Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? (Supplementary Material)

Abhishek Das[1*] **Harsh Agrawal**[1*] **C. Lawrence Zitnick**[2] **Devi Parikh**[1,3] **Dhruv Batra**[1,3]
[1]Virginia Tech [2]Facebook AI Research [3]Georgia Institute of Technology
{abhshkdz, harsh92, parikh, dbatra}@vt.edu, zitnick@fb.com

1 Attention Annotation Interface

We design and test multiple game-inspired novel interfaces for conducting large-scale human studies on Amazon Mechanical Turk (AMT). Our basic interface design consists of a "deblurring" exercise for answering visual questions. Specifically, we present subjects with a blurred image and a question about the image, and ask subjects to sharpen regions of the image that will help them answer the question correctly, in a smooth, click-and-drag, 'coloring' motion with the mouse. The sharpening is gradual: successively scrubbing the same region progressively sharpens it. Figure 1 shows intermediate steps in our attention annotation interface, from a completely blurry image to a deblurred attention map.

Our interface starts by showing a low-resolution blurry version of the image. This is to convey a partial 'holistic' understanding of the scene to the subjects so they may intelligently choose which regions to sharpen. Gradual sharpening with strokes was aimed to capture initial exploration as they tried to get a better sense of the scene, and eventually focussed sharpening to answer the question. Next we describe the three variants of our attention annotation interface that we experimented with.

1.1 Blurred Image without Answer

In our first interface, subjects were shown a blurred image and a question without the answer, and were asked to deblur regions and enter the answer. We found that this interface sometimes resulted in 'exploratory attention', where the subject lightly sharpens large regions of an image to find salient regions that eventually lead them to the answer. However, subjects often ended up with 'incomplete' attention maps since they did not see the high-resolution image and the answer, so they did not know when to stop deblurring or exploring. For instance, for an image with 3 players playing a sport, if the question is "How many players are visible in the image?", the subject might sharpen a region that seems to have the players, count the 2 players in there and answer 2, and completely miss another region of the image that had 1 more. The resulting attention map in this case is incomplete since there are 3 players in the image. This effect of incomplete human attention maps was seen in counting ("How many ...") and binary ("Is there ...") types of questions, and as a result, the answers to these were often incorrect.

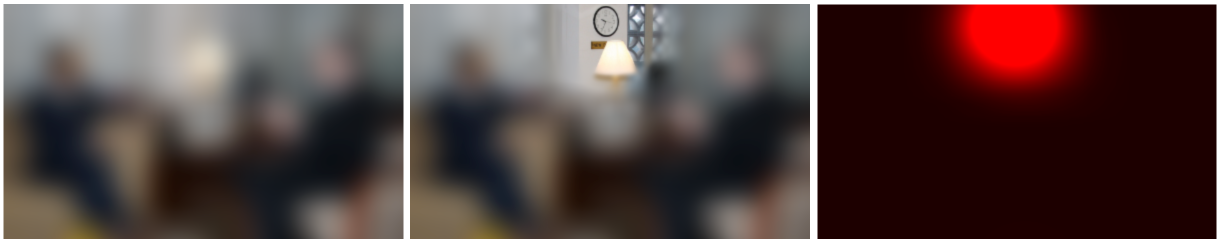
1.2 Blurred Image with Answer

In our second interface, subjects were shown the correct answer in addition to the question and blurred image. They were asked to sharpen as few regions as possible such that someone can answer the question just by looking at the blurred image with sharpened regions. This interface is shown in Figure 2b. Providing the answer fixed the failure cases from the 1st interface, i.e. for counting and binary questions, since the subjects now knew the answer, they continued to explore till they found the answer region in the image.

1.3 Blurred and Original Image with Answer

To encourage exploitation instead of exploration, in our third interface, subjects were shown the question-answer pair and full-resolution original image. In principle, seeing the original (full-resolution) image, the question, and answer provides most information to subjects, thus enabling them to provide the most 'accurate' attention maps. However, this task turns out to be fairly counter-

[*]Denotes equal contribution.



