
On the Efficacy of Co-Attention Transformer Layers in Visual Question Answering

Ankur Sikarwar
Department of ECE
Birla Institute of Technology, Mesra
ankursikarwardc@gmail.com

Gabriel Kreiman
Center for Brains, Minds and Machines
Harvard Medical School
gabriel.kreiman@tch.harvard.edu

Abstract

In recent years, multi-modal transformers have shown significant progress in Vision-Language tasks, such as Visual Question Answering (VQA), outperforming previous architectures by a considerable margin. This improvement in VQA is often attributed to the rich interactions between vision and language streams. In this work, we investigate the efficacy of co-attention transformer layers in helping the network focus on relevant regions while answering the question. We generate visual attention maps using the question-conditioned image attention scores in these co-attention layers. We evaluate the effect of the following critical components on visual attention of a state-of-the-art VQA model: (i) number of object region proposals, (ii) question part of speech (POS) tags, (iii) question semantics, (iv) number of co-attention layers, and (v) answer accuracy. We compare the neural network attention maps against human attention maps both qualitatively and quantitatively. Our findings indicate that co-attention transformer modules are crucial in attending to relevant regions of the image given a question. Importantly, we observe that the semantic meaning of the question is *not* what drives visual attention, but specific keywords in the question do. Our work sheds light on the function and interpretation of co-attention transformer layers, highlights gaps in current networks, and can guide the development of future VQA models and networks that simultaneously process visual and language streams.

1 Introduction

The ability of humans to efficiently ground information across different modalities, such as vision and language, plays a central role in cognitive function. The interactions between vision and language are highlighted in visual question answering (VQA) tasks, where attentional allocation is naturally routed by combination of sensory and semantic cues. For instance, given an image of people playing football and the question 'What color shirt is the person behind the referee wearing?', subjects rapidly identify the referee, saccade to the player behind the referee, and process the relevant regions of the image to find the answer. A four-year old can easily answer such questions and seamlessly direct visual attention to the relevant regions based on the question.

In contrast, such multi-modal tasks are quite challenging for current AI systems because the solution encompasses several increasingly complex subtasks. First of all, the system has to interpret the key elements in the question for attention allocation, in this case, referees, players, and shirt. Distinguishing the referee from the players is complicated in itself, as it requires further background knowledge about sports. Next, the system has to make sense of prepositions like 'behind' to capture spatial relationships between objects or agents, in this case, to attend to one specific player. Finally, the system needs to visually attend to the task-relevant regions, distill the type of information required (shirt color), and produce the answer.

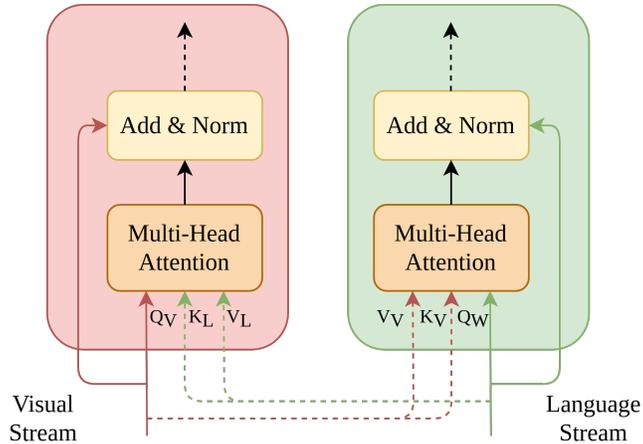


Figure 1: Co-attention transformer layer [Lu et al., 2019]

Recently, there has been an exciting trend of extending the successful transformer architecture [Vaswani et al., 2017] to solve multi-modal tasks combining modalities including text, audio, images, and videos [Chuang et al., 2019, Gabeur et al., 2020, Sun et al., 2019]. This trend has led to significant improvements in state-of-the-art models for Vision-Language tasks like visual grounding, referring expressions, and visual question answering. These families of models are based on either single-stream or two-stream architectures. The former shares the parameters across both modalities, while the latter has separate processing stacks for vision and language. In [Lu et al., 2019], Co-Attention Transformer Layers (Fig. 1) are used to facilitate interactions between the visual and language streams of the network. The task-relevant representations from the language stream modulate processing in the visual stream in the form of attention.

