Fantastic Answers and Where to Find Them: Immersive Question-Directed Visual Attention

Ming Jiang[*] Shi Chen[*] Jinhui Yang Qi Zhao

University of Minnesota

{mjiang, chen4595, yang7004, qzhao}@umn.edu

Abstract

While most visual attention studies focus on bottom-up attention with restricted field-of-view, real-life situations are filled with embodied vision tasks. The role of attention is more significant in the latter due to the information overload, and attention to the most important regions is critical to the success of tasks. The effects of visual attention on task performance in this context have also been widely ignored. This research addresses a number of challenges to bridge this research gap, on both the data and model aspects.

*Specifically, we introduce the first dataset of top-down attention in immersive scenes. The Immersive Question-directed Visual Attention (IQVA) dataset features visual attention and corresponding task performance (*i.e.*, answer correctness). It consists of 975 questions and answers collected from people viewing 360° videos in a head-mounted display. Analyses of the data demonstrate a significant correlation between people's task performance and their eye movements, suggesting the role of attention in task performance. With that, a neural network is developed to encode the differences of correct and incorrect attention and jointly predict the two. The proposed attention model for the first time takes into account answer correctness, whose outputs naturally distinguish important regions from distractions. This study with new data and features may enable new tasks that leverage attention and answer correctness, and inspire new research that reveals the process behind decision making in performing various tasks.*

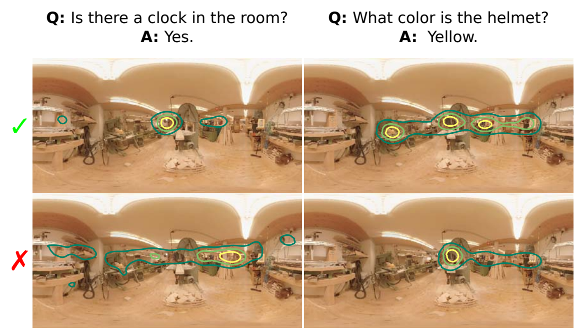

Figure 1: Visual attention is driven by tasks. The correct attention (row 1) provides essential information for answering the question, while the incorrect attention (row 2) helps identify the distracting features to be avoided when designing intelligent visual systems. Contours represent different fixation densities (0.25, 0.5, and 0.75), and brighter contours indicate higher fixation densities.

1. Introduction

Visual attention provides humans and machines with the ability to rapidly understand a scene by selectively processing the incoming information. Understanding the roles of attention is of significant importance for many applications. In the past decades, many eye-tracking datasets and attention prediction models have been developed to study attention in regular images and videos. Due to the limited field of view (FOV) and the passive viewing (PV) paradigm, however, these studies are difficult to be transferred to solve real-world problems. Furthermore, despite the popularity of aggregating all human attention patterns for attention modeling, the effects of different patterns on task performances have been mostly unstudied (see Figure 1 for an example). Such differences reveal important visual features to focus on or to avoid, providing insights for the understanding and modeling of attention for tasks of interest. To push forward the research frontier of visual attention, we aim at investigating two unstudied problems in computer vision: task-driven attention in immersive scenes, and the relationship between attention and task performance.

In this work, we introduce Immersive Question-directed Visual Attention (IQVA), a new dataset of eye-tracking data

[*]Equal contribution.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00305

collected from humans answering questions in immersive scenes. It consists of 975 questions on 360° video clips, each annotated with 14 answers (either correct or incorrect) and the corresponding eye-tracking data. Different from previous eye-tracking datasets, IQVA is built upon a more general and realistic paradigm where people actively explore the immersive scenes with time limits and answer questions. It highlights the importance of attention to the task outcomes and enables a fine-grained comparison between the attention patterns associated with different task performances. To the best of our knowledge, IQVA is the first attention dataset that explicitly verifies the correctness of ground-truth labels and differentiates between correct and incorrect ones. It demonstrates the significant impacts of attention on task performance, which can benefit the modeling of both human and machine vision systems. Based on the new dataset and analyses, we further introduce a novel attention model to predict the correct and incorrect attention maps with an emphasis on their differences. Considering the incorrect attention as a hard negative sample, we show that jointly predicting correct and incorrect attention can increase the accuracy of both. In sum, the main contributions of this work are three-fold:

First, we introduce and highlight a new research problem: Immersive Question-directed Visual Attention. To study this problem, we propose the IQVA dataset with an emphasis on the differences between attention patterns of correct and incorrect answers.

Second, with extensive data analyses, we demonstrate correlations between visual attention and task performance. People who answer correctly exhibit consistent attention patterns, while those who answer incorrectly are affected by diverse factors.

Finally, we propose a neural network model to jointly predict the correct and incorrect attentions. A semantic working memory and a fine-grained difference loss are proposed to model the top-down task guidance and to learn features that distinguish both attentions.