(a) Initial blurred image (b) Regions sharpened by subject (c) Attention map

Figure 1: Deblurring procedure to collect attention maps. We present subjects with a blurred image and ask them to sharpen regions of the image that will help them answer the question correctly, in a smooth, click-and-drag, 'coloring' motion with the mouse.

intuitive – subjects are shown full-resolution images and the answer, and asked to imagine a scenario where someone else has to answer the question without looking at the original image.

Figure 2 shows screen-captures of the 3 data collection interfaces.

Interface Type	Human Accuracy
Blurred Image without Answer	75.2
Blurred Image with Answer	78.7
Blurred & Original Image with Answer	71.2
Original Image	80.0

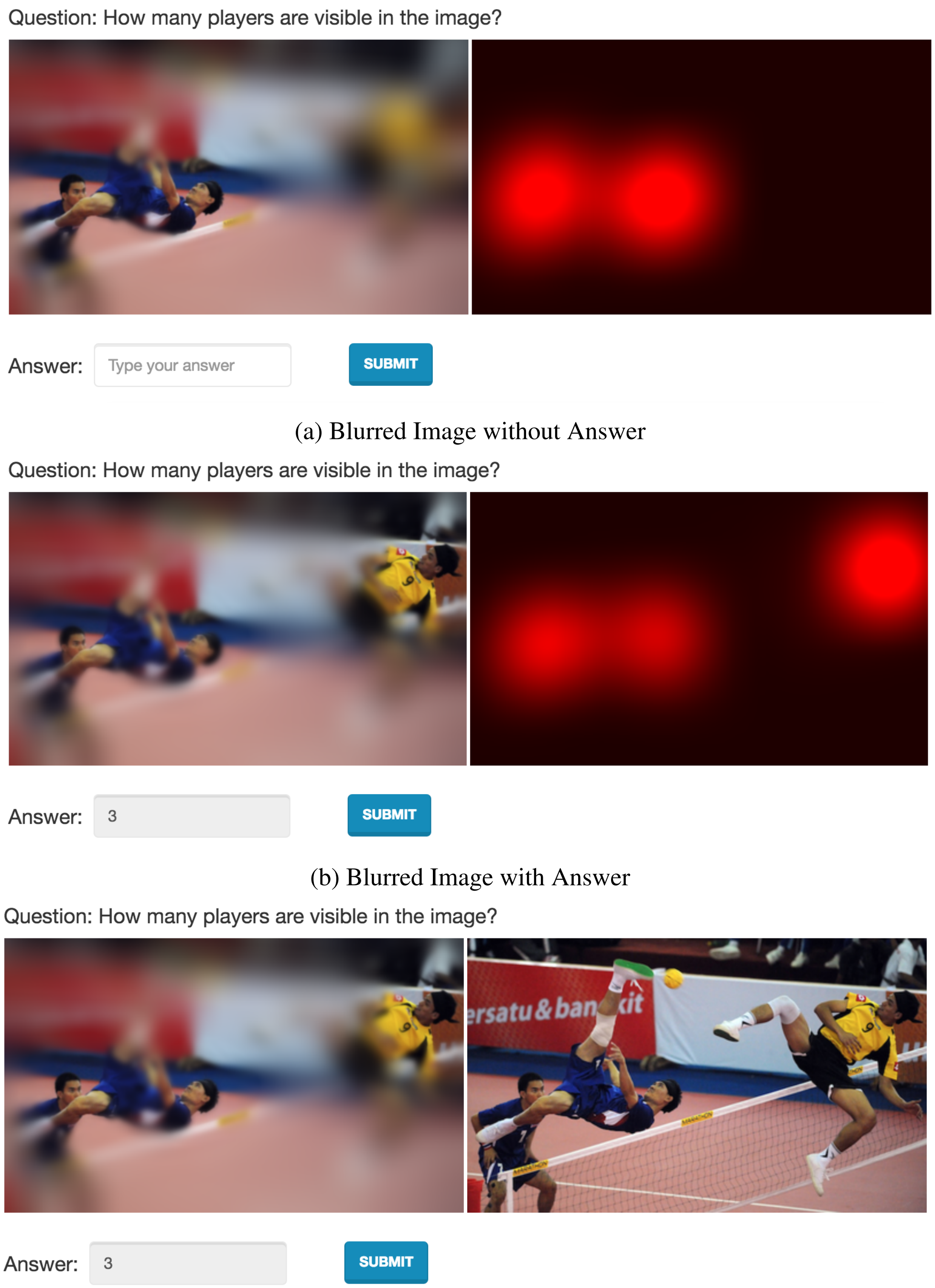
Table 1: Human accuracies to compare the quality of human attention maps collected by different interfaces. Subjects were shown deblurred images from each of these interfaces and asked to answer the visual question.