In this work, we assess the capabilities of co-attention transformer layers in guiding visual attention to task-relevant regions. We focus specifically on the Visual Question Answering task and conduct experiments to gain insight into the attention mechanisms of these layers and compare these mechanisms to human attention. Given an image/question pair, we generate attention maps for different co-attention layers based on the question-conditioned image attention scores and evaluate these maps against human attention maps, quantitatively and qualitatively, via rank-correlation and visualizations. We ask the following questions: 1) Does the use of object-based region proposals act as a bottleneck? 2) Is the model more likely to correctly answer a question when its attention map is better correlated to humans? 3) What is the role of question semantics in driving the model’s visual attention? 4) What is the importance of different parts of speech in guiding the model to attend to task-relevant regions? Our experiments demonstrate that object-based region proposals often restrict the model from focusing on task-relevant regions. We show that rank-correlation between human and machine attention is considerably higher in current state-of-the-art transformer-based architectures compared to previous CNN/LSTM networks. Lastly, we find that question semantics have little influence on the model’s visual attention, and only specific keywords in the question are responsible for driving attention.

2 Related Work

The Visual Question Answering (VQA) v1 dataset containing images from the MSCOCO dataset [Lin et al., 2014] with over 760K questions and 10M answers was introduced in [Antol et al., 2015], and a more balanced VQA v2 dataset was introduced in [Goyal et al., 2017]. The initial model for VQA [Antol et al., 2015] employed deep convolutional neural networks and recurrent neural networks to compute image and question representations separately. These were then fused using point-wise multiplication and fed to a Multi-Layer Perceptron (MLP) to predict the answer. Later, [Yang et al., 2016] proposed Stacked Attention Networks (SAN), in which the question representation from an LSTM was used for predicting an attention distribution over different parts of the image. Based on this attention and the question representation, another level of attention was performed over the image. The Hierarchical Co-Attention Model [Lu et al., 2016] introduced co-attention, where the

model attends to parts of the image along with parts of the question. Given a question about an image, this model hierarchically uses word-level, phrase-level, and question-level co-attention.

The VQA-HAT dataset consisting of human attention maps for question/image pairs from the VQA v1 dataset was introduced in [Das et al., 2016]. These maps were collected by asking humans to deblur different image regions by clicking on those regions to answer the question. Attention-based VQA models [Yang et al., 2016, Lu et al., 2016] based on convolutional neural networks and LSTM modules, but not transformer-based models, were compared against human attention maps [Das et al., 2016]. The authors concluded that these models did *not* attend to the same regions as humans while answering the question. However, increased performance was weakly associated with a better correlation between human and model attention maps. Later, [Goyal et al., 2016] used guided backpropagation and occlusion techniques to generate image importance maps for a VQA model and then compared those with human attention maps.

Various transformer-based VQA models [Li et al., 2020, Chen et al., 2020, Su et al., 2019, Li et al., 2019b,a, Zhou et al., 2019, Chefer et al., 2021] have been introduced in the last few years. Among them, [Tan and Bansal, 2019] and [Lu et al., 2019] are two-stream transformer architectures that use cross-attention layers and co-attention layers, respectively, to allow information exchange across modalities. There are several studies on the interpretability of VQA models [Goyal et al., 2016, Agrawal et al., 2016, Kafle and Kanan, 2017, Jabri et al., 2016], and yet very few have focused on the co-attention transformer layers used in recent VQA models. In this work, we use ViLBERT [Lu et al., 2019] for our study as it employs these co-attention layers.

3 Methods

We study the co-attention module between language and vision and the interactions within this module. To study co-attention in two-stream vision-language transformer architectures, we evaluated visual attention in the model by comparing it against human attention maps. ViLBERT [Lu et al., 2019] is an extension of the BERT architecture [Devlin et al., 2018] to process visual inputs. Given a question and an image, the model processes them separately in the language and visual streams, respectively. Both visual and language streams contain a stack of transformer and co-attention transformer layers. The embeddings for the word tokens and other special tokens are fed to the language stream after adding positional embeddings. The image is processed through the Faster RCNN network [Ren et al., 2016] to generate features for different region proposals. The feature representations of region proposals with the highest objectness score are fed to the visual stream. The model then processes these inputs through the two streams while fusing information within them using subsequent *co-attention layers* (Fig. 1).