2. Related Work

Visual attention datasets. For decades, visual attention has been extensively studied in the fields of computer vision [4, 5, 23, 38] and cognitive vision [1, 32, 50]. Datasets have been built using eye-tracking [21, 48] or simulated alternatives [20] to facilitate the development of attention models [4, 5, 45, 50]. While much research has focused on the bottom-up attention driven by stimulus [4, 5, 49, 50], top-down attention driven by tasks is less studied [2, 23, 47]. Moreover, the highly controlled settings and the rectangular limited FOV in conventional image of video viewing prevent eye-tracking data from accurately representing human attention in everyday tasks. To collect attention data in a natural FOV, several works [10, 27, 28] use wearable eye-trackers to record attention in daily activities (*e.g.,* cooking), where people can move and act freely in the environment. Another line of research utilizes omnidirectional cameras and head-mounted displays (HMDs) to study how people explore virtual environments. Attention data in this type of immersive scenes are captured by tracking people's head movements [18, 44] or eye movements [17, 43]. While enabling the tracking of more natural gaze behaviors, existing datasets either have insufficient variability in scenes, or ignore the impact from top-down tasks. As a result, understanding and modeling task-driven attention remain an open challenge. To address these issues, our dataset places an emphasis on the variety of attention for question answering in immersive scenes, and the correctness of answers. The dataset enables the study of how people's attention is driven by tasks and subsequently determines task performance.

Human and machine attention in top-down tasks. Many computer vision models use model attention to prioritize information in vision tasks. Despite their widespread acceptance and contributions to task performance, model attention does not always agree with humans in where to look at given the same tasks [4, 8, 49]. For example, in visual question answering (VQA) [3, 15], where attention plays an important role, analysis [12] has shown a low correlation between model and human attention. Such misalignment may be caused by the dataset bias that directs the model attention to certain priors [15, 31, 40], or the insufficient correctness verification of the ground-truth annotations [22, 30]. In this work, we study human attention under general top-down tasks, such as counting objects, identifying object characteristics, or finding inter-object relationships. To reduce the data bias, we increase the task difficulty by asking more challenging questions and providing broad-FOV visual inputs (*i.e.,* immersive scenes). Thus, both humans and machines need to attend correctly in order to answer the questions. Furthermore, we explicitly verify the correctness of ground-truth answers, so the proposed dataset and model can provide insights into how correct and incorrect attentions affect the task performance.

3. Data Collection

In this section, we introduce the procedure of data collection and post-processing. Featuring task-driven attention in immersive viewing of 360° videos, our IQVA dataset contains a total of 975 video clips and eye-tracking data of 14 participants each. Table 1 compares IQVA with other related datasets. Our dataset will be publicly available.

3.1. Stimuli and Annotations

Our stimuli are 360° YouTube videos. We manually select 392 videos with a wide variety of 360° scenes and rich contexts. Most videos depict human activities such as touring, gathering, driving, and sports activities, while others

Dataset	Modality	Scenes	Scanpaths	Task	TPA
Corbillon *et al.* [11]	Head	5	0.3k	PV	✗
Wu *et al.* [46]	Head	9	0.4k	PV	✗
Lo *et al.* [29]	Head	10	0.5k	PV	✗
Nguyen&Yan [34]	Head	24	1k	PV	✗
David *et al.* [13]	Eye	19	1k	PV	✗
Sitzmann *et al.* [43]	Head/Eye	22^{I}	2k	PV	✗
Zhang *et al.* [51]	Eye	104	2k	PV	✗
Rai *et al.* [39]	Eye	98^{I}	4k	PV	✗
IQVA	Eye	975	14k	VQA	✓

Table 1: A comparison between IQVA and related immersive visual attention datasets. TPA: with task performance annotation. *I*: image datasets. PV: passive viewing.

	Questions	Answers
	What is the woman playing with	
	How many lions are there drinking	
	Is the white car passing by after the green motorcycle	
	What shape are the red sunglasses worn by the boy	
	Does the man in gray walk before the woman walks	
	What color is the leftmost bucket on the truck	
	Is there a plant in front of the black painting	
	Who puts the first chair under a table	
	What animals are swimming	
	How many different people have kicked the ball	

Table 2: Examples of questions and common words.

present animals or natural landscapes. All of the videos are in 4K equirectangular format (3840×1920 pixels) with various frame rates between 24 and 60 fps.

A total of 975 clips are cropped from these videos, where each clip is annotated with a question. The questions are proposed by the authors and two trained graduate assistents. All questions are reviewed by the first author to make sure they have little to no ambiguity, and be reasonably difficult (*i.e.*, an active observer can answer correctly given the time limit). The level of difficulty is determined by the time limit, complexity of the scene, number and size of related objects, *etc.*While the questions represent a wide variety of general tasks, to better structure the data collection and analyses, we group the questions into three categories: **query** (*e.g.,* 'What ...' and 'Who ...') , **count** (*e.g.,* 'How many ...'), and **verify** (*e.g.,* 'Is ...' and 'Does ...'). Many of the questions require exhaustive search, spatial and temporal reasoning, or fine-grained recognition. Depending on their requirement of attention and reasoning skills, the difficulty of each question is rated on a scale of 0 to 2. Table 2 presents examples of the questions, and common words used in questions and answers.

While the VQA datasets consider the most frequent answers from annotators to be correct, this hypothesis does not always hold true [16, 22, 30]. To differentiate correct and incorrect attentions, we annotate each question with a correct answer by exhaustively examining all videos with at least two authors. If the authors do not agree on the answers due to ambiguity, the questions are revised or deleted.

