Fantastic Answers and Where to Find Them: Immersive Question-Directed Visual Attention (Supplementary Materials)

Ming Jiang* Shi Chen* Jinhui Yang Qi Zhao
University of Minnesota
{mjiang, chen4595, yang7004, qzhao}@umn.edu

The supplementary materials provide additional analyses and results on the proposed Immersive Question-directed Visual Attention (IQVA) dataset, together with more details on the model design. Specifically, we 1) present additional examples of correct and incorrect attention patterns, and video illustrations for all the examples; 2) provide additional results on the effects of different hyperparameters and more qualitative examples; and 3) elaborate the design of our Map Aggregation module for general 360° video saliency prediction, and the modifications of existing bottom-up saliency models to take into account task information for a fair comparison.

1. Additional Data Analyses and Supplementary Video

In this section, we extend the analyses of the human fixation data and its influence in answer correctness with additional examples. In Figure 1, we show the key frames in which the correct attention was on the important visual cues. At the same time, the incorrect attention either missed the cues or did not last for enough time to understand the observed information.

Missing important cues. Figures 1(a)-(d) show examples of incorrect attention missing important cues that answer the questions. Figures 1(a)-(c) are the same examples shown in the main paper. Figure 1(d) shows that the white backpack is not attended by people who answer incorrectly.

Looking, but not seeing. Figures 1(e)-(f) show examples of incorrect attention looking at the visual cues but failed to spend enough time to understand them. Figure 1(e) is the same example shown in the main paper, and Figure 1(f) shows that people who answered incorrectly do not see the black bowl, even though they also look at the white table at some point.

Wrong timing. Figures 1(g)-(h) show examples of incorrect attention missing important moments of the scenes. Figure 1(g) is the same example shown in the main paper. Figure 1(h) shows that people who answered incorrectly miss the train driver when the train passes by.

For more details, please refer to the **supplementary video**.

2. Additional Results

2.1. Effects of Hyperparameters

The learning objective presented in the main paper consists of two hyperparameters, *i.e.* β and γ. They balance the losses for independent attention maps and their difference, and determine the contributions of the two terms in our Fine-grained Difference (FGD) loss. We empirically set $\beta = 0.5$ and $\gamma = 2$ based on ablation studies. Table 1 reports the model performances under different settings.