2 Dataset Evaluation

We ran pilot studies on AMT to experiment with the above described three interfaces. In order to quantitatively evaluate the interfaces, we conducted a second human study where (a second set of) subjects where shown the attention-sharpened images generated from each of the attention interfaces from the first experiment and asked to answer the question. The intuition behind this experiment is that if the attention map revealed too little information, this second set of subjects would answer the question incorrectly. Table 1 shows VQA accuracies of the answers given by human subjects under these 3 interfaces. We can see that the "Blurred Image with Answer" interface (section 1.2) gives the highest accuracy on evaluation by humans. The payment structure on AMT encourages completing tasks as quickly as possible, this implicitly incentivizes subjects to deblur as few regions as possible, and our human study shows that humans can still answer questions. Thus, overall we achieve a balance between highlighting too little or too much. The "Blurred Image with Answer" interface gives highest accuracy in the human evaluation study.

3 Qualitative Examples

Figure 3 and Figure 4 show randomly sampled visualizations of machine-generated and human attention maps.



Question: How many players are visible in the image?

Answer: Type your answer

(a) Blurred Image without Answer

Question: How many players are visible in the image?

Answer: 3

(b) Blurred Image with Answer

Question: How many players are visible in the image?

Answer: 3

(c) Blurred & Original Image with Answer

Figure 2: Attention annotation interface variants. (a) In our first interface, subjects were shown a blurred image and a question without the answer, and were asked to deblur regions and enter the answer. (b) In our second interface, subjects were shown the correct answer in addition to the question and blurred image. They were asked to sharpen as few regions as possible such that someone can answer the question just by looking at the blurred image with sharpened regions. (c) To encourage exploitation instead of exploration, in our third interface, subjects were shown the question-answer pair and full-resolution original image. Out of the three interfaces, Blurred Image with Answer (b) struck the right balance between exploration and exploitation, and gives the highest accuracy on evaluation by humans as described in section 2.