Figure 2 presents statistics of videos and questions, in-

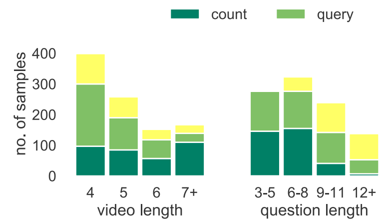

Figure 2: Distribution of data annotations regarding question type, video length (in seconds), question length (in words), and difficulty level.

cluding the length of video clips (4-15 seconds, 5.26±1.56), the length of questions (3-17 words, 7.94±2.98), and the difficulty level (0-2, 0.70±0.72). The three general question types make up 40.78% (query), 35.76% (count) and 23.46% (verify) of the data, respectively.

3.2. Eye Tracking

Apparatus. The 360° videos are displayed in an HMD (HTC VIVE Pro Eye, HTC, Valve corporation). This HMD allows sampling of scenes by approximately 110° horizontal FOV (2880×1600 pixels) at 90 frames per second. An integrated eye-tracker in the HMD samples gaze data at 120 Hz with a precision of 0.5°-1.1°. The experiment is running on a computer with an NVIDIA GTX 2070 GPU. A custom Unity3D (Unity Engine, CA, USA) scene is created to display the equirectangular videos in 360° and record the pixel coordinates of the eye-fixations.

Participants. A total of 18 males and 10 females, aged 19 to 38, participate in the eye-tracking experiment under the approval of the Institutional Review Board (IRB). All participants receive monetary compensation. The videos and questions are randomly grouped into 10 blocks for an one-hour session each. On average, each participant observes around 500 video clips and answers the corresponding questions. Each question is answered by 14 participants.

Procedure. The eye-tracker is 5-point calibrated before each session. The order of trials and the starting longitudinal position of each video are randomly initialized. Each trial begins with a question displayed on an empty background. Having completely understood the question, the participants push a controller button to start playing the corresponding video. All videos are played without sound. The participants actively explore the scenes and search for the correct answer. When the video ends, the question is displayed again. The participants either respond with their answer, or say "I don't know" to indicate a failure. The experimenter records the responses in a spreadsheet. Finally, the participants press another controller button to proceed to the next trial. To avoid HMD hazards (*e.g.,* dizziness, collision, falling), the participants or the experimenter can interrupt

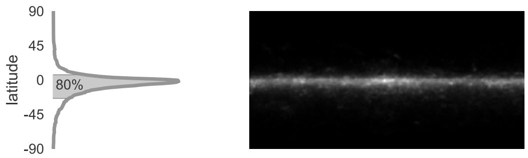

Figure 3: The average fixation map of the dataset demonstrates a skewed equator bias.

or terminate the experiment at any time.

3.3. Post-Experiment Processing

Answer verification. The authors review the responses from the participants, and compare them with the previously annotated answers. Since question ambiguity has been either reduced or eliminated, all responses can be classified to be either *correct* or *incorrect*. Cases where the participants fail to provide an answer are also classified to be *incorrect*.
Fixation map computation. The experiments produce a set of visual scanpaths for each video and question. A fixation map is generated for each video frame from raw gaze positions of all participants. The fixation map for a frame at time t is computed by accumulating gaze points in a temporal sliding window of 400 ms centered in t. The fixation maps are further smoothed using a spherical convolution with a Gaussian kernel (σ=9°) to obtain the final fixation maps $\{F_t\}$. For computational efficiency, we compute the maps at the reduced resolution of 256×128 following [52].

4. Data Analysis

In this section, we conduct and report statistical analyses to gather insights from the eye-tracking data and annotations. We present observations about human attention and VQA performances in immersive scenes.

4.1. Human gaze is biased towards the equator

Similar to previous literature in eye tracking that report different types of spatial bias [7, 35, 36, 42] in perceptive images or 360° scenes, we observe a strong equator bias in our data as shown in Figure 3. In terms of latitude, 95% of the gaze points are between -43° and 18.5°, and 80% are between -24° and 6.5°. This bias is jointly caused by the positioning of camera (*i.e.,* always in an upright position with the camera facing forward), the participants' motor bias (*i.e.,* turning around horizontally), as well as their viewing strategy (*i.e.,* expecting interesting objects to be placed near the ground). The downward skew is likely caused by the camera position, as the cameras are usually mounted at a relatively higher altitude (*e.g.,* on top of a car or a pole, *etc.*). Because of the random longitude initialization, no significant horizontal bias is introduced by the experiment. Further, by separating correct and incorrect attentions, we observe that their equator biases are highly similar (Pearson's r=0.95, p≈0), so the equator bias does not affect the task performance of humans.

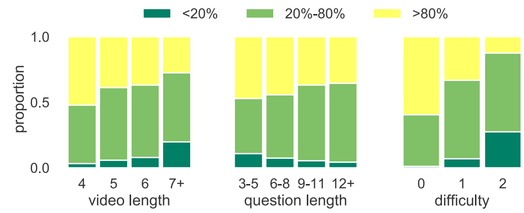

Figure 4: Distribution of human answer accuracy over different video lengths, question lengths, and difficulty levels.

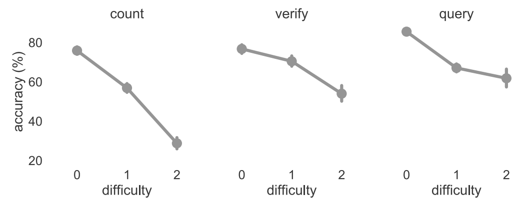

Figure 5: Human answer accuracy over different question types and difficulty levels. Error bars indicate the standard error of the mean.