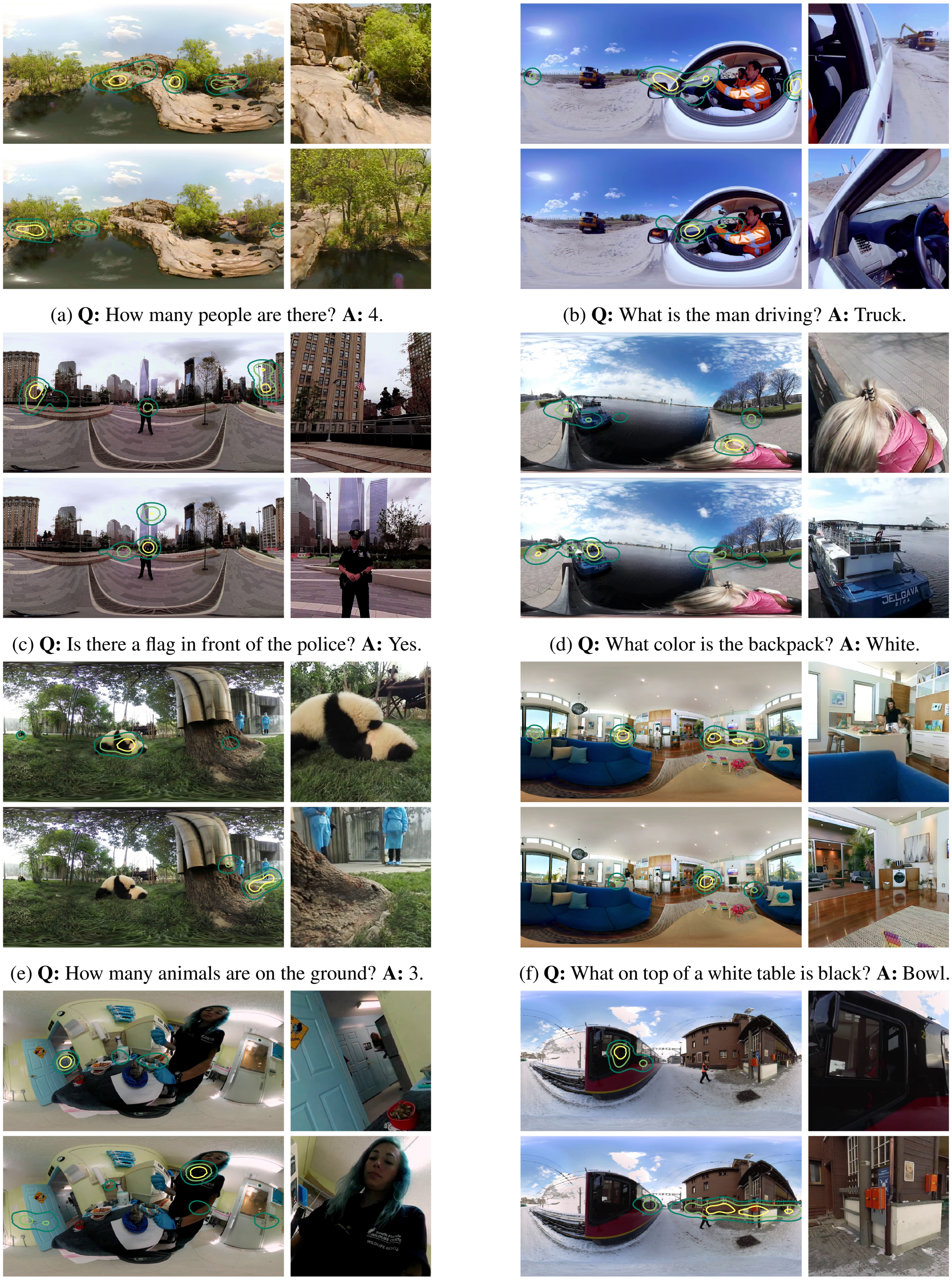
We first investigate the effects of β that balances the proposed FGD loss, and the losses on each independent attention map, with γ fixed to 2. On the one hand, with a small β (*i.e.* $\beta = 0.1$ and $\beta = 0.3$) that lessens the contribution of the FGD loss, the models have difficulties differentiating the two attention maps and tend to provide sub-optimal results. On the other hand, assigning a large value to β (*i.e.* $\beta = 0.8$ and $\beta = 1$) overemphasizes the difference between the attention maps, resulting in difficulties in fitting each independent attention map. Setting $\beta = 0.5$ leads to a reasonable trade-off between learning independent attention maps and their difference, and produces the best results.

Next, we study the effects of γ that determines the contributions of different terms in the FGD loss. According to the results, a sub-optimal value can result in difficulties separating the two attention maps (*i.e.* $\gamma = 3$ and $\gamma = 4$) or fitting to the ground truth difference (*i.e.* $\gamma = 0.5$ and $\gamma = 1$). On the contrary, $\gamma = 2$ provides a good balance between the two terms and leads to the best results.

2.2. Qualitative Examples

Additional qualitative results in Figure 2 further demonstrate the effectiveness of our model for the correctness-aware attention prediction. Similar to the qualitative results shown in the main paper, we can see that existing models

[*] Equal contribution.



(a) **Q:** How many people are there? **A:** 4.

(b) **Q:** What is the man driving? **A:** Truck.

(c) **Q:** Is there a flag in front of the police? **A:** Yes.

(d) **Q:** What color is the backpack? **A:** White.

(e) **Q:** How many animals are on the ground? **A:** 3.

(f) **Q:** What on top of a white table is black? **A:** Bowl.

(g) **Q:** How many people are there? **A:** 2.

(h) **Q:** Who is wearing sunglasses? **A:** Man.