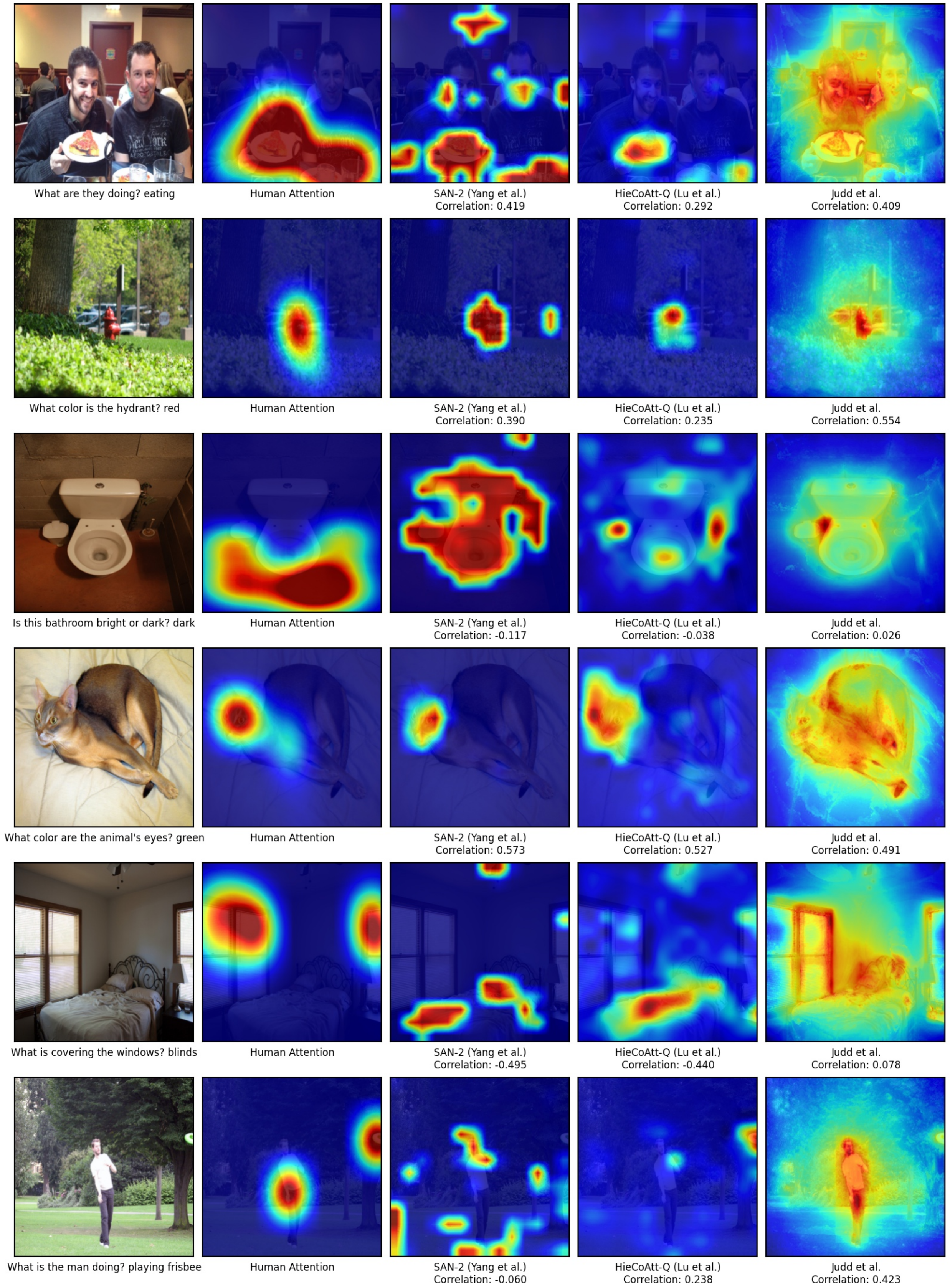

What are they doing? eating
Human Attention
SAN-2 (Yang et al.) Correlation: 0.419
HieCoAtt-Q (Lu et al.) Correlation: 0.292
Judd et al. Correlation: 0.409
What color is the hydrant? red
Human Attention
SAN-2 (Yang et al.) Correlation: 0.390
HieCoAtt-Q (Lu et al.) Correlation: 0.235
Judd et al. Correlation: 0.554
Is this bathroom bright or dark? dark
Human Attention
SAN-2 (Yang et al.) Correlation: -0.117
HieCoAtt-Q (Lu et al.) Correlation: -0.038
Judd et al. Correlation: 0.026
What color are the animal's eyes? green
Human Attention
SAN-2 (Yang et al.) Correlation: 0.573
HieCoAtt-Q (Lu et al.) Correlation: 0.527
Judd et al. Correlation: 0.491
What is covering the windows? blinds
Human Attention
SAN-2 (Yang et al.) Correlation: -0.495
HieCoAtt-Q (Lu et al.) Correlation: -0.440
Judd et al. Correlation: 0.078
What is the man doing? playing frisbee
Human Attention
SAN-2 (Yang et al.) Correlation: -0.060
HieCoAtt-Q (Lu et al.) Correlation: 0.238
Judd et al. Correlation: 0.423

Figure 3: Random samples of human attention (column 2) v/s machine-generated attention (columns 3-5)