4.2. Human answers have a broad range of accuracy

The overall accuracy of the participants' answers is 68.45%. Due to the unique characteristics of our questions and videos, only 15.78% of the questions have all correct answers, and 50.51% questions have an accuracy between 20% and 80%. As shown in Figure 4, the accuracy of the participants' answers decreases with increasing video length and difficulty, while the answers become more diverse as question length increases. Note that the video length is correlated with the difficulty by design (*i.e.,* more difficult questions have longer time limit). Figure 5 shows a decrease in accuracy with increased difficulty for different question types. In general, it is easier to correctly answer a **query** (77.03% accuracy) or **verify** question (69.84% accuracy) than a **count** question (57.76% accuracy). This may be because both **query** and **verify** questions require fewer targets to be observed, and targets tend to be provided with additional descriptions (*e.g.,* "woman in blue" instead of "woman"), which also makes the search easier.

4.3. Correct attentions are alike

To study how attention influences task performance, we measure the spatio-temporal distance between each pair of visual scanpaths, and classify them into three groups based on the correctness of the two answers: both correct, both incorrect, and between correct and incorrect. The distance is measured with a spherical Edit Distance on Real sequence

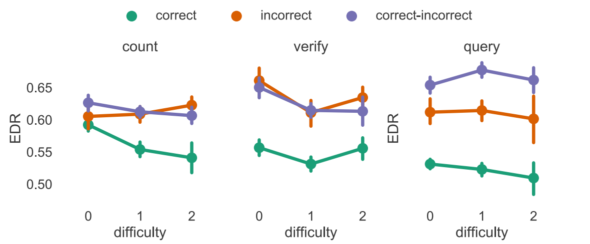

Figure 6: The EDR scores compared over different question types and difficulty levels. Error bars indicate the standard error of the mean.

(EDR) [9] with a distance threshold at 9° (*i.e.,* half fovea size [14]), and lower EDR scores indicate more similar scanpaths. For robustness, this comparison only includes questions with accuracy between 20% and 80%.

According to Figure 6, people who answer correctly have consistently similar attention patterns, whereas the attention patterns leading to incorrect answers are less similar to each other. The between-group similarity is also lower than that within the correct group. This holds true across difficulty levels for all question types. Given that the distance gap is evidently smaller for **count** questions, we hypothesize that this is because the order of counting each target can be different for participants who all count correctly.

4.4. Incorrect attentions fail with different patterns

We further analyze qualitative examples of correct and incorrect attentions to understand why their differences lead to different answers. In particular, Figure 7 illustrates examples of typical cases of wrong answers. While all of the correct fixation maps highlight the important regions where the answer is grounded, incorrect attentions differ from the correct attentions due to diverse reasons:

Missing important cues. Figures 7a-7c show typical examples of missing task-relevant cues. Causes of such misses can be three-fold: first, the answer can be grounded in a less salient region and difficult to find (*e.g.,* Figure 7a, some people are walking under the trees); second, people's subjective bias may lead to biased attention (*e.g.,* Figure 7b, some people answer "car" without looking to the back of the vehicle); third, people's attention can be distracted by visually or semantically similar objects (*e.g.,* Figure 7c, the street lamp looks like a flag pole). All these different factors can lead to the failure of finding the correct answers.

Looking, but not seeing. Many questions require paying close attention to the visual cues. For example, in Figure 7d there are two pandas in front of the camera and another one behind. The two pandas are very close to each other, and people can easily miscount them as one if not paying enough attention to them. In these cases, the amount of attention or time spent on observing the visual cues can influence the correctness of an answer.

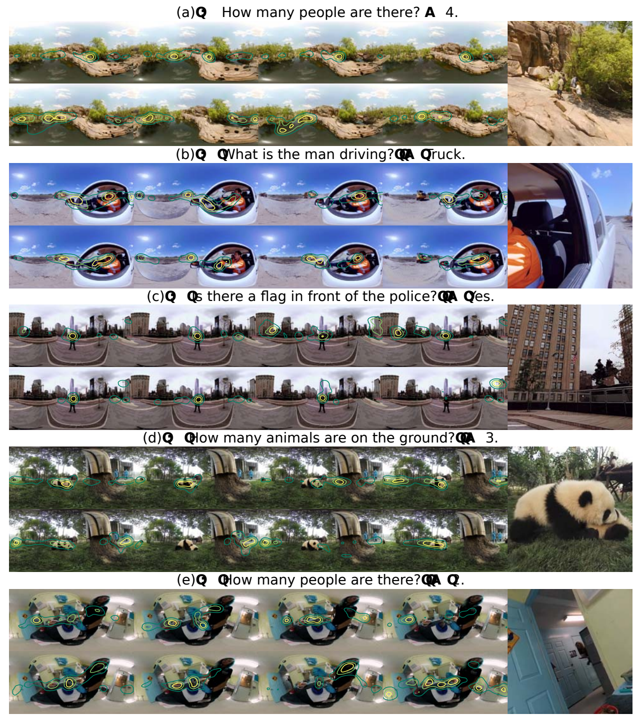

Figure 7: Examples of correct (row 1) and incorrect (row 2) attention patterns. Fixation maps overlaid as contour maps are averaged across all frames (column 1) and every third of the frames (columns 2-4). Column 5 shows the local regions of interest for answering correctly.

