Attention in Reasoning: Dataset, Analysis, and Modeling

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao

Abstract—While attention has been an increasingly popular component in deep neural networks to both interpret and boost the 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 a 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 attention mechanisms on their reasoning capability and how they impact task performance. To improve the attention and reasoning ability of visual question answering models, we propose to supervise the learning of attention progressively along the reasoning process and to differentiate the correct and incorrect attention patterns. 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.

Index Terms—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 [1], [2], [3], [4], [5], [6], [6], [7], [8], [9], to improve their ability of localizing important parts of the inputs, as well as task performance. It also enables fine-grained analyses and understanding of the black-box DNN models, by highlighting important information in their decision-making process. Recent studies explored different machine attention mechanisms and showed varied degrees of agreement on where humans consider important in various vision tasks, such as image captioning [10], [11] and visual question answering (VQA) [12].

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 regions 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 in Fig. 1a), and avoid the distraction from visually salient but task-irrelevant information (*i.e.,* the skirt in Fig. 1b). Guiding attention to explicitly look for these objects following the reasoning process has the potential to improve both the interpretability and the performance of a computer vision model.

To understand the role of visual attention in VQA, and leverage attention 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 that enables the quantitative evaluation of attention, and proposes two supervision methods that guide attention based on the differentiation of attention patterns and 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 four distinctions from previous attention evaluation [12], [13], [14], [15] and supervision [16], [17], [18] methods: (1) We go beyond the existing evaluation methods that are either qualitative or focused only on the alignment of attention outputs, and propose a measure that encodes the progressive attention and reasoning defined by a set of atomic operations. (2) Focusing on the tight correlation among attention, reasoning, and task performance, we conduct fine-grained analyses to answer various research questions about different types of attention. (3) We jointly supervise machine attention with the reasoning data, so that it can progressively focus on different regions of interest in each step of the reasoning process. (4) We help machines avoid salient distractors by guiding their attention with both correct and incorrect attention patterns. (5) 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 progressive attention supervision method (AiR-M) to optimize the reasoning operations and the allocation

- *The authors are with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455. E-mail: see http://www-users.cs.umn.edu/~qzhao/. The first two authors have equal contributions.*

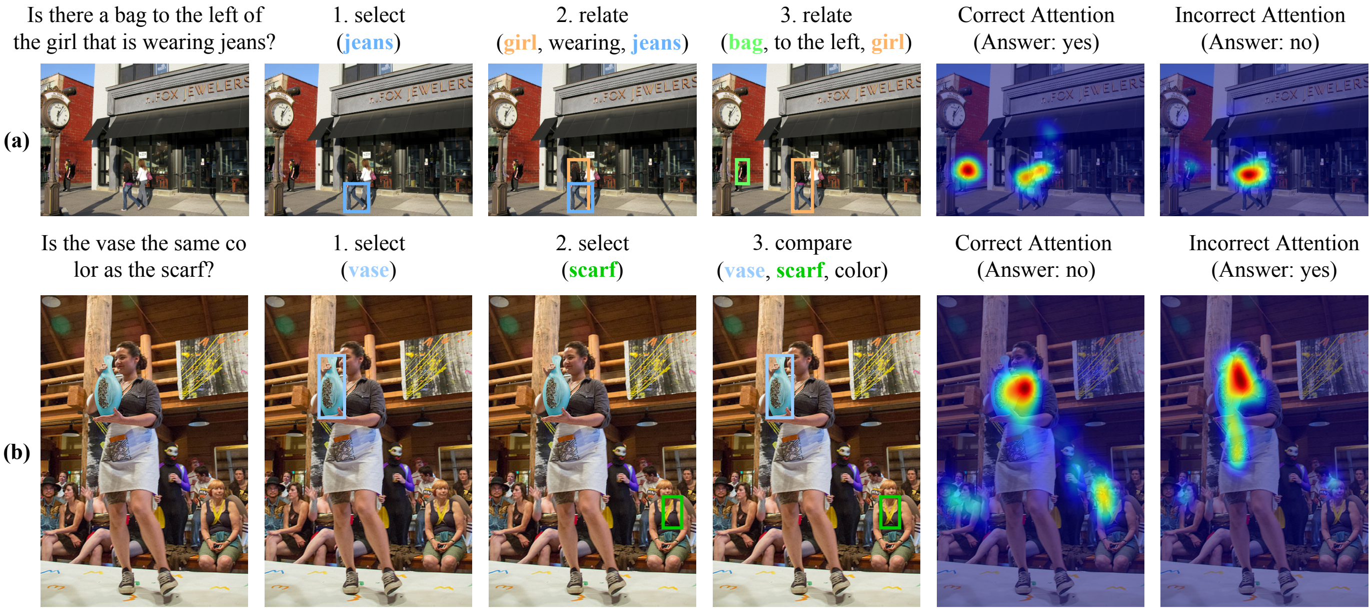

Is there a bag to the left of the girl that is wearing jeans?
1. select (jeans)
2. relate (girl, wearing, jeans)
3. relate (bag, to the left, girl)
Correct Attention (Answer: yes)
Incorrect Attention (Answer: no)
(a)
Is the vase the same co lor as the scarf?
1. select (vase)
2. select (scarf)
3. compare (vase, scarf, color)
Correct Attention (Answer: no)
Incorrect Attention (Answer: yes)
(b)

Fig. 1: Attention is an essential mechanism that affects task performances in visual question answering. (a) People who answer correctly look at the most relevant ROIs in the reasoning process (*i.e.*, jeans, girl, and bag). (b) Incorrect answers can be caused by misdirected attention towards salient distractors (*i.e.*, the skirt).

of machine attention throughout the entire reasoning process,

3) a correctness-aware attention supervision method (AiR-C) that for the first time incorporates both correct and incorrect attention to guide the learning of machine attention,
4) an eye-tracking dataset (AiR-D) featuring high-quality attention and reasoning labels as well as ground truth answer correctness,
5) extensive analyses of various human and machine attention mechanisms with respect to attention accuracy and task performance. Multiple factors of attention in the reasoning process have been examined and discussed. Experiments show the significance of the proposed attention dataset, evaluation metric, and supervision methods.

In particular, this paper extends our preliminary study [19] in the following aspects:

1) We propose a new AiR-C method that for the first time considers attention and answer correctness during the learning of attention. It jointly leverages both correct and incorrect attention patterns as positive and negative guidance to supervise machine attention (Section 3.4 and Section 4.6).
2) We introduce a new hold-out test set for AiR-D to facilitate future research on attention modeling. It consists of new images, questions, answers, and eye-tracking data. The 406 questions of this new dataset are on a different set of 319 images. It provides a new benchmark for attention studies and can be used to test generalizability of models (Section 3.5).
3) We conduct new analyses about the inter-subject consistency of the eye-tracking data, and find that human attention in the VQA task is highly consistent (Section 3.5).
4) To demonstrate the impacts of task information on attention allocation, we conduct a new quantitative study to investigate the attention difference when answering different questions about the same image (Section 4.1).
5) To understand how attention affects task performance, we explicitly compare the attention between correct and incorrect answers to the same question, which shows interesting observations and motivates the use of incorrect attention for models. (Section 4.4).
6) We conduct a new experiment to analyze the correlation as well as inconsistency between attention accuracy and reasoning performance, which suggests the significance of learning high-quality attention for visual reasoning (Section 4.7).
7) We extend the ablation study for the proposed AiR-M method for more complete and thorough discussion (Section 4.5). We have also explicitly discussed the advantages of the new AiR-C method with new evaluation results, in terms of improving the attention accuracy and answer accuracy (Section 4.6).
8) We extend and reorganize Section 2 to include a more comprehensive review of related studies. In particular, on human visual attention, we review attention datasets, models, and their applications in computer vision tasks (Section 2.1).

The rest of the paper is organized as follows. In Section 2, we introduce the related studies on visual attention and reasoning. In Section 3, we present the details of the proposed framework. Section 4 reports the experiments and analyses on various attention mechanisms. Finally, in Section 5, we conclude this paper and provide directions for future work.

2 Related Works

In this section, we briefly review related literature on human attention (Section 2.1), evaluation of machine attention in

Dataset	No. of Scenes	Task	HPA	RP
MIT-1003 [20]	1003	PV	✗	✗
EMOd [21]	1019	PV	✗	✗
DHF1K [22]	1000	PV	✗	✗
CAMO [23]	120	PV	✗	✗
Webpage Saliency [24]	149	Web browsing	✗	✗
EGTEA Gaze+ [25]	86	Cooking	✗	✗
DR(eye)VE [26]	74	Driving	✗	✗
IQVA [27]	975	VQA	✓	✗
AiR-D	1828	VQA	✓	✓

TABLE 1: A comparison between different eye-tracking datasets. PV: passive viewing. HPE: human performance annotation. RP: reasoning process.

VQA (Section 2.2), supervision of machine attention in VQA (Section 2.3), and visual reasoning datasets (Section 2.4).