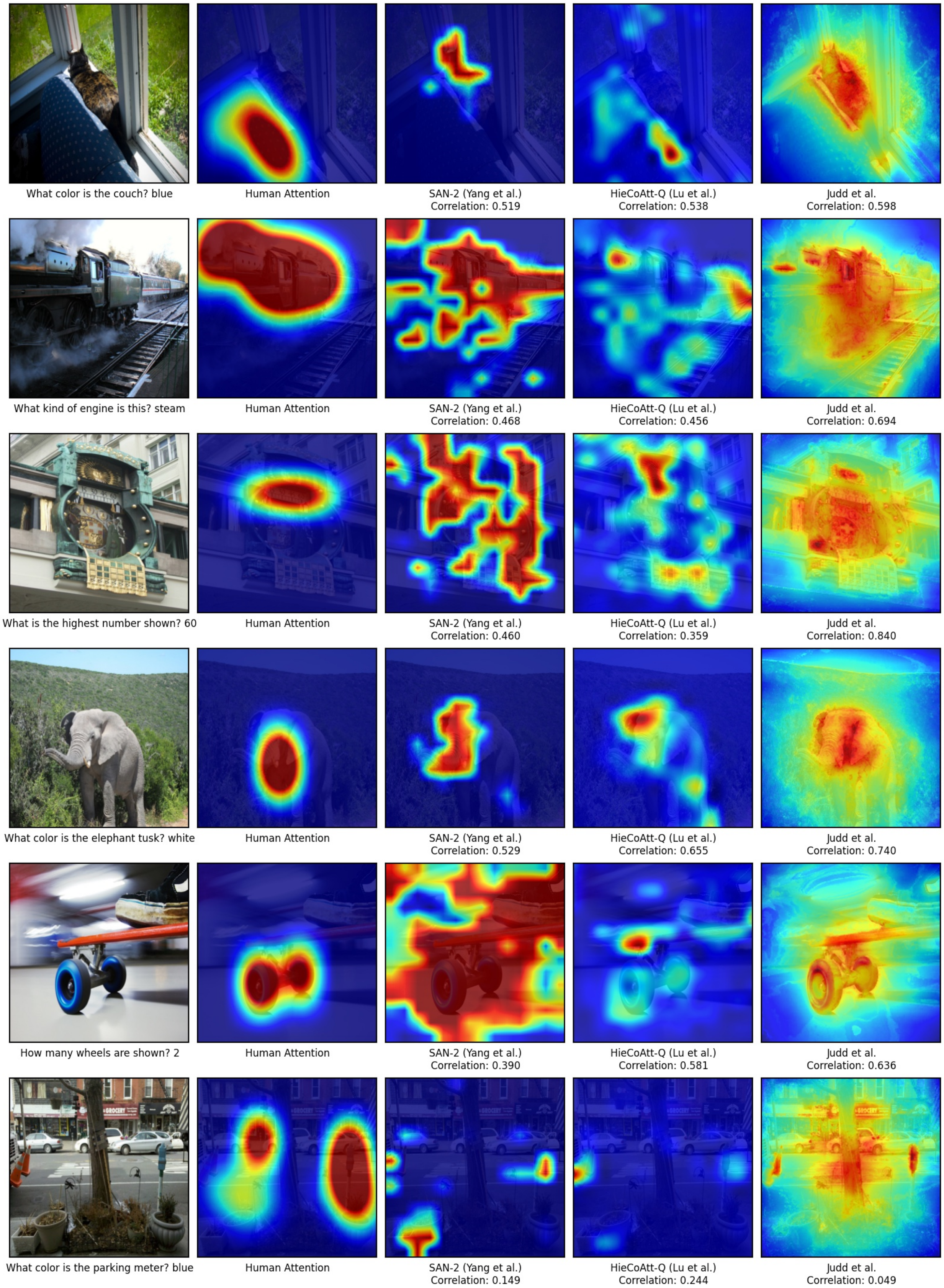


Figure 4: Random samples of human attention (column 2) v/s machine-generated attention (columns 3-5)