Wrong timing. Timing is also a critical factor. Since the scene is changing, looking at the right places yet missing the key moments will lead to incorrect answers. As shown in Figure 7e, the second person only appears at the door for a short interval of time (see column 3). People who answer correctly consistently look at the door at the key moment, while those with incorrect answers are either early or late.

Our analyses suggest strong correlations between attention and task performance, as well as fine-grained differences between correct and incorrect attentions. More examples are shown in the Supplementary Materials.

5. Predicting Correct and Incorrect Attentions

Understanding correct and incorrect attention patterns can play an essential role on distinguishing the important visual features from the hard-negative priors and distracters. In this section, we present a new attention prediction model with the awareness of answer correctness, to further demonstrate the major impacts of our dataset.

Most attention prediction models simulate the bottom-up pathway of human vision [19, 36, 45]. Though some can be trained with gaze data recorded in top-down tasks,

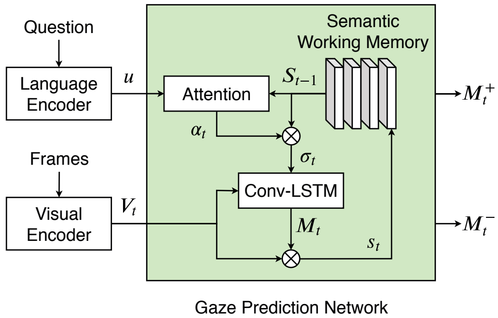

Figure 8: Architecture of the proposed correctness-aware attention prediction model.

few efforts have been made to explicitly model the impact of top-down factors or characterize fine-grained differences between correct and incorrect attention patterns. In this section, we propose a novel correctness-aware attention prediction network to addresses both issues.

As shown in Figure 8, the proposed model consists of a *Visual Encoder* (attentive VGG [45]) and a *Language Encoder* (Skip-Thought model [25]) to extract visual features V_t and language features u from the video and question inputs, and a new *Gaze Prediction Network* that predicts the correct and incorrect attention maps. Different from conventional models, our model simultaneously computes the two attentions and enables knowledge sharing among them. Moreover, the semantic working memory (SWM) takes into account the question information and the visual semantics attended over time, characterizing the role of top-down task in affecting the spatial distributions as well as temporal order of eye fixations. In addition to the model design, to capture the differences between the correct and incorrect attentions, we further propose a new fine-grained difference (FGD) loss to better differentiate the two types of attention.

5.1. Semantic Working Memory

Previous gaze prediction networks implicitly model temporal dynamics [45] or rely on short-term correlation between consecutive frames [52]. Differently, the proposed SWM explicitly and selectively memorizes the most task-relevant semantics attended over time. Specifically, we define the SWM at time t as $S_t = [s_1, \ldots, s_t]$ where $s_t \in \mathbb{R}^d$ is the memorized visual semantics at time t. In order to simultaneously predict both the correct and incorrect attentions, two SWM blocks (S_t^+ and S_t^-) are used in the proposed Gaze Prediction Network to memorize visual semantics attended by correct and incorrect attentions.

Specifically, we first develop a selective mechanism to recall the most relevant information σ_t from the previously memorized semantics S_{t-1}. With the language features $u \in \mathbb{R}^d$ to incorporate the task information, and the semantics attended at the previous time step s_{t-1} to capture the temporal dynamics, such selection is achieved via $\sigma_t = \alpha_t S_{t-1}$, where

$$\alpha_t = W_\alpha(W_S S_{t-1} + W_s s_{t-1} + W_u u) \tag{1}$$

is a temporal attention vector indicating the dynamic importance of each historical time step t. It determines what visual semantics to recall from the memory for the computation of σ_t. Here, W_S, W_s and W_u are all trainable weights of the corresponding factors, and W_α is trained to optimize the temporal attention α_t. The weights are shared between the two SWM blocks to allow knowledge sharing between both attentions.

The recalled semantics σ_t^+ and σ_t^- (corresponding to correct and incorrect attentions) are then combined with the visual features $V_t \in \mathbb{R}^{d\times w\times h}$ and processed with a convolutional LSTM, where w and h are the width and height of the visual features respectively. They are used to adaptively control the gate functions of the LSTM:

$$\begin{aligned} i_t = {} & W_{v_i}V_t + W_{h_i}h_{t-1} + W_{c_i}c_{t-1} \\ & + W_{\sigma_i}^+\sigma_t^+ + W_{\sigma_i}^-\sigma_t^- + b_i \end{aligned} \tag{2}$$

$$\begin{aligned} f_t = {} & W_{v_f}V_t + W_{h_f}h_{t-1} + W_{c_f}c_{t-1} \\ & + W_{\sigma_f}^+\sigma_t^+ + W_{\sigma_f}^-\sigma_t^- + b_f \end{aligned} \tag{3}$$

$$\begin{aligned} o_t = {} & W_{v_o}V_t + W_{h_o}h_{t-1} + W_{c_o}c_{t-1} \\ & + W_{\sigma_o}^+\sigma_t^+ + W_{\sigma_o}^-\sigma_t^- + b_o \end{aligned} \tag{4}$$

where i_t, f_t, o_t are the input, forget and output gates. The h_{t-1} and c_{t-1} are the hidden states. W_{v_i}, W_{h_i}, W_{c_i}, W_{v_f}, W_{h_f}, W_{c_f}, W_{v_o}, W_{h_o} are the weights of the corresponding factors in the gate functions, while $W_{\sigma_i}^+$, $W_{\sigma_f}^+$, $W_{\sigma_o}^+$, $W_{\sigma_i}^-$, $W_{\sigma_f}^-$, $W_{\sigma_o}^-$ are the weights for incorporating the recalled semantics from the memory.

