIs, instead of shifting their attentions progressively.

4.3 Does Progressive Attention Supervision Improve Attention and Task Performance?

Experiments in Section 4.1 and Section 4.2 suggest that attention towards ROIs relevant to the reasoning process contributes to task performance, and furthermore, the order of attention matters. Therefore, we propose to guide models to look at places important to reasoning in a progressive manner. Specifically, we propose to supervise machine attention along the reasoning process by jointly optimizing attention, reasoning operations, and task performance (AiR-M, see Section 3.3). Here we investigate the effectiveness of the AiR-M supervision method on three VQA models, *i.e.,* UpDown [2], MUTAN [4], and BAN [24]. We compare AiR-M with a number of state-of-the-art attention supervision methods, including supervision from human-like attention (HAN) [30], attention supervision mining (ASM) [45] and adversarial learning (PAAN) [29]. Note that while the other compared methods are typically limited to supervision on a single attention map, our AiR-M method is generally applicable to various VQA models with single or multiple attention maps (*e.g.,* BAN [24]).

Table 4: Comparative results on GQA test sets (test-dev and test-standard). We report the single-model performance trained on the balanced training set of GQA

	UpDown [2]		MUTAN [4]		BAN [24]	
	dev	standard	dev	standard	dev	standard
w/o Supervision	51.31	52.31	50.78	51.16	50.14	50.38
PAAN [29]	48.03	48.92	46.40	47.22	n/a	n/a
HAN [30]	49.96	50.58	48.76	48.99	n/a	n/a
ASM [45]	52.96	53.57	51.46	52.36	n/a	n/a
AiR-M	**53.46**	**54.10**	**51.81**	**52.42**	**53.36**	**54.15**

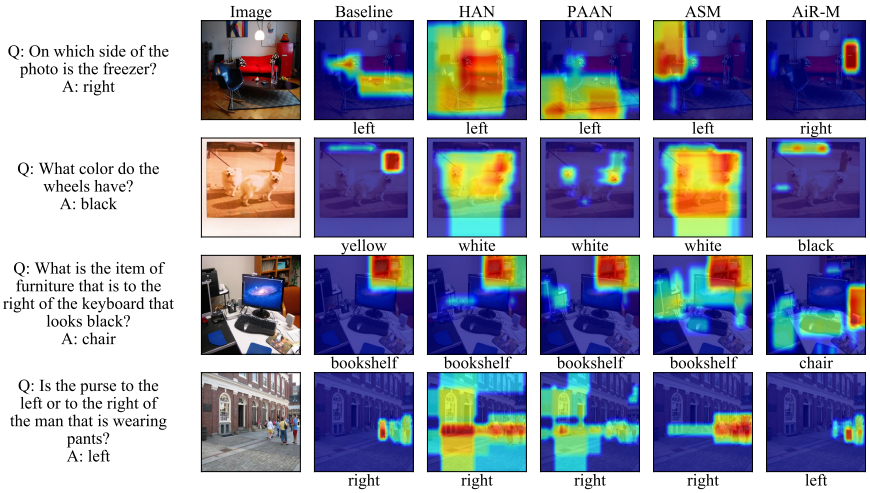

Fig. 6: Qualitative comparison between attention supervision methods, where Baseline refers to UpDown [2]. For each row, from left to right are the questions and the correct answers, input images, and attention maps learned by different methods. The predicted answers associated with the attentions are shown below its respective attention map

According to Table 4, the proposed AiR-M supervision significantly improves the performance of all baselines and consistently outperforms the other attention supervision methods. Two of the compared methods, HAN and PAAN, fail to improve the performance of object-based attention. Supervising attention with knowledge from objects mined from language, ASM [45] is able to consistently improve the performance of models. However, without considering the intermediate steps of reasoning, it is not as effective as the proposed method.

Fig. 6 shows the qualitative comparison between supervision methods. The proposed AiR-M not only directs attention to the ROIs most related to the answers (*i.e.,* freezer, wheel, chair, purse), but also highlights other important ROIs mentioned in the questions (*i.e.,* keyboard, man), thus reflecting the entire reasoning process, while attentions in other methods fail to localize these ROIs.

Table 5 reports the AiR-E scores across operations. It shows that the AiR-M supervision method significantly improves attention accuracy (attention aggre-

Table 5: AiR-E scores of the supervised attentions

Attention	and	compare	filter	or	query	relate	select	verify
Human	2.197	2.669	2.810	2.429	3.951	3.516	2.913	3.629
AiR-M	**2.396**	**2.553**	**2.383**	**2.380**	3.340	**2.862**	**2.611**	**4.052**
Baseline [2]	1.859	1.375	1.717	2.271	**3.651**	2.448	1.796	2.719
ASM	1.415	1.334	1.443	1.752	2.447	1.884	1.584	2.265
HAN	0.581	0.428	0.468	0.607	1.576	0.923	0.638	0.680
PAAN	1.017	0.872	1.039	1.181	2.656	1.592	1.138	1.221

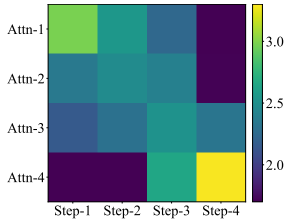

Fig. 7: Alignment between the proposed attention and reasoning process

gated across different steps), especially on those typically positioned in early steps (*e.g., select*, *compare*). In addition, the AiR-M supervision method also aligns the multi-glimpse attentions better according to their chronological order in the reasoning process (see Fig. 7 and the supplementary video), showing progressive improvement of attention throughout the entire process.