2.1 Human Visual Attention

This paper is related to a collection of human visual attention studies. Leveraging biologically-inspired filters, attention models compute a probability map that indicates where humans look when freely observing an image or a video [28]. Early computational models of attention focus on studying the bottom-up mechanism driven by visual stimuli. To evaluate attention models and train data-driven algorithms for attention prediction, many eye-tracking datasets [20], [21], [22], [29], [30], [31], [32], [33], [34], [35], [36], [37] have been developed. Unlike the bottom-up mechanism, the top-down mechanism directs human visual attention using a task, which attracts growing research interests. Eye-tracking datasets have been built to study where humans look in various vision tasks (see Table 1), including visual search in 2D [38], [39], [40], [41] or 3D images [42], and dynamic videos [22], [23], action recognition [43], [44], web-browsing [24], [45], cooking [25], driving [26], and video-gaming [46]. The above attention datasets have empowered data-driven models, especially deep neural networks with remarkable learning abilities [47], so that the performance gap between attention prediction models and humans has been significantly reduced. Human attention datasets and models have also contributed to the development of many computer vision applications [48], such as object recognition [49], [50], scene classification [51], salient object segmentation [52], [53], video summarization [54], *etc.* In this work, to facilitate the analysis of human attention in VQA, we construct this new eye-tracking dataset collected from humans performing the VQA tasks.

2.2 Evaluation of Machine Attention in VQA

This paper is closely related to prior studies on the evaluation of machine attention mechanisms in VQA [12], [13], [14], [15]. In particular, the pioneering work by Das *et al.* [12] is the only one that collected human attention data for VQA 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 related to different objects. (2) While most VQA studies assume human answers to be accurate, it is not always the case [55]. We collect ground truth correctness labels to examine the effects of attention and reasoning on task performance, and investigate how humans and machines prioritize their attention in search of diverse answers. (3) The only available dataset [12], 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 [56]. Our work addresses these limitations by using on-site eye-tracking data and QA annotations collected from the same participants. (4) Das *et al.* [12] only compared spatial attention with human attention. Since recent studies [13], [15] suggest that attention based on object proposals is more semantically meaningful, we conduct the first quantitative and principled evaluation of object-based attention.

2.3 Supervision of Machine Attention in VQA

This paper presents supervision approaches for the learning of attention mechanisms for VQA, which is related to the recent efforts in improving machine attention accuracy with explicit supervision. Several studies use different sources of attention ground truth, such as human visual attention [17], adversarial learning [16], and objects mined from textual descriptions [18], to explicitly supervise the learning of machine attention. Similar to the attention evaluation studies introduced above, these attention supervision studies only consider attention as a single-output mechanism, and optimize models to attend to all ROIs with a single glimpse. They typically lead to outspread attention maps that cannot focus on the most relevant regions. They are also agnostic to the reasoning process and fail to acquire sufficient information from intermediate reasoning steps. Besides, these methods only consider the attention ground truths that positively contribute to the correct answers, but do not explicitly identify salient distractors that may lead to incorrect answers. Our work addresses these challenges from two distinct perspectives: (1) jointly predicting the reasoning operations and the desired attention throughout the entire process, enabling the learning of progressive and reasoning-aware attention, and (2) supervising models with both correct and incorrect attention to improve their attention outputs and answers.

2.4 Visual Reasoning Datasets

Several visual reasoning datasets [13], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66] have been collected in the form of VQA. Some are annotated with human-generated questions and answers [57], [60], while others are developed with synthetic scenes and rule-based templates to remove the subjectivity of human answers and the language bias [13], [58], [59], [61]. The one most closely related to this work is GQA [13], 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 [5], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80]. In this work, we use a selection of GQA

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).

TABLE 2: Semantic operations of the reasoning process.

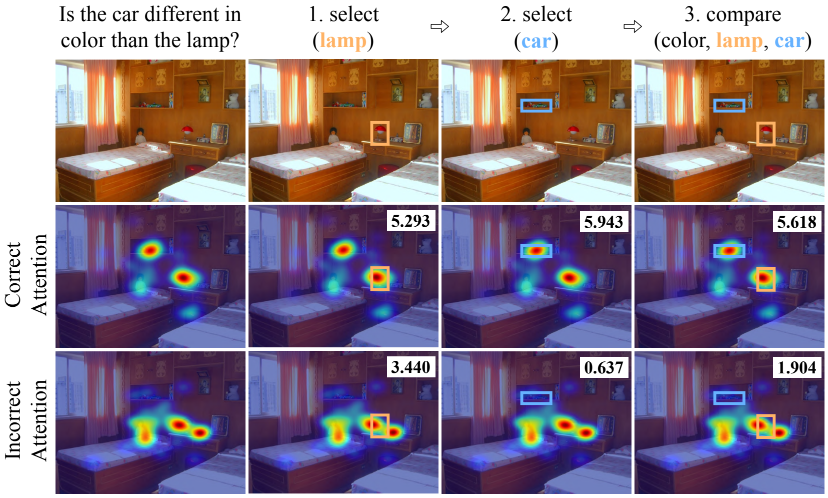


Fig. 2: AiR-E scores of Correct and Incorrect human attention maps. They measure the alignment between attention and the bounding boxes of ROIs.

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 for visual reasoning, (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 method relies on accurate attention for various reasoning operations, so that the most important information is not filtered out but passed throughout to the final step.

To represent the reasoning process and obtain the corresponding ROIs, we define a vocabulary of atomic operations emphasizing the role of attention. These operations are grounded on the 127 types of operations of GQA [13] that completely represent all questions. We define the operations by characterizing and abstracting the complex functional programs of the GQA dataset. Specifically, we define each operation as a triplet, *i.e.,* <operation, attribute, category>, and categorize the original operations in the GQA program based on their semantic meanings: (1) For the original operations that exactly align with our definitions, we 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 operations with similar semantic meanings. For example, we convert "different color object A and object B" to <compare, color, category A and category B>. As described in Table 2, our operations cover various situations for attention allocation: some require attention to a specific object (*query*, *verify*); some require attention to objects of the same category (*select*), attribute (*filter*), or relationship (*relate*); and others require attention to any (*or*) or all (*and*, *compare*) ROIs from the previous operations.

Upon obtaining the operation for each reasoning step, we determine its corresponding ROIs by jointly considering both the semantic meaning of operation and the scene information (*i.e.,* object categories, attributes, and relationships in the scene graphs):

- **Select**: The ROIs belong to a specific category of objects. We query all objects in the scene graph and select those

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3114582, IEEE Transactions on Pattern Analysis and Machine Intelligence

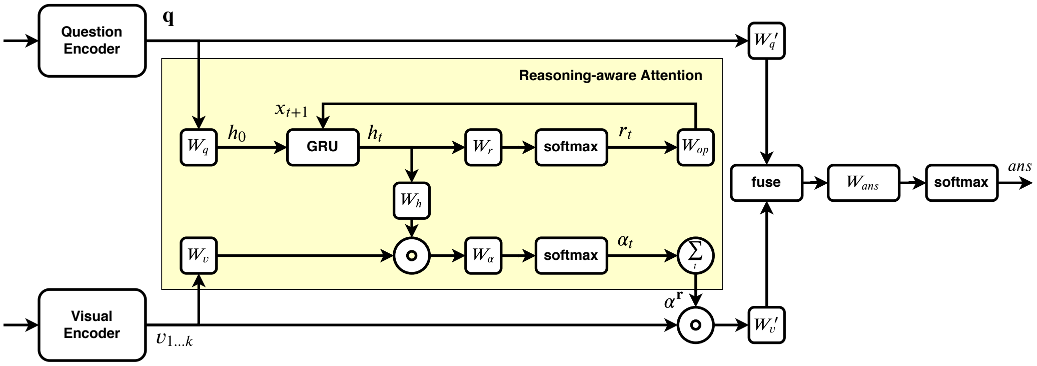


Fig. 3: Network architecture of the proposed AiR-M method.

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 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 [13], *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 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 a nonexistent object, the top-k ($k = 20$) co-existing objects in the scene are selected as the corresponding ROIs.

The aforementioned paradigm allows us to efficiently traverse the reasoning process, starting with all objects in the scene, and sequentially investigating the operation and ROIs at each step. It enables the evaluation of attention accuracy throughout the continuous reasoning process (Section 3.2), and the progressive supervision that guides models to learn where to look following the process (Section 3.3).

3.2 Measuring Attention Accuracy with ROIs

Decomposing the reasoning process into a sequence of operations allows us to evaluate the quality of machine or human attention 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 a 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 [81] (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: 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*) *or* that requires attention to one of the 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 throughout 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 [16], [17], [18], 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 throughout the entire reasoning process for deriving the correct answers.