Figure 1: Correct (row 1) and incorrect (row 2) attentions at a key moment when the correct attention was paid to the most important visual cues that lead to correct answers. **Left:** equirectangular fixation maps are overlaid as contours. **Right:** the most fixated regions shown in the perspective view.

	Correct					Incorrect				
	CC	NSS	KLD	SIM	sAUC	CC	NSS	KLD	SIM	sAUC
$\beta = 0.1$	0.431	2.263	1.435	0.359	**0.480**	0.417	2.228	1.528	0.345	**0.474**
$\beta = 0.3$	0.431	2.239	1.457	0.361	0.461	0.411	2.155	1.565	0.337	0.457
$\beta = 0.5$	**0.441**	**2.375**	**1.429**	**0.371**	0.462	**0.424**	**2.267**	**1.524**	**0.345**	0.469
$\beta = 0.8$	0.416	2.243	1.533	0.359	0.438	0.413	2.150	1.591	0.338	0.467
$\beta = 1$	0.386	2.201	1.673	0.351	0.406	0.369	1.769	1.652	0.324	0.409
$\gamma = 0.5$	0.432	2.294	1.463	0.364	0.455	0.408	2.197	1.596	0.342	0.435
$\gamma = 1$	0.437	2.257	1.448	0.364	0.465	0.420	2.134	1.536	0.338	0.474
$\gamma = 2$	**0.441**	**2.375**	**1.429**	**0.371**	0.462	**0.424**	**2.267**	**1.524**	**0.345**	0.469
$\gamma = 3$	0.437	2.237	1.433	0.355	**0.501**	0.417	2.179	1.533	0.330	**0.502**
$\gamma = 4$	0.425	2.108	1.472	0.350	0.472	0.412	2.092	1.552	0.332	0.469

Table 1: Quantitative comparisons under different settings of hyperparameters β and γ. Best results are highlighted in bold.