5 Conclusion

We introduce AiR, a novel framework with a quantitative evaluation metric (AiR-E), a supervision method (AiR-M), and an eye-tracking dataset (AiR-D) for understanding and improving attention in the reasoning context. Our analyses show that accurate attention deployment can lead to improved task performance, which is related to both the task outcome and the intermediate reasoning steps. Our experiments also highlight the significant gap between models and humans on the alignment of attention and reasoning process. With the proposed attention supervision method, we further demonstrate that incorporating the progressive reasoning process in attention can improve the task performance by a considerable margin. We hope that this work will be helpful for future development of visual attention and reasoning method, and inspire the analysis of model interpretability throughout the decision-making process.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

1. Alers, H., Liu, H., Redi, J., Heynderickx, I.: Studying the effect of optimizing the image quality in saliency regions at the expense of background content. In: Image Quality and System Performance VII. vol. 7529, p. 752907. International Society for Optics and Photonics (2010)
2. 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)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
4. Ben-Younes, H., Cadène, R., Thome, N., Cord, M.: Mutan: Multimodal tucker fusion for visual question answering. ICCV (2017)
5. Borji, A., Itti, L.: Cat2000: A large scale fixation dataset for boosting saliency research. arXiv preprint arXiv:1505.03581 (2015)
6. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
7. Das, A., Agrawal, H., Zitnick, C.L., Parikh, D., Batra, D.: Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2016)
8. Do, T., Do, T.T., Tran, H., Tjiputra, E., Tran, Q.D.: Compact trilinear interaction for visual question answering. In: ICCV (2019)
9. Ehinger, K.A., Hidalgo-Sotelo, B., Torralba, A., Oliva, A.: Modelling search for people in 900 scenes: A combined source model of eye guidance. Visual cognition **17**(6-7), 945–978 (2009)
10. Fan, S., Shen, Z., Jiang, M., Koenig, B.L., Xu, J., Kankanhalli, M.S., Zhao, Q.: Emotional attention: A study of image sentiment and visual attention. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp. 7521–7531 (2018)
11. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 457–468 (2016)
12. Gao, P., Jiang, Z., You, H., Lu, P., Hoi, S.C.H., Wang, X., Li, H.: Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: CVPR (2019)
13. 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)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. He, S., Tavakoli, H.R., Borji, A., Pugeault, N.: Human attention in image captioning: Dataset and analysis. In: ICCV (2019)
16. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. In: ICCV (2017)
17. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning (2018)

18. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
19. Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: CVPR (2018)
20. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
21. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
22. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th international conference on computer vision. pp. 2106–2113. IEEE (2009)
23. König, S.D., Buffalo, E.A.: A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. Journal of Neuroscience Methods **227**, 121 – 131 (2014)
24. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear Attention Networks. In: NeurIPS. pp. 1571–1581 (2018)
25. Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? Journal of vision **14**(3), 14–14 (2014)
26. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
27. Li, W., Yuan, Z., Fang, X., Wang, C.: Knowing where to look? analysis on attention of visual question answering system. In: ECCV Workshops (2018)
28. Mascharka, D., Tran, P., Soklaski, R., Majumdar, A.: Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In: CVPR (2018)
29. Patro, B.N., Anupriy, Namboodiri, V.P.: Explanation vs attention: A two-player game to obtain attention for vqa. In: AAAI (2020)
30. Qiao, T., Dong, J., Xu, D.: Exploring human-like attention supervision in visual question answering. In: AAAI (2018)
31. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: ICCV (2019)
32. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR (2016)
33. Tavakoli, H.R., Ahmed, F., Borji, A., Laaksonen, J.: Saliency revisited: Analysis of mouse movements versus fixations. In: CVPR (2017)
34. Tavakoli, H.R., Shetty, R., Borji, A., Laaksonen, J.: Paying attention to descriptions generated by image captioning models. In: ICCV (2017)
35. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)
36. Wu, J., Mooney, R.: Self-critical reasoning for robust visual question answering. In: NeurIPS (2019)
37. Xu, J., Jiang, M., Wang, S., Kankanhalli, M.S., Zhao, Q.: Predicting human gaze beyond pixels. Journal of vision **14**(1), 28–28 (2014)

38. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057 (2015)
39. Yang, C.J., Grauman, K., Gurari, D.: Visual question answer diversity. In: Sixth AAAI Conference on Human Computation and Crowdsourcing (2018)
40. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)
41. Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., Tenenbaum, J.: Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In: NeurIPS, pp. 1031–1042 (2018)
42. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: CVPR (2019)
43. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: ICCV (2017)
44. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR (2019)
45. Zhang, Y., Niebles, J.C., Soto, A.: Interpretable visual question answering by visual grounding from attention supervision mining. In: WACV. pp. 349–357 (2019)