As shown in Fig. 3, 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, the proposed method takes the question features q and the visual features V as the inputs, and uses a Gated Recurrent Unit [82] (GRU) to sequentially predict the operations $\boldsymbol{r}_t$ and the desired attention weights $\boldsymbol{\alpha}_t$ at the t-th step. At the beginning of the reasoning process, the hidden state of GRU $\boldsymbol{h}_0$ with the question features $\boldsymbol{q}$ is defined as:

$$\boldsymbol{h}_0 = \boldsymbol{W}_q \boldsymbol{q}, \tag{1}$$

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 $\boldsymbol{\alpha}_t$:

$$\boldsymbol{r}_t = \text{softmax}(\boldsymbol{W}_r \boldsymbol{h}_t), \tag{2}$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\boldsymbol{W}_\alpha(\boldsymbol{W}_v \boldsymbol{v} \circ \boldsymbol{W}_h \boldsymbol{h}_t)) \tag{3}$$

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:

$$\boldsymbol{x}_{t+1} = \boldsymbol{W}_{op} \boldsymbol{r}_t \tag{4}$$

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

$$\boldsymbol{\alpha}^r = \sum_t \boldsymbol{\alpha}_t / T \tag{5}$$

that takes into account all 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 can adaptively aggregate attention over time to perform complex visual reasoning.

With the joint prediction of the operation $\boldsymbol{r}_t$ and the attention $\boldsymbol{\alpha}_t$, models learn desirable attention for capturing the ROIs throughout the reasoning process and deriving the answer. The predicted operations and attention outputs are supervised together with the prediction of answers:

$$L = L_{ans} + \theta \sum_t L_{\boldsymbol{\alpha_t}} + \phi \sum_t L_{\boldsymbol{r}_t} \tag{6}$$

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 ground-truth attention map $L_{\boldsymbol{\alpha_t}}$ is derived from our decomposed reasoning process. Specifically, we first extract ROIs for the current reasoning step t, and then compute the Intersection of Union (IoU) between each ROI and each input region proposal [5]. The attention weight for each region proposal is defined as the sum of its IoUs with all ROIs. Finally, the ground truth attention map is constructed by normalizing the attention weights with the sum of weights for all input region proposals.

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 Correctness-aware Attention Supervision

Successful visual reasoning requires not only attention to regions of interest throughout the decision-making process, but also avoiding visually salient distractors that commonly lead to problematic answers. To address this need, we further propose a Correctness-aware Attention Supervision method (AiR-C) that uses both correct and incorrect attention patterns to guide the learning of machine attention. The differentiation between the correct and incorrect attention patterns reveals important cues for visual reasoning: on the one hand, correct attention captures the ROIs most relevant to the task, providing essential information that leads to the correct answer. On the other hand, the incorrect attention highlights the salient distractors that commonly lead to wrong answers (Section 4.4), and enables the models to avoid these hard-negative regions. To our best knowledge, despite many efforts on the learning of attention [16], [17], [18], we are the first to propose the usage of incorrect attention in the supervision of machine attention.

Specifically, we introduce supervision of the incorrect attention using a negative cross-entropy loss:

$$L^-_{att} = \sum_p M^-_p \log \alpha_p \tag{7}$$

where M^- denotes the incorrect attention map, α represents the predicted model attention, and p corresponds to different positions within the maps. The loss encourages models to avoid the distractors, while at the same time allows them to freely explore the other positions. The overall training objective can be formulated as follows:

$$L = L_{ans} + \theta L^+_{att} + \phi L^-_{att} \tag{8}$$

where L_{ans} is the answer prediction loss, L^+_{att} is the attention loss for correct attention (*e.g.*, [18]), θ and ϕ are the hyperparameters.

Given a question and the corresponding image, we construct the ground truth incorrect attention maps by mining the top-k frequently mentioned ROIs in other questions on the same image. We empirically set $k = 3$ in our experiments and exclude those highly overlapping with the relevant ROIs. The rest are considered as hard-negative ROIs used to supervise the incorrect attention. The overlapping area between two ROIs is measured as the proportion of intersection I:

$$I_{j,k} = \frac{O^-_j \cap O^+_k}{min(O^-_j, O^+_k)} \tag{9}$$

where O^- denotes the mined ROIs and O^+ represents the ROIs relevant to the question. For the k_{th} mined ROI O^-_k, we iteratively compute its overlapping areas with every ROI in O^+. If the maximum area is smaller than a threshold (*i.e.,* 0.3), we consider O^-_k as a valid hard-negative ROI. The aforementioned method efficiently locates the hard-negative ROIs that are visually salient, but irrelevant to the given question. Finally, the selected hard-negative ROIs are aggregated into an incorrect attention map to guide the training of models.

3.5 Evaluation Benchmark and Human Attention Baseline

Previous attention data collected under passive image viewing [83], approximations with post-hoc mouse clicks [12], or visually grounded answers [14] 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.

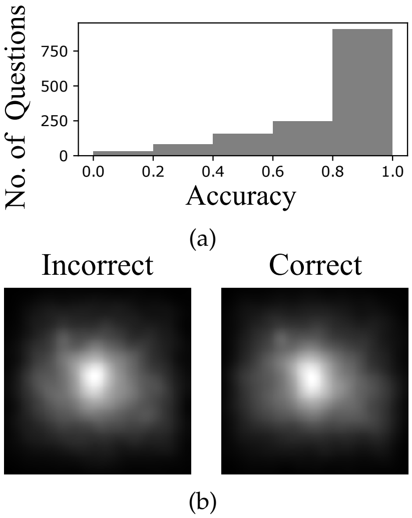
Specifically, we (1) select images and questions that require humans to actively look and reason; (2) remove ambiguous or ill-formed questions and verify the ground truth answer to be correct and unique; (3) collect eye-tracking data and answers from the same human participants, and evaluate their correctness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

Images and questions. Our images and questions are selected from the balanced validation set of GQA [13]. Since the questions of the GQA dataset are automatically generated from a number of templates based on scene graphs [84], 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 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 requires paying close attention to the scene and actively searching for the answer.

In addition, to facilitate future research on task-driven attention modeling, we also introduce a new hold-out test set that contains 319 images and 406 questions. The average answer accuracy of the questions is 65.42%, with a 26.38% standard deviation. Eye-tracking data on this test set will not be released to the public. This test set will provide a new benchmark for gaze prediction in the visual reasoning context and will enable studies on the generalizability of attention modeling across different questions and answers.

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 [85], 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. 4a. To quantify the inter-subject consistency in eye fixations, following [86], we randomly select data from half of the subjects and evaluate their fixation maps against the other half using the AUC-Judd [87] metric. We observe a high inter-subject consistency (*i.e.,* AUC-Judd=0.895) of eye fixations in the VQA task, which suggests the existence of consistently important visual cues that attract human attention in order to answer the questions. 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 number of fixations for the correct answers has a lower standard deviation across trials (0.99 *vs.* 1.54). Fig. 4b 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 attention outputs, and also play a role in validating the effectiveness of the proposed AiR-E metric. Illustrative examples are presented in the supplementary video.



(a)

(b)

Fig. 4: Distributions of answer accuracy and eye fixations of humans. (a) Histogram of human answer accuracy (b) Center biases of the correct and incorrect attention.

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) How do questions affect human attention? (Section 4.1)
2) Do machines or humans look at places relevant to the reasoning process? How does the attention process influence task performances? (Section 4.2)

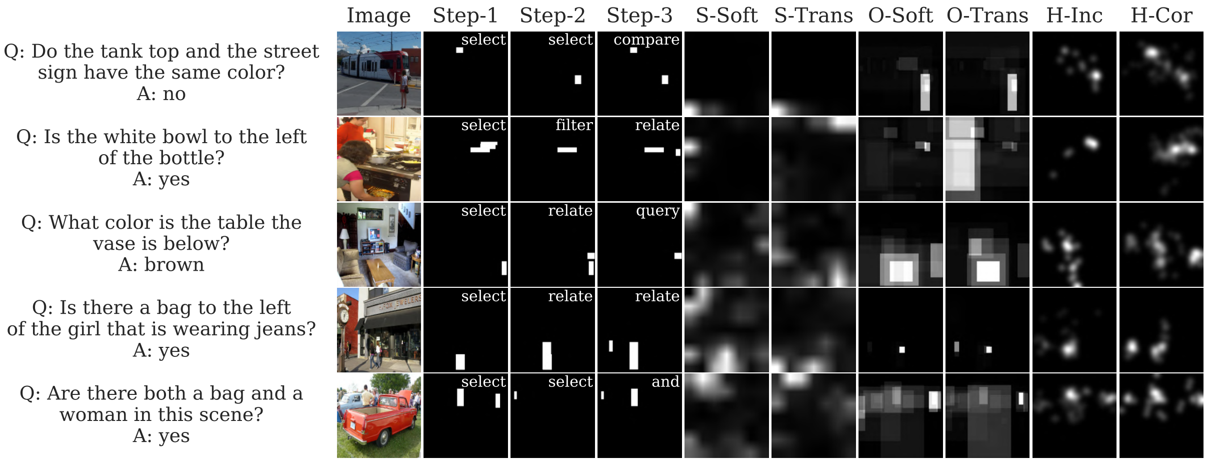


Image
Step-1
Step-2
Step-3
S-Soft
S-Trans
O-Soft
O-Trans
H-Inc
H-Cor
Q: Do the tank top and the street sign have the same color?
A: no
select
select
compare
Q: Is the white bowl to the left of the bottle?
A: yes
select
filter
relate
Q: What color is the table the vase is below?
A: brown
select
relate
query
Q: Is there a bag to the left of the girl that is wearing jeans?
A: yes
select
relate
relate
Q: Are there both a bag and a woman in this scene?
A: yes
select
select
and

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

3) How does attention accuracy evolve, and what about its correlation with the reasoning process? (Section 4.3)
4) Do machines and humans with diverse answers look differently? (Section 4.4)
5) Does progressive attention supervision improve attention and task performance? (Section 4.5)
6) Is incorporating the incorrect attention beneficial for attention learning? (Section 4.6)
7) Do attention accuracy and reasoning performance agree? (Section 4.7)

4.1 How Do Questions Affect Human Attention?

Human attention is driven by both the bottom-up visual stimuli and the top-down task information (*e.g.*, question in the VQA task). Our AiR-D dataset has 299 images with at least two corresponding questions. To study how questions affect human attention in the VQA task, on this subset of eye-tracking data, we study the agreement between the fixation maps when answering two different questions about the same image.