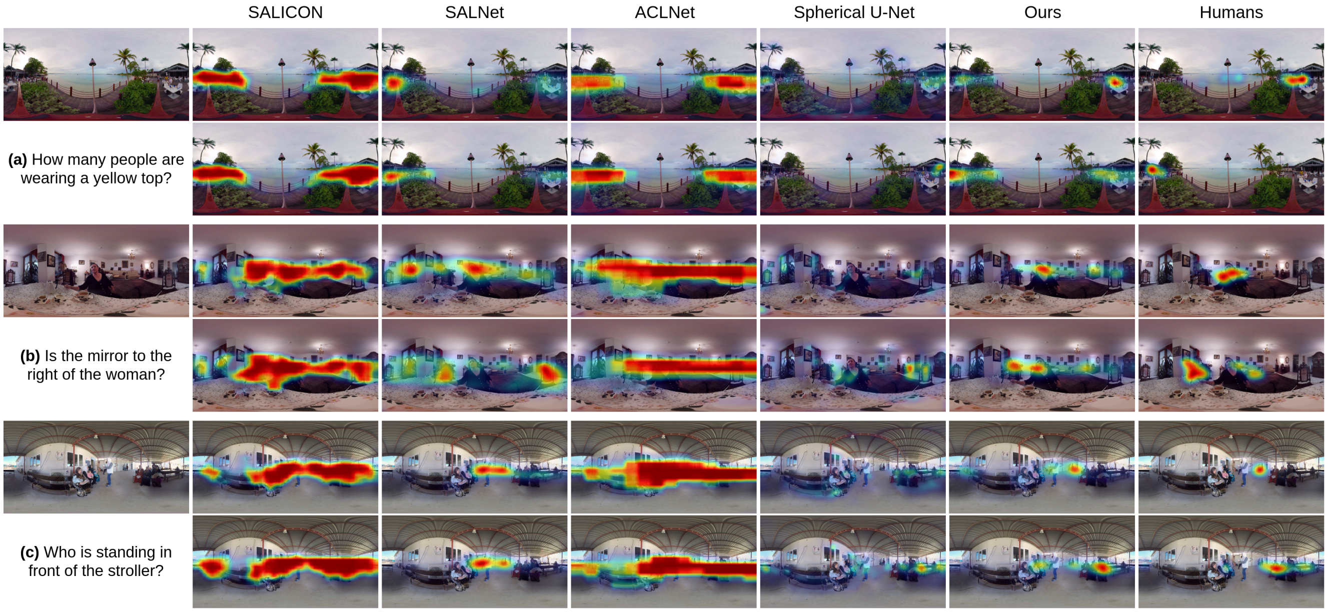

SALICON
SALNet
ACLNet
Spherical U-Net
Ours
Humans
(a) How many people are wearing a yellow top?
(b) Is the mirror to the right of the woman?
(c) Who is standing in front of the stroller?

Figure 2: Additional qualitative comparisons of the predicted correct (row 1) and incorrect (row 2) attention maps.

lack the capability of capturing the key differences between the correct and incorrect attentions. They either fail to locate the regions of interest as well as distracting objects in the corresponding attention maps (*i.e.* example (b) and (c) for SALNet), or predict identical maps for two attentions that have significant differences (*i.e.* in most of the cases). In comparison, our model is capable of identifying the key differences between two attentions and providing better results. Specifically, in all of the examples, our model successfully distinguishes the regions of interest that lead to the correct answers (*i.e.* the man in a yellow top, the woman, and the stroller in the back) from the distracting objects (*i.e.* a group of people not wearing a yellow top, a mirror-like object near the woman, and the woman with a baby).

The aforementioned results highlight the necessity of learning the differences between correct and incorrect attentions, and further validate the effectiveness of the proposed model.

3. Additional Modeling Details

3.1. Map Aggregation Module

In addition to the correctness-aware attention prediction, we also carry out experiments on general 360° video saliency prediction. Though our model proposed in the main paper is primarily designed for correctness-aware attention prediction, it can be seamlessly adapted to general saliency prediction by introducing a Map Aggregation module. The Map Aggregation module takes the correct and incorrect attention maps predicted by our model as inputs, and learns adaptive weights for combining the two maps, based on both the question and visual information.

More specifically, given the language features u and the visual semantics recalled from the semantic working memory σ_t, we first compute the adaptive weights λ for the two attention maps as follows:

$$\lambda = W_\lambda(W_{u'}u + W_\sigma[\sigma_t^+, \sigma_t^-]) \quad (1)$$

where $[\sigma_t^+, \sigma_t^-]$ is the concatenation of semantics for correct and incorrect attentions, $W_{u'}$ and W_σ are trainable weights for the corresponding factors, W_λ is used for computing the adaptive weights that integrate the two attention maps. Then, the final saliency map M_{Sal} is computed as a linear combination of the correct and incorrect attention maps $M_t = [M_t^+, M_t^-]$ based on the adaptive weights as $M_{Sal} = \lambda M_t$. Similar to [2], the saliency map is normalized by dividing by its maximum.

3.2. Modifications of Existing Bottom-up Models

Existing models [3, 4, 5, 6] focus on predicting bottom-up saliency maps without taking into account the influences of top-down factors, *e.g.* questions in the proposed dataset. For a fair comparison with our model, we slightly modify them to incorporate the question features. Specifically, the question information is processed with the same Language Encoder used in our model, and integrated with the bottom-up visual features via a standard element-wise addition. The features from different modalities, *i.e.* visual features and language features, are combined before the last layer to predict the final correctness-aware attention maps. Note that since the Spherical U-Net [6] requires extensive computational resources before incorporating the question information (over 12 GB GPU memory for processing a single image), for better efficiency and generalization, we replace their spherical convolution layers with the ones proposed in [1].

References

[1] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images. In *ECCV*, 2018. 4

[2] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *TIP*, 2018. 4

[3] 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. 4

[4] 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. 4

[5] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji. Revisiting Video Saliency Prediction in the Deep Learning Era. *TPAMI*, 2019. 4

[6] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency Detection in 360° Videos. In *ECCV*, 2018. 4