Finally, the predicted attention maps $M_t = [M_t^+, M_t^-]$ are computed as $M_t = W_{out}h_t$, where W_{out} indicates the output-layer parameters. The memories for the two attentions are updated with the newly attended semantics:

$$S_t[t]^{+/-} = W_{att}(M_t^{+/-} \odot V_t) \tag{5}$$

where W_{att} are the learned weights to further encode the attended semantics in the visual features V_t, and $\circ$ indicates the Hadamard product.

By incorporating the SWM blocks, our model is able to associate task information with the visual inputs, and adaptively aggregate important semantics over time to benefit the attention prediction across all video frames.

5.2. Fine-Grained Difference Loss

We propose a fine-grained difference (FGD) loss to encourage the model to differentiate the two outputs. First, we compute the difference between the two ground-truth fixation maps $\Delta F_t = F_t^+ - F_t^-$ and those between the two outputs $\Delta M_t = M_t^+ - M_t^-$. The FGD loss is denoted as

$$L_{FGD} = \sum L_{CC}(M_t^+ \circ |\Delta F_t|, M_t^- \circ |\Delta F_t|) + \gamma \sum \left[(\Delta M_t - \Delta F_t)^2 \circ |\Delta F_t^{'}| \right] \quad (6)$$

where L_{CC} represents the Correlation Coefficient [33]). The first term of the loss normalizes the attentions based on the magnitude of differences in the ground truth, paying more attention to the positions where two ground truth have larger differences, and then enforces the model to predict differently by minimizing their correlation. The second term further minimizes the discrepancies between the differences in the predicted and ground truth attentions. To characterize the spatial distribution and accurate positions of fixations, we follow ACLNet [45] and use both smoothed ΔF_t and unsmoothed fixation maps $\Delta F_t^{'}$ in our loss. The hyperparameter γ balances the contributions of the two loss terms.

Our final loss is defined as a linear combination of the FGD loss and the loss terms that independently optimize the two outputs:

$$L = L^+ + L^- + \beta \cdot L_{FGD} \quad (7)$$

where $L^{+/-}$ are defined a combination of attention evaluation metrics [45] to measure the distances between $M^{+/-}$ and $F^{+/-}$. The hyperparameter β balances the contributions of the loss terms.

6. Experiments and Results

Dataset. For our experiments, we split the dataset into 658 training samples, 96 validation samples, and 221 test samples. We train and evaluate models on the IQVA dataset to perform two different tasks: correctness-aware attention prediction and aggregated attention prediction regardless of correctness. Given a video clip and a question, the goal of the former is to predict both the correct and incorrect attentions for each video frame, while the latter predicts an aggregated fixation map. To reduce the bias caused by imbalanced numbers of correct and incorrect answers, for the first task we only consider samples with answer accuracy between 20% and 80% (*i.e.,* 50.51% of the samples). For the second task, we use all of the available data. Following [52], all videos are temporally downsampled by 5.

Evaluation Protocols. We use five popular attention evaluation metrics in our experiments, including Correlation Coefficient (CC) [33], Normalized Scanpath Saliency (NSS) [37], Kullback–Leibler Divergence (KLD) [26], Similarity (SIM) [41] and shuffled AUC [6]. The distortions of equirectangular projections are corrected with a sine weighting function following [13]. As existing state-of-the-art models are designed only for bottom-up attention, to accommodate our dataset with top-down attention, we slightly modify them to efficiently take into account the question information similarly to our model. More details are provided in the Supplementary Materials.

Training. We train our model with the proposed objective using Adam [24] optimizer with learning rate 10^{-4} and weight decay 10^{-3}. The hyperparameters β and γ are empirically set to 0.5 and 2 respectively, based on the validation set performance. Resolution of the visual input is set to 512×256. For the existing models, we follow their original settings and train two independent models using correct or incorrect data respectively. All of the models are initialized with weights pre-trained on ImageNet classification. Batch size 1 is used for training all models similar to [19], since larger batch sizes require higher computational cost, and do not result in obvious improvement. The best models are selected based on their performance on the validation set.

6.1. Predicting Correct and Incorrect Attentions

We first evaluate our model on predicting correct and incorrect attentions. Quantitatively, Table 3 shows that our baseline model (*i.e.,* Multi-Att) that predicts two attentions without memory and the proposed loss significantly outperforms the existing state-of-the-art, indicating the importance of knowledge sharing in developing better understanding of the task. Moreover, the increased performance achieved with the SWM (*i.e.,* Multi-Att + SWM) demonstrates the effectiveness of adaptively incorporating the visual semantics attended over time. Finally, our complete model with both the SWM and the FGD loss (*i.e.,* Multi-Att + SWM + FGD) achieves the best results on predicting both attentions among all evaluation metrics.