Our experiments consider three distinct aspects of human attention (*i.e.*, temporal dynamics, spatial alignment, and semantic alignment). For temporal dynamics, we group the fixations into three temporal bins (0-1s, 1-2s, and 2-3s), and compare the fixation maps for each bin. For spatial alignment, we compare fixation maps using Spearman's Rank correlation r to measure the similarity of fixation distributions. For semantic alignment, we measure the average attention value in each object category of the scene, compare the top-5 object categories with the highest attention values, and evaluate the proportion of overlapping categories with Intersection over Union (IoU).

Table 3 shows the spatial and semantic alignments of attention between different questions and their temporal evolutions. Two key observations can be drawn from the results: (1) There exists a considerable overlap between human attentions when answering different questions on the same image. This is validated by the relatively high spatial and semantic alignment scores (*i.e.*, 0.6) of their overall attention. (2) Both the spatial and semantic alignments decrease monotonically over time, suggesting that the question information progressively affects attention. At the beginning of visual exploration, people answering different questions focus on similar regions to quickly and broadly understand the image. After that, they gradually shift their attention towards the ROIs specific to each question, which results in low alignments between their attention patterns.

	Spatial	Semantic
Aggregated	0.709	0.605
0-1s	0.678	0.675
1-2s	0.536	0.590
2-3s	0.439	0.564

TABLE 3: Spatial and semantic alignment scores between aggregated attention and attention over time.

These observations show that human eye fixations have generally strong agreements when looking at the same image, even when answering different questions. However, the question information affects human attention in a dynamic manner, as the spatial and semantic agreements between attention patterns in different questions decrease monotonically over time.

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

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

We evaluate four types of attention that are commonly used in VQA models, including spatial soft attention (*i.e.*, S-Soft), spatial Transformer attention (*i.e.*, S-Trans), object-based soft attention (*i.e.*, O-Soft), and object-based Transformer attention (*i.e.*, O-Trans). Spatial and object-based

	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 4: Quantitative evaluation of AiR-E scores and task performance.

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**

TABLE 5: Pearson's r between attention accuracy (AiR-E) and task performance. Bold numbers indicate significant positive correlations ($p<0.05$).

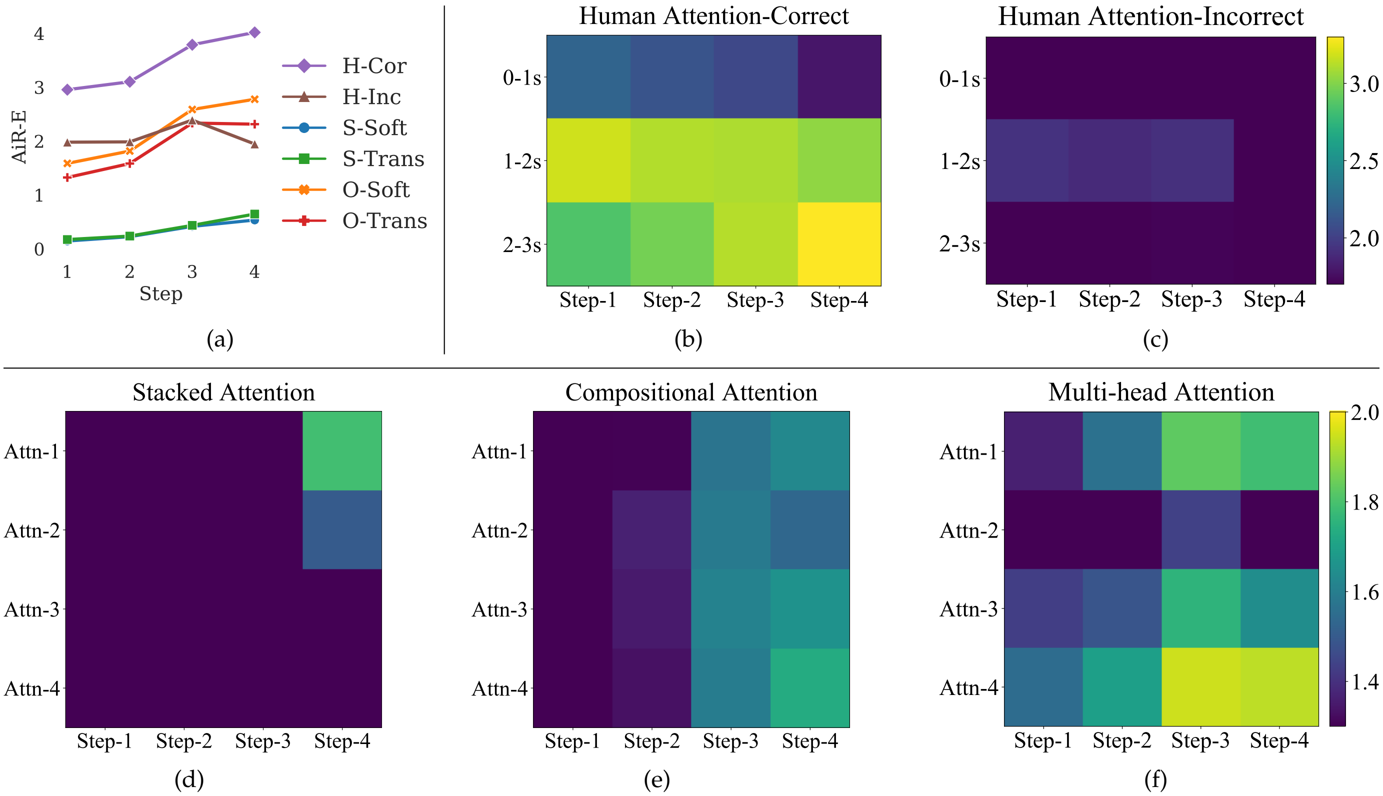
attention differ in terms of their inputs (*i.e.,* image features or regional features), while soft and Transformer attention methods differ in terms of the computational methods of attention (*i.e.,* with convolutional layers or matrix multiplication). We use spatial features extracted from ResNet-101 [88] and object-based features from [5] as the two types of inputs, and follow the implementations of [5] and [89] for the soft attention [90] and Transformer attention [91] computation, respectively. We integrate the aforementioned attention mechanisms 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 [5] for illustration (results for the other backbones are provided in the supplementary materials). For human attention, we denote the fixation maps associated with correct and incorrect answers as H-Cor and H-Inc, and the overall fixation map regardless of correctness is denoted as H-Tot. Fig. 5 presents examples of ROIs for different reasoning operations and the compared attention maps.

Attention accuracy and task performance of humans and models. Table 4 quantitatively compares the AiR-E scores and VQA task performance across humans and models with different types of attention. 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 attention mechanisms attend much more accurately than spatial attention. The low AiR-E of spatial attention confirms the previous conclusion drawn from the VQA-HAT dataset [12]. 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 attention, the soft attention slightly outperforms its Transformer counterpart. Since the Transformer attention explicitly learns 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 [91], they do not perform as well as soft attention. The ranks of different attention mechanisms 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 4 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 ability of the 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 Pearson's r between the attention accuracy and task performance across different questions.

As shown in Table 5, human attention accuracy and task performance are correlated for most of the operations (up



(a) (b) (c)

(d) (e) (f)

Fig. 6: Spatiotemporal accuracy of attention throughout the reasoning process. (a) shows the AiR-E of different reasoning steps for human aggregated attention and single-glimpse machine attention, (b)-(c) AiR-E scores for decomposed human attention with correct and incorrect answers, (d)-(f) AiR-E for multi-glimpse machine attention. For heat maps shown in (b)-(f), the x-axis denotes the different reasoning steps while the y-axis corresponds to the indices of attention maps.

to $r = 0.329$). The correlation is higher than most of the compared machine attention mechanisms, suggesting that humans' task performance is more consistent with their attention quality. In contrast, though commonly referred to as an interface for interpreting models' decisions [12], [14], [15], 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 attention mechanisms achieve higher attention-answer correlations.

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

4.3 How Does Attention Accuracy Evolve Throughout the Reasoning Process?

To complement our previous analysis on the spatial allocation of attention, we move forward to analyze the spatiotemporal alignment of attention. Specifically, we analyze the AiR-E scores according to the chronological order of reasoning operations. We show in Fig. 6a that the AiR-E scores peak at the 3rd or 4th steps, suggesting that human attention and machine attention 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 attention. Our analysis reveals the significant spatiotemporal discrepancy between human attention and machine attention.

Does human attention follow the reasoning process? First, to analyze the spatiotemporal deployment of human attention in visual reasoning, we conduct a time course analysis by grouping fixations into three temporal bins (*i.e.*, 0-1s, 1-2s, and 2-3s) and analyzing both the attention and the allocation of attention. We measure the accuracy of attention with AiR-E scores for each fixation map and reasoning step (see Fig. 6b-c), and study the allocation of attention by computing the Correlation Coefficient (CC) [81] between fixation maps and a center prior baseline created by placing a Gaussian ($\sigma = 15$) at the image center [92]. Our results show that humans start exploring the visual scene (*i.e.*, 0-1s) with relatively low attention accuracy because it takes time to understand the visual scene and locate the correct ROIs. Their attention is also more biased towards the central regions at the beginning because of the experimental setting that aligns the initial fixation with the image center, and the advantage of rapidly extracting the gist of the scene [93], *i.e.*, 0.47 CC score for the first second compared to 0.15 CC score for the latter periods. After the initial exploration, human attention shows improved accuracy across all reasoning steps (*i.e.*, 1-2s), and particularly focuses on the early-step ROIs. In the final steps (*i.e.*, 2-3s), depending on the correctness of the answers, human attention either shifts to the ROIs at later stages (*i.e.,* correct), or becomes less accurate with lowered AiR-E scores (*i.e.,* incorrect). Such observations suggest a high spatiotemporal alignment between human attention and the sequence of reasoning operations.

Does machine attention follow the reasoning process? Similarly, we evaluate multi-glimpse machine attention mechanisms. We compare the stacked attention from SAN [74], compositional attention from MAC [94] and the multi-head attention [69], [77], which all adopt object-based attention. Fig. 6d-f shows that multi-glimpse attention mechanisms do not evolve with the reasoning process. Stacked attention's first glimpse already attends to the ROIs at the 4th step, and the other glimpses contribute little to the attention accuracy. Compositional attention and multi-head attention consistently align best with the ROIs at the 3rd or 4th 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. In contrast, models tend to directly attend to the final ROIs, instead of shifting their attention progressively.

4.4 Do Machines and Humans with Diverse Answers Look at Input Images Differently?

Our previous analyses show various degrees of alignment between the attention, the task outcome, and the intermediate decision-making process. The results motivate us to further study the correlation between attention and task performance, and how different attention patterns lead to diverse answers. In this subsection, we conduct pairwise comparisons between humans or VQA models, and organize the questions into different groups based on the correctness of the two answers.

Specifically, for each pair of models/humans, questions in the AiR-D dataset can fall into three distinct groups: questions where both humans/models answer them correctly (Correct) or incorrectly (Incorrect), or those where only one human/model answers correctly (Inter).

For the comparison of human attention, we follow [27] and measure the alignment between gaze sequences using the edit distance on real sequence (EDR) [95], and use the AUC-Judd [87] to measure the inter-subject consistency in gaze distributions. Lower EDR and higher AUC-Judd measures suggest more consistent attention. Our experiments suggest certain agreements between the correct and incorrect human attention (*i.e.,* EDR=0.641, AUC-Judd=0.872). The inter-subject agreement within the correct attention group is high (*i.e.,* EDR=0.592, AUC-Judd=0.895) while that with the incorrect attention group is relatively low (*i.e.,* EDR=0.635, AUC-Judd=0.864).

For the comparison of machine attention, for each question, we measure the Spearman's Rank correlation r between attentions corresponding to the two models. Table 6 reports the results for machine attention. We choose UpDown and MUTAN with soft attention (S) and Transformer attention (T) as the backbone models, as they have comparable VQA accuracy on the GQA validation set. All models adopt the object-based attention.

Three observations can be made from the experimental results: (1) Both humans and models have higher diversity of attention if their answers are different. Compared to the alignment scores for both the correct attention and the incorrect attention, the inter-correctness alignment scores are consistently lower. (2) Humans tend to converge on similar ROIs to answer questions, while machines tend to have more diverse focuses, depending on both the backbone models and attention types. This is validated by the high variance of attention alignment scores across different models being compared. (3) Compared with humans, models are more vulnerable to the most salient distractors, as they have higher alignment scores for incorrect attention.

	Inter	Correct	Incorrect
UpDown (S) - MUTAN (S)	0.569	0.610	0.698
UpDown (T) - MUTAN (T)	0.308	0.460	0.546
UpDown (S) - UpDown (T)	0.440	0.575	0.634
MUTAN (S) - MUTAN (T)	0.397	0.444	0.528
UpDown (S) - MUTAN (T)	0.440	0.475	0.556
UpDown (T) - MUTAN (S)	0.422	0.523	0.602

TABLE 6: Spearman's Rank Correlation between machine attention mechanisms for different answers. For each group separated by the horizontal lines, from top to bottom are results on different VQA backbones but the same type of attention, the same backbone but with different types of attention, and different backbones and attention mechanisms.

The aforementioned observations reveal the visual behaviors of humans and machines when deriving different answers. More importantly, it shows that, unlike humans, models are vulnerable to similar hard-negative distractors when answering a question, suggesting the usefulness of incorporating the negative attention to encourage models to avoid these distractors.

4.5 Does Progressive Attention Supervision Improve Attention and Task Performance?

Experiments in Section 4.2 and Section 4.3 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 throughout the reasoning process by jointly optimizing attention, reasoning operations, and task performance (*i.e.,* AiR-M, Section 3.3). Here we investigate the effectiveness of the AiR-M supervision method on three VQA models, *i.e.,* UpDown [5], MUTAN [96], and BAN [71]. We compare AiR-M with a number of state-of-the-art attention supervision methods, including supervision with human-like attention (HAN) [17], attention supervision mining (ASM) [18] and adversarial learning (PAAN) [16]. 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 [71]).

As shown in Table 7, the proposed AiR-M method significantly improves the performance of all baselines and consistently outperforms the other attention supervision methods. Two of the compared methods, HAN and PAAN,

	UpDown [5]		MUTAN [96]		BAN [71]	
	dev	standard	dev	standard	dev	standard
w/o Supervision	51.31	52.31	50.78	51.16	50.14	50.38
PAAN [16]	48.03	48.92	46.40	47.22	n/a	n/a
HAN [17]	49.96	50.58	48.76	48.99	n/a	n/a
ASM [18]	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**

TABLE 7: Comparative results on GQA test sets (test-dev and test-standard). All the compared results are from single models trained on the balanced training set of GQA.

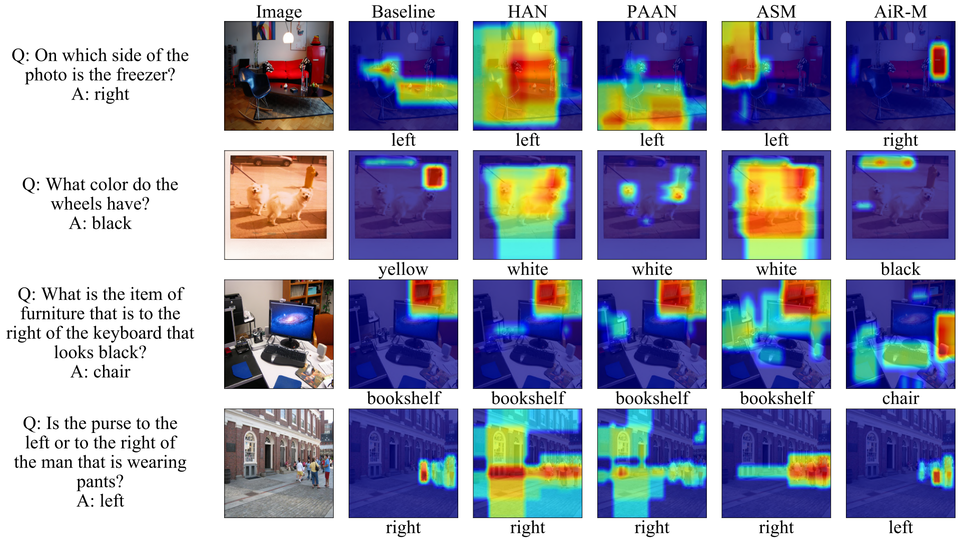


Image
Baseline
HAN
PAAN
ASM
AiR-M
Q: On which side of the photo is the freezer?
A: right
left
left
left
left
right
Q: What color do the wheels have?
A: black
yellow
white
white
white
black
Q: What is the item of furniture that is to the right of the keyboard that looks black?
A: chair
bookshelf
bookshelf
bookshelf
bookshelf
chair
Q: Is the purse to the left or to the right of the man that is wearing pants?
A: left
right
right
right
right
left

Fig. 7: Qualitative comparison between attention supervision methods, where Baseline refers to UpDown [5]. 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 each attention mechanism are shown below its respective attention map.