Qualitatively, as shown in Figure 9, ground truth attentions corresponding to correct and incorrect answers (the rightmost column) show distinct differences, indicating that attention plays a role in these cases (more details and discussions in Section 4 and the Supplementary Materials). From the modeling aspect, while most existing models (see columns 2-5) highlight regions of interest (*i.e.,* people in both examples) to some degree, they all fail to differentiate attention patterns leading to correct and incorrect answers (*i.e.,* predicted attention patterns in both rows are similar). In comparison, the proposed model (see column 6) not only captures the regions of interest related to the question, but also differentiates the regions crucial for correct answers (*i.e.,* the people skateboarding far from the camera and the man with a mic on the right) from the others (*i.e.,* people not matching these descriptions). Note that predictions of correct and incorrect attentions from existing models are trained with the respective data. The lack of capability in differentiating the difference demonstrates the needs in model designs to close this gap.

Results above show the effectiveness of our model architecture, semantic memory, and loss in differentiating the at-

	Correct					Incorrect				
	CC	NSS	KLD	SIM	sAUC	CC	NSS	KLD	SIM	sAUC
SALICON [19]	0.407	2.010	1.645	0.350	0.429	0.389	1.914	1.689	0.326	0.431
SALNet [36]	0.412	2.028	1.560	0.347	0.451	0.380	1.946	1.703	0.329	0.397
ACLNet [45]	0.402	1.938	1.606	0.341	0.448	0.378	1.900	1.717	0.322	0.424
Spherical U-Net [52]	0.268	1.225	1.955	0.262	0.333	0.247	1.167	2.085	0.234	0.343
Multi-Att	0.426	2.293	1.479	0.365	0.446	0.411	2.225	1.570	0.344	0.447
Multi-Att + SWM	0.439	2.316	1.434	0.368	0.456	0.422	2.205	1.561	0.344	0.455
Multi-Att + SWM + FGD	**0.441**	**2.375**	**1.429**	**0.371**	**0.462**	**0.424**	**2.267**	**1.524**	**0.345**	**0.469**

Table 3: Comparison of attention prediction performances. Best results are highlighted in bold.

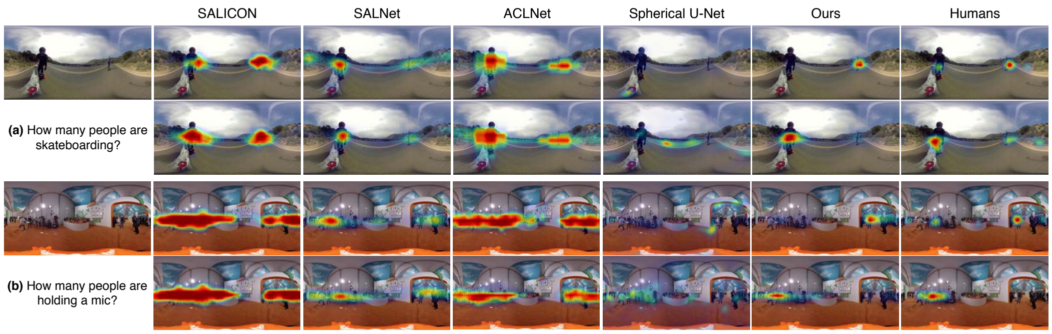

Figure 9: Qualitative comparison of the predicted correct (row 1) and incorrect (row 2) fixation maps.

tentions that lead to different task performance. It opens up a new paradigm in attention modeling by considering task performance. In addition, the difference in output naturally highlights regions to be fixated (*e.g.,* visual cues relevant to the task) or to be avoided (*e.g.,* visual distractors), which has direct benefits to a variety of applications.

6.2. Predicting Aggregated Attention

The proposed dataset can also be utilized for predicting aggregated attention regardless of correctness. In this section, we benchmark the existing models and the proposed one for predicting the aggregated attention on our dataset. For the proposed model, we adopt our pre-trained model in the previous experiments and develop a Map Aggregation module that adaptively integrates the predicted correct and incorrect attention maps into an aggregated attention map. As shown in Table 4, with an understanding of the correct and incorrect attentions developed in the previous task, the proposed model is able to consistently outperform the existing models on aggregated attention prediction. Please refer to our Supplementary Materials for details.

	CC	NSS	KLD	SIM	sAUC
SALICON [19]	0.514	2.098	1.103	0.449	0.483
SALNet [36]	0.498	2.083	1.128	0.439	0.463
ACLNet [45]	0.493	2.022	1.146	0.438	0.466
Spherical U-Net [52]	0.343	1.309	1.547	0.331	0.408
Ours	**0.538**	**2.409**	**1.047**	**0.466**	**0.498**

Table 4: Comparative results of predicting aggregated attention for 360° videos. Best results are highlighted in bold.

7. Conclusion

We introduce a new dataset for task-driven attention in immersive scenes. With the new paradigm featuring diverse immersive scenes and questions, as well as manual annotations of answer correctness, the proposed dataset not only serves as a new benchmark for top-down visual attention modeling, but also opens up new research opportunities by taking into account task performance. Our analyses demonstrate a strong correlation between attention and task performance, opening a new avenue for research in performance-aware human attention in real-life scenarios. Furthermore, we propose a correctness-aware attention prediction model together with a new loss for jointly predicting the correct and incorrect attention patterns. Our model highlights the importance of incorporating knowledge from both types of attentions for capturing their fine-grained differences as well as predicting the aggregated attention. Future efforts will be made towards two research directions: characterizing the attention patterns of individuals to understand and predict their task performances, and improving the performance and interpretability of neural networks with improved attention mechanism.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Alan Allport. *Visual attention.* The MIT Press, 1989.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[4] Ali Borji. Saliency prediction in the deep learning era: An empirical investigation. *arXiv preprint arXiv:1810.03716*, 2018.