Method	GQA test-dev
AiR-M w/o L_α	50.01
AiR-M w/o L_r	50.33
AiR-M Single	52.84
AiR-M	**53.46**

TABLE 8: Experimental results of AiR-M under different supervision strategies. All reported results are on the GQA [13] test-dev set. Bold numbers indicate the best performance.

fail to improve the performance of object-based attention. Supervising attention with knowledge from objects mined from language, ASM [18] can consistently improve the performance of models. However, without considering the intermediate steps of reasoning, it is not as effective as the proposed method. In addition to the enhanced VQA performance, our method also predicts the reasoning steps with high accuracy (96.2% validation accuracy on reasoning step prediction). It shows that our method can accurately capture the correct reasoning process and learn reasoning-aware attention to improve the performance of visual reasoning.

Fig. 7 shows the qualitative comparison between supervision methods. As the previous supervision methods (*i.e.*, HAN, PAAN and ASM) are optimized to simultaneously capture all important regions, their attention outputs tend to spread over multiple ROIs, in which some are less relevant. On the contrary, by progressively supervising the attention throughout the reasoning process, our proposed AiR-M learns focused attention towards the most relevant ROIs (*i.e.*, freezer, wheel, chair, purse). Moreover, unlike reasoning-agnostic methods that commonly ignore ROIs for intermediate decision-making steps (*i.e.*, keyboard, man), our method can capture diverse ROIs with regard to the entire reasoning process.

Our method jointly and progressively optimizes attention and reasoning operations. To further demonstrate its advantages, we compare it with three alternatives: two models trained with either attention ground truth or reasoning operations (AiR-M w/o L_r and AiR-M w/o L_α), and a single-glimpse model jointly optimized on both types of ground truth (AiR-M Single, the attention is supervised with ground truth aggregated across all reasoning steps). Three observations can be drawn from the results in Table 8: (1) Due to the lack of knowledge about the correlation between attention and reasoning process, individually optimizing the reasoning semantics or fine-grained grounding fails to improve the performance; (2) Jointly optimizing attention and the reasoning process with the same types of ground truth (AiR-M Single) leads to significant improvements, demonstrating the need of learning reasoning-aware attention; and (3) Compared to single-step attention optimization (AiR-M Single), AiR-M with multi-step progressive supervision can learn more fine-grained attention specific to each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2021.3114582, IEEE Transactions on Pattern Analysis and Machine Intelligence

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 [5]	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

TABLE 9: AiR-E scores of the supervised attention mechanisms.

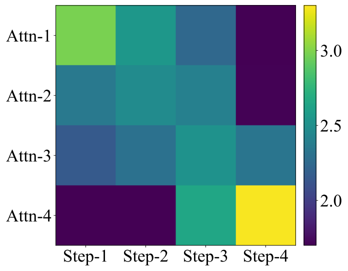


Fig. 8: Alignment between the attention and reasoning process supervised with the AiR-M method.

reasoning step, resulting in better performance.

To further demonstrate the impact of our AiR-M method on the attention accuracy, Table 9 reports the AiR-E scores across different operations. It shows that the AiR-M supervision method significantly improves attention accuracy (attention aggregated 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 attention better according to their chronological order in the reasoning process (see Fig. 8 and the supplementary video), showing progressive improvement of attention throughout the entire process.

4.6 Does Incorporating the Incorrect Attention Benefit Attention Learning?

Analyses in Section 4.4 show that VQA models tend to predict wrong answers because of hard-negative distractors. In this subsection, we investigate if explicitly supervising models with both correct and incorrect attention (AiR-C) can help them avoid such hard-negative distractors and improve answer accuracy. We utilize UpDown [5] as our backbone model, and conduct experiments on the GQA dataset by supervising the machine attention with the correct and incorrect attention mined from the annotations. Further, to demonstrate its generalizability, we experiment our method on the IQVA [27] dataset with eye-tracking data on 360° videos. Following [97], we decompose the 360° visual frames into perspective cubemaps, and apply the UpDown [5] backbone on each cubemap. Features from different cubemaps and time steps are combined with trainable attention to derive the final answer. The new model (UpDown-360) is first pre-trained on the GQA dataset, and then fine-tuned on the training set of IQVA.

Table 10 shows quantitative results of AiR-C compared with two alternatives (*i.e.,* w/o Supervision and supervision with Correct answers). By incorporating both correct and incorrect attention, our AiR-C method can outperform these two counterparts, suggesting that avoiding hard-negative distractors is complementary to the supervision from the correct attention. The improvement brought by the incorrect attention supervision is consistent across different datasets. In addition to the increase in VQA performance, the improvement of attention accuracy (*i.e.,* AiR-E) is also significant. Compared to model supervised by only the correct attention (AiR-E=1.74), our AiR-C method can alleviate the distraction from visually salient yet question-irrelevant regions and achieve much higher attention accuracy (AiR-E=2.02).

	UpDown [5]		UpDown-360
	GQA-dev	GQA-standard	IQVA
w/o Supervision	51.31	52.31	39.73
Correct	52.96	53.57	40.55
AiR-C	**53.74**	**53.85**	**41.10**

TABLE 10: VQA accuracy on the GQA test sets (test-dev and test-standard) and IQVA test set. AiR-C denotes our full method incorporating both the correct attention and the incorrect attention.

Qualitatively, Fig. 9 shows that supervising the models (*i.e.,* UpDown and UpDown-360) with incorrect attention leads to focused attention on the correct ROIs. In the 1st and 2nd examples (perspective images from GQA), other models (*i.e.,* w/o Supervision and Correct) either are distracted by the dominant objects (*i.e.,* cabinet and bathtub), or fail to focus on the correct ROIs (*i.e.,* toilet, towel, and vase), while our AiR-C method helps the model avoid these distractors and focus on the most relevant ROIs to generate correct answers. In the 3rd and 4th examples (*i.e.,* 360° video frames from IQVA), without knowledge of the visual distractors, other models do not have a clear focus due to the complexity of scenes, while our method develops focused attention.

Theses results demonstrate the effectiveness of incorporating knowledge from hard-negative distractors and suggest the generalizability of the proposed AiR-C method.

4.7 Do Attention Accuracy and Reasoning Performance Agree?

Our analyses in the previous sections demonstrate the positive correlation between attention accuracy and reasoning performance (Section 4.2), and show that learning more accurate attention leads to a considerable improvement in reasoning performance (Section 4.5 and Section 4.6). To further analyze the impacts of attention on visual reasoning, we conduct an ablation study by replacing the model

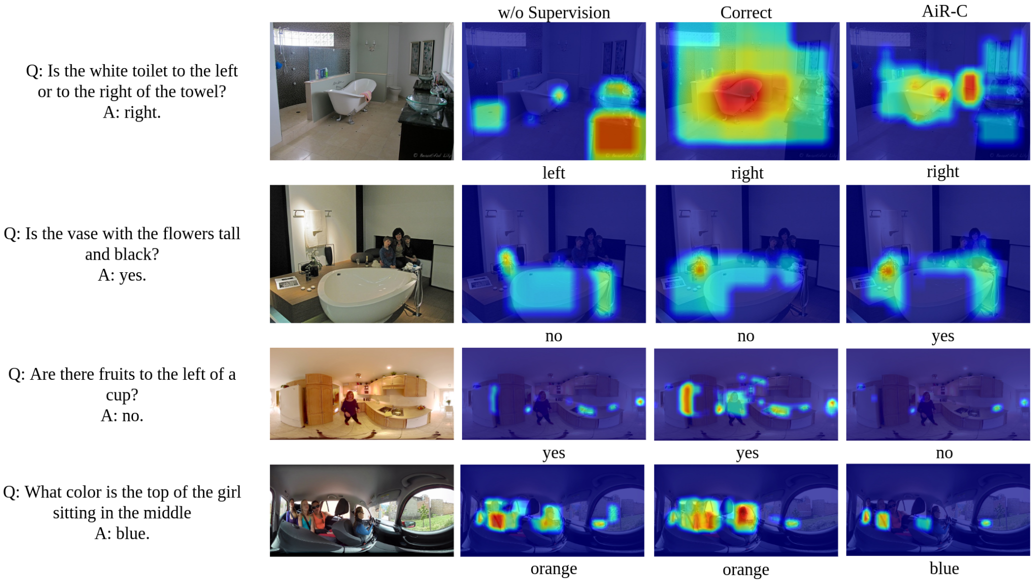

w/o Supervision
Correct
AiR-C
Q: Is the white toilet to the left or to the right of the towel?
A: right.
left
right
right
Q: Is the vase with the flowers tall and black?
A: yes.
no
no
yes
Q: Are there fruits to the left of a cup?
A: no.
yes
yes
no
Q: What color is the top of the girl sitting in the middle
A: blue.
orange
orange
blue

Fig. 9: Qualitative results for attention supervision with incorrect attention. For sample show in each row, from left to right are input image with ground truth question and answer, model attention learned without supervision, model attention learned with our correct attention [18], model attention learned with our AiR-C method. The model predicted answers are shown at the bottom.

attention outputs with two extreme types of attention: random attention and ground-truth attention. Therefore, the reasoning accuracy based on the random attention can be seen as the performance lower bound, while the reasoning accuracy based on the ground-truth attention can be seen as the performance upper bound. Specifically, to evaluate the performance lower bound, we replace the attention computed from the pre-trained UpDown [5] model with randomly sampled attention maps following a uniform distribution. Similarly, by replacing the model's attention with the ground-truth attention, we can evaluate its performance upper bound. With this experiment, we find that the random attention leads to a significant drop in the answer accuracy (-7.39%) over the pre-trained baseline, while the ground-truth attention improves the answer accuracy by a large margin (+8.00%). These performance bounds suggest the significant role of attention in visual reasoning.

However, attention is not the only important factor for achieving high reasoning accuracy. For instance, visual recognition is also consequential. To correctly answer a question, even with correct attention, one must recognize the attributes of the attended objects and the relationship between them. As a result, there are cases where attention accuracy does not agree with the reasoning performance. Fig. 10 shows typical cases where the attention accuracy and reasoning performance are inconsistent. These cases include (1) when the model answers correctly but with wrong attention (*i.e.*, AiR-E < 1), and (2) when the model answers incorrectly but with reasonable attention (*i.e.*, AiR-E > 2.5). We use our AiR-M and AiR-C methods for demonstration due to their high attention accuracy.

In some cases, the model answers correctly with incorrect attention, which is resulted from various reasons: (1) **Biased prior distribution of questions and answers**. Language biases are prevalent in VQA datasets due to the imbalanced prior distributions of questions and answers. For example, paper bowls are less common, as shown in the 1st example of Fig. 10. Models leveraging such biases can predict the correct answers without attending to the correct ROIs. (2) **Attending to wrong objects that coincidentally relate to the answers**. Many images contain abundant objects that may share similar characteristics. As a result, even with incorrect attention, models can still answer correctly by coincidentally looking at another object that relates to the answer. *E.g.*, looking at boys not wearing glasses also leads to the correct answer, as shown in the 2nd example of Fig. 10. (3) **Capturing the ROIs without focused attention**. For scenes cluttered with various semantics, models may not focus on the correct ROIs. *E.g.*, boxes with bright colors attract more attention than the shelf behind them, as shown in the 3rd example of Fig. 10. However, since the features of the ROIs can be extracted without strong attention, they can still answer correctly despite the low attention accuracy.

There are also cases where the model answers incorrectly but with reasonable attention: (1) **Missing the ROIs directly related to the answers**. Many questions in our dataset require reasoning over multiple ROIs, even if models focus on most of the ROIs, *E.g.*, as shown in the 4th example of Fig. 10, AiR-M looks at people with the bag but not the van, while AiR-C looks at the van but not the people. They both answer incorrectly due to the failure of capturing both ROIs. (2) **Failing to recognize the ROIs**. Some of the ROIs could be small and difficult to recognize, so models looking at the correct ROIs can still answer incorrectly. *E.g.*, as shown in the 5th example of Fig. 10), models fail to describe the policeman due to erroneous recognition.

In sum, these results suggest that the attention accuracy strongly correlates with the reasoning performance in general, but answer correctness is not completely dependent on the accuracy of attention.

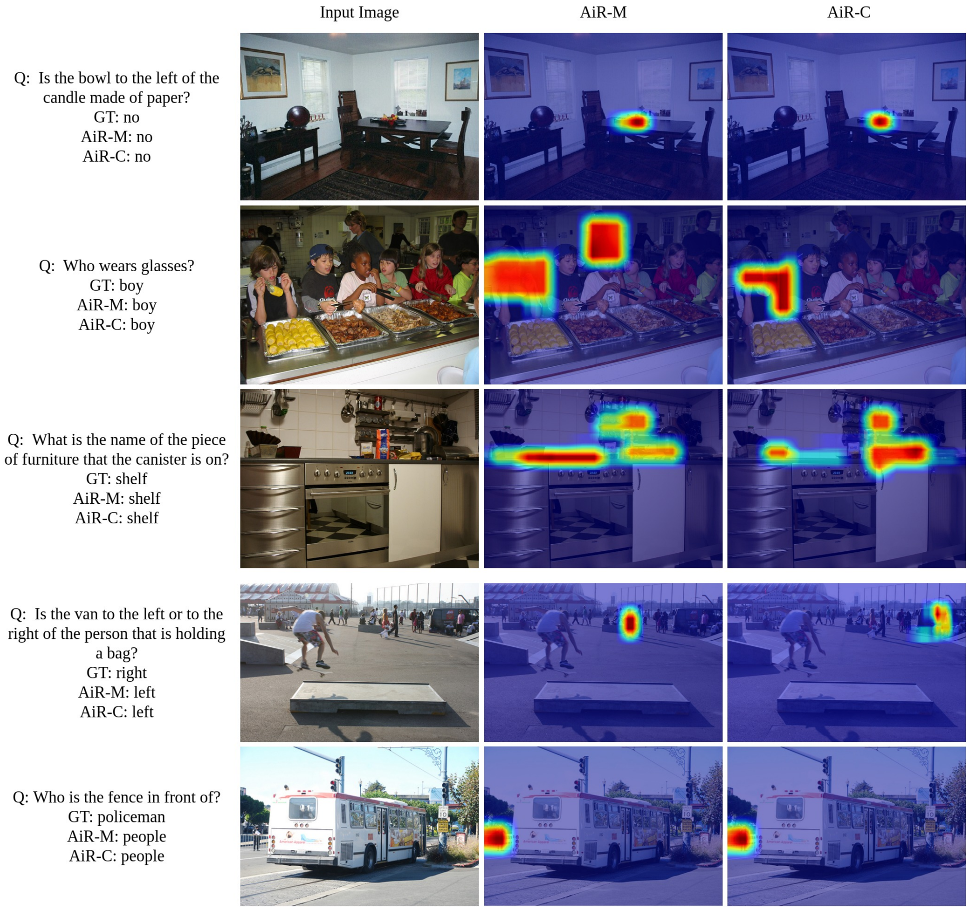

Input Image
AiR-M
AiR-C
Q: Is the bowl to the left of the candle made of paper?
GT: no
AiR-M: no
AiR-C: no
Q: Who wears glasses?
GT: boy
AiR-M: boy
AiR-C: boy
Q: What is the name of the piece of furniture that the canister is on?
GT: shelf
AiR-M: shelf
AiR-C: shelf
Q: Is the van to the left or to the right of the person that is holding a bag?
GT: right
AiR-M: left
AiR-C: left
Q: Who is the fence in front of?
GT: policeman
AiR-M: people
AiR-C: people

Fig. 10: Examples for studying the inconsistency between attention accuracy and reasoning performance. From left to right are questions with ground truth (GT) and predicted answers, input images, and attention maps for the two models.

5 Conclusion

We introduce AiR, a novel framework with a quantitative evaluation metric (AiR-E), two supervision methods (AiR-M and AiR-C), and an eye-tracking dataset (AiR-D) for understanding and improving attention in the reasoning context. Our experiments analyze the correlation between attention and task performance in various aspects, and highlight the significant gap between machines and humans on the alignment of attention and reasoning process. With the newly proposed supervision methods, we show that accurate attention deployment can lead to improved task performance, which is related to both the task outcome and the intermediate reasoning steps. We hope that this work will be helpful for the future development of visual attention and reasoning method, and inspire the analysis of model interpretability throughout the reasoning process.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *TPAMI*, vol. 42, no. 8, pp. 2011–2023, 2020.

[2] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CC-Net: Criss-cross attention for semantic segmentation," in *ICCV*, 2019, pp. 603–612.

[3] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *CVPR*, 2019, pp. 10 697–10 706.

[4] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018.

[6] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *CVPR*, 2019, pp. 1448–1457.

[7] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical lstms with adaptive attention for visual captioning," *TPAMI*, vol. 42, no. 5, pp. 1112–1131, 2020.

[8] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4633–4642.

[9] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *CVPR*, 2017, pp. 6298–6306.

[10] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault, "Human attention in image captioning: Dataset and analysis," in *ICCV*, 2019.

[11] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *ICCV*, 2017.

[12] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *EMNLP*, 2016.

[13] D. A. Hudson and C. D. Manning, "GQA: a new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019.

[14] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach,

B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *CVPR*, 2018.

[15] W. Li, Z. Yuan, X. Fang, and C. Wang, "Knowing where to look? analysis on attention of visual question answering system," in *ECCV Workshops*, 2018.

[16] B. N. Patro, Anupriy, and V. P. Namboodiri, "Explanation vs attention: A two-player game to obtain attention for VQA," in *AAAI*, 2020.

[17] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," in *AAAI*, 2018.

[18] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *WACV*, 2019, pp. 349–357.

[19] S. Chen, M. Jiang, J. Yang, and Q. Zhao, "AiR: Attention with reasoning capability," in *ECCV*, 2020.

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*. IEEE, 2009, pp. 2106–2113.

[21] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *CVPR*, 2018, pp. 7521–7531.

[22] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *TPAMI*, 2019.

[23] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, "Static saliency vs. dynamic saliency: A comparative study," in *ACM MM*, 2013, p. 987–996.

[24] C. Shen and Q. Zhao, "Webpage saliency," in *ECCV*. Springer, 2014, pp. 33–46.