[5] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *TPAMI*, 35(1):185–207, 2013.

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

[7] Georg Buscher, Ed Cutrell, and Meredith Ringel Morris. What do you see when you're surfing? using eye tracking to predict salient regions of web pages. In *CHI*, 2009.

[8] Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, and Frédo Durand. Where should saliency models look next? In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 809–824, 2016.

[9] Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *SIGMOD*, pages 491–502, 2005.

[10] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *ECCV*, pages 397–412, 2018.

[11] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 360-degree video head movement dataset. In *MMsys*, pages 199–204, 2017.

[12] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 163:90–100, 2017.

[13] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A dataset of head and eye movements for 360 videos. In *MMsys*, pages 432–437, 2018.

[14] Donald H Edwards. Neuroscience. third edition. edited by dale purves, george j augustine, david fitzpatrick, william c hall, anthony-samuel lamantia, james o mcnamara , and s mark williams. *The Quarterly Review of Biology*, 81(1):86–87, 2006.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[16] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In *CHI*, pages 3511–3522, 2017.

[17] Jesús Gutiérrez, Erwan J David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360° contents. In *QoMEX*, pages 1–3, 2018.

[18] Brian Hu, Ishmael Johnson-Bey, Mansi Sharma, and Ernst Niebur. Head movements during visual exploration of natural images in virtual reality. In *CISS*, pages 1–6, 2017.

[19] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In *ICCV*, 2015.

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

[21] Tilke Judd, Krista A. Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. *ICCV*, pages 2106–2113, 2009.

[22] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *CVIU*, 163:3 – 20, 2017.

[23] Christopher Kanan, Mathew H. Tong, Lingyun Zhang, and Garrison W. Cottrell. SUN: Top-down saliency using natural statistics. *Visual Cognition*, 17(6-7):979–1003, 2009.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[25] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *NIPS*, pages 3294–3302, 2015.

[26] Solomon Kullback. *Information Theory and Statistics*. Wiley, 1959.

[27] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, pages 639–655, 2018.

[28] Y. Li, Zhefan Ye, and J. M. Rehg. Delving into egocentric actions. In *CVPR*, pages 287–295, 2015.

[29] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 360 video viewing dataset in head-mounted virtual reality. In *MMsys*, pages 211–216, 2017.

[30] Mateusz Malinowski and Mario Fritz. Towards a visual turing challenge. *arXiv preprint arXiv:1410.8027*, 2014.

[31] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *CVPR*, pages 9562–9571, 2019.

[32] Vinod Menon and Lucina Q Uddin. Saliency, switching, attention and control: a network model of insula function. *Brain Structure and Function*, 214(5-6):655–667, 2010.

[33] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[34] Anh Nguyen and Zhisheng Yan. A saliency dataset for 360-degree videos. In *MMsys*, pages 279–284, 2019.

[35] Antje Nuthmann, Wolfgang Einhäuser, and Immo Schütz. How well can saliency models predict fixation selection in scenes beyond central bias? a new approach to model evaluation using generalized linear mixed models. *Frontiers in human neuroscience*, 11:491, 2017.

[36] Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor. Shallow and Deep Convolutional Networks for Saliency Prediction. In *CVPR*, 2016.

[37] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.

[38] Yao Qin, Mengyang Feng, Huchuan Lu, and Garrison W. Cottrell. Hierarchical cellular automata for visual saliency. *IJCV*, 126(7):751–770, 2018.

[39] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. A dataset of head and eye movements for 360 degree images. In *MM-sys*, pages 205–210, 2017.

[40] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NIPS*, pages 1548–1558, 2018.

[41] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000.

[42] Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV*, pages 33–46, 2014.

[43] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. Saliency in vr: How do people explore virtual environments? *TVCG*, 24(4):1633–1642, 2018.

[44] Evgeniy Upenik and Touradj Ebrahimi. A simple method to obtain visual attention data in head mounted virtual reality. In *ICME*, pages 73–78, 2017.

[45] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji. Revisiting Video Saliency Prediction in the Deep Learning Era. *TPAMI*, 2019.

[46] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. A dataset for exploring user behaviors in vr spherical video streaming. In *MMsys*, pages 193–198, 2017.

[47] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, pages 451–466, 2016.

[48] Yingyue Xu, Xiaopeng Hong, Qiuhai He, Guoying Zhao, and Matti Pietikäinen. A task-driven eye tracking dataset for visual attention analysis. In Sebastiano Battiato, Jacques Blanc-Talon, Giovanni Gallo, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *ACIVS*, pages 637–648, 2015.

[49] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013.

[50] Lingyun Zhang, Matthew H Tong, and Garrison W Cottrell. SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In *AAAI*, pages 2944–2949, 2009.

[51] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *ECCV*, pages 488–503, 2018.

[52] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency Detection in 360° Videos. In *ECCV*, 2018.