[25] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *ECCV*, 2018, pp. 619–635.

[26] A. Palazzi, D. Abati, F. Solera, R. Cucchiara *et al.*, "Predicting the driver's focus of attention: the dr (eye) ve project," *TPAMI*, vol. 41, no. 7, pp. 1720–1733, 2018.

[27] M. Jiang, S. Chen, J. Yang, and Q. Zhao, "Fantastic answers and where to find them: Immersive question-directed visual attention," in *CVPR*, 2020, pp. 2980–2989.

[28] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *TPAMI*, vol. 35, no. 1, pp. 185–207, 2012.

[29] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NeurIPS*, 2005, pp. 155–162.

[30] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *NeurIPS*, vol. 20, pp. 1–7, 2008.

[31] A. Bovik, L. Cormack, I. Van Der Linde, and U. Rajashekar, "Doves: a database of visual eye movements," *Spatial Vision*, vol. 22, no. 2, pp. 161–177, 2009.

[32] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *ECCV*. Springer, 2010, pp. 30–43.

[33] G. Kootstra, B. de Boer, and L. R. Schomaker, "Predicting eye fixations on complex visual stimuli using local symmetry," *Cognitive Computation*, vol. 3, no. 1, pp. 223–240, 2011.

[34] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1, pp. 28–28, 2014.

[35] M. Jiang, J. Xu, and Q. Zhao, "Saliency in crowd," in *ECCV*. Springer, 2014, pp. 17–32.

[36] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *CVPR*, 2015, pp. 1072–1080.

[37] A. Borji and L. Itti, "CAT2000: a large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[38] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Visual Cognition*, vol. 17, no. 6-7, pp. 945–978, 2009.

[39] S. O. Gilani, R. Subramanian, Y. Yan, D. Melcher, N. Sebe, and S. Winkler, "Pet: An eye-tracking dataset for animal-centric pascal object classes," in *ICME*. IEEE, 2015, pp. 1–6.

[40] G. Zelinsky, Z. Yang, L. Huang, Y. Chen, S. Ahn, Z. Wei, H. Adeli, D. Samaras, and M. Hoai, "Benchmarking gaze prediction for categorical visual search," in *CVPR Workshops*, 2019, pp. 0–0.

[41] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai, "Predicting goal-directed human attention using inverse reinforcement learning," in *CVPR*, 2020, pp. 193–202.

[42] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., 2012, pp. 101–115.

[43] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *TPAMI*, vol. 37, no. 7, pp. 1408–1424, 2014.

[44] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *TIP*, vol. 28, no. 8, pp. 3703–3713, 2019.

[45] Q. Zheng, J. Jiao, Y. Cao, and R. W. Lau, "Task-driven webpage saliency," in *ECCV*, 2018, pp. 287–302.

[46] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. Muller, J. Whritner, L. Zhang, M. Hayhoe, and D. Ballard, "Atari-head: Atari human eye-tracking and demonstration dataset," in *AAAI*, vol. 34, no. 04, 2020, pp. 6811–6820.

[47] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *TPAMI*, 2019.

[48] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," in *ACM MM*, 2018, p. 86–110.

[49] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *TPAMI*, vol. 24, no. 3, pp. 420–425, 2002.

[50] S. Han and N. Vasconcelos, "Biologically plausible saliency mechanisms improve feedforward object recognition," *Vision Research*, vol. 50, no. 22, pp. 2295–2307, 2010.

[51] A. Borji and L. Itti, "Scene classification with a sparse set of salient regions," in *ICRA*. IEEE, 2011, pp. 1902–1908.

[52] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *TPAMI*, vol. 42, no. 8, pp. 1913–1927, 2019.

[53] W. Wang, J. Shen, X. Lu, S. C. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *TPAMI*, 2020.

[54] J. Xu, L. Mukherjee, Y. Li, J. Warner, J. M. Rehg, and V. Singh, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *CVPR*, 2015, pp. 2235–2244.

[55] C.-J. Yang, K. Grauman, and D. Gurari, "Visual question answer diversity," in *HCOMP*, 2018.

[56] H. R. Tavakoli, F. Ahmed, A. Borji, and J. Laaksonen, "Saliency revisited: Analysis of mouse movements versus fixations," in *CVPR*, 2017.

[57] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *ICCV*, 2015.

[58] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *CVPR*, 2017.

[59] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017.

[60] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: understanding stories in movies through question-answering," in *CVPR*, 2016.

[61] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, 2019.

[62] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *CVPR*, 2019, pp. 3190–3199.

[63] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, "Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension," in *CVPR*, 2017, pp. 5376–5384.

[64] X. Wang, Y. Liu, C. Shen, C. C. Ng, C. Luo, L. Jin, C. S. Chan, A. v. d. Hengel, and L. Wang, "On the general value of evidence, and bilingual scene-text visual question answering," in *CVPR*, 2020, pp. 10 123–10 132.

[65] N. Garcia, M. Otani, C. Chu, and Y. Nakashima, "Knowit VQA: Answering knowledge-based questions about videos," *AAAI*, vol. 34, no. 07, pp. 10 826–10 834, 2020.

[66] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "TVQA: Localized, compositional video question answering," in *EMNLP*, 2018.

[67] I. Ilievski and J. Feng, "Multimodal learning and reasoning for visual question answering," in *NeurIPS*, vol. 30, 2017, pp. 551–562.

[68] T. Do, T.-T. Do, H. Tran, E. Tjiputra, and Q. D. Tran, "Compact trilinear interaction for visual question answering," in *ICCV*, 2019.

[69] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual

question answering and visual grounding," in *EMNLP*, 2016, pp. 457–468.

[70] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *ICCV*, 2017.

[71] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018, pp. 1571–1581.

[72] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *CVPR*, 2018.

[73] J. Wu and R. Mooney, "Self-critical reasoning for robust visual question answering," in *NeurIPS*, 2019.

[74] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *CVPR*, 2016.

[75] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *NeurIPS*, 2018, pp. 1031–1042.

[76] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019.

[77] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017.

[78] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *ICCV*, 2019.

[79] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *CVPR*, 2020, pp. 10 797–10 806.

[80] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *ICCV*, 2019, pp. 10 312–10 321.

[81] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *TPAMI*, 2019.

[82] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.

[83] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[84] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[85] S. D. König and E. A. Buffalo, "A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds," *Journal of Neuroscience Methods*, vol. 227, pp. 121 – 131, 2014.

[86] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014, pp. 280–287.

[87] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," 2015.

[88] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[89] P. Gao, Z. Jiang, H. You, P. Lu, S. C. H. Hoi, X. Wang, and H. Li, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *CVPR*, 2019.

[90] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[91] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[92] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *ICCV*, 2013, pp. 921–928.

[93] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 4–4, 11 2007.

[94] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," 2018.

[95] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *SIGMOD*, 2005, pp. 491–502.

[96] H. Ben-Younes, R. Cadène, N. Thome, and M. Cord, "MUTAN: Multimodal tucker fusion for visual question answering," *ICCV*, 2017.

[97] S.-H. Chou, W.-L. Chao, W.-S. Lai, M. Sun, and M.-H. Yang, "Visual question answering on 360○ images," *WACV*, 2020.

Shi Chen received the BE degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2015, and the MS degree from the University of Minnesota, Minneapolis, in 2017. He is currently working toward the PhD degree in the Department of Computer Science, University of Minnesota. His research interests include computer vision, pattern recognition, and deep learning.

Ming Jiang is a postdoctoral researcher at the Department of Computer Science and Engineering, University of Minnesota. He obtained his Ph.D. degree in Electrical and Computer Engineering from the National University of Singapore. His M.E and B.E degrees were received from Zhejiang University, Hangzhou, China. His research interests cover computer vision, cognitive vision, machine learning, psychophysics, and brain-machine interface.

Jinhui Yang is a Ph.D. student at the Department of Computer Science and Engineering, University of Minnesota. He graduated from Carleton College in 2019 with a BA degree in Computer Science and Statistics. His current research interests include computer vision, interpretable machine learning, and deep neural networks.

Qi Zhao is an assistant professor in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. Her main research interests include computer vision, machine learning, cognitive neuroscience, and healthcare. She received her Ph.D. in computer engineering from the University of California, Santa Cruz in 2009. She was a postdoctoral researcher in the Computation & Neural Systems, and Division of Biology at the California Institute of Technology from 2009 to 2011. Before joining the University of Minnesota, Qi was an assistant professor in the Department of Electrical and Computer Engineering and the Department of Ophthalmology at the National University of Singapore. She has published more than 80 journal and conference papers in top computer vision, machine learning, and cognitive neuroscience venues, and edited a book with Springer, titled Computational and Cognitive Neuroscience of Vision, that provides a systematic and comprehensive overview of vision from various perspectives. She serves as an Associate Editor of IEEE Transactions on Neural Networks and Learning Systems (TNNLS), as a program chair for IEEE Winter Conference on Applications of Computer Vision (WACV), and as an organizer and/or area chair for IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and other major venues in computer vision and AI. She is a member of the IEEE since 2004.