3.1 Setup

The ViLBERT [Lu et al., 2019, 2020] network variant in our study uses the BERT_{BASE} model [Devlin et al., 2018] for the language part, composed of 12 transformer blocks. The latter 6 blocks have co-attention transformer modules stacked between them. The visual stream comprises 6 transformer and co-attention transformer modules. The co-attention transformer layer uses 8 parallel attention heads. All experiments were performed on a single NVIDIA 1080 Ti GPU. The source code will be publicly available upon publication.

3.2 Attention Map Generation

Given an image and a question, the inputs to the visual stream are the region features v_0, v_1, \dots, v_T and the input to the language stream are w_0, w_1, \dots, w_N . We generate an attention map for each co-attention transformer layer in the model as shown in Fig. 2. Inside the multi-head attention block in each co-attention transformer layer, the key and value matrices from one stream are projected onto another stream and vice versa. Consequently, inside the language stream, the multiplication of the Query matrix (Q_L) from the language stream and the Key matrix (K_V) from the visual stream produces attention scores over the different image regions based on the question. These attention scores are then passed through a softmax operation to generate respective attention probabilities

$$a_h^i = \text{softmax}\left(\frac{Q_L K_V^T}{\sqrt{d_k}}\right),$$

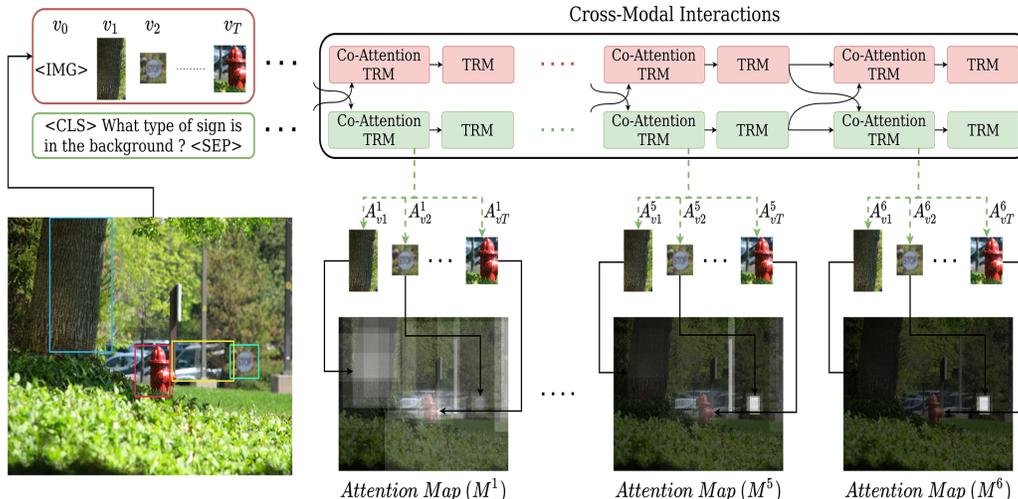


Figure 2: **Illustration of our attention map generation process.**

where i is the co-attention layer number, h is the attention head number, and $\sqrt{d_k}$ is a scaling factor [Vaswani et al., 2017]. These probabilities over the 8 attention heads capture the modulations from each text token to different image regions. To generate question-level attention maps, we first average these attention probabilities (before dropout) over all the attention heads and then across the words present in the question. This gives us attention data $\mathbf{A}^1, \dots, \mathbf{A}^6$ for the 6 co-attention layers, where $\mathbf{A}^i = \{A_{v_1}^i, \dots, A_{v_T}^i\}$. Based on the attention probability of different region proposal, i.e., $A_{v_1}^i, \dots, A_{v_T}^i$, we weigh the corresponding pixel intensities in an image matrix and then normalize this image matrix to get the final attention map over the image, conditioned on the question. We do this for all 6 co-attention layers to get attention maps M^1, \dots, M^6 .

3.3 Comparison Metric

We use rank-correlation (denoted by ρ in the visualization figures) to compare ViLBERT’s attention with human attention [Das et al., 2016]. Both attention maps are scaled to 14×14 and then flattened to get a 196 dimensional vector. These two vectors are then ranked based on their spatial attention and then we compute the correlation between the two rank vectors. All reported rank-correlation values except Question POS tag experiments (Sec. 4.3), show averages over 1,374 question/image pairs from the VQA-HAT [Das et al., 2016] validation set.

4 Experiments

4.1 Similarity to human attention shows a small dependence on the number of region proposals

We investigated the influence of the number of region proposals on the model’s ability to examine task-relevant regions. Since humans rely on context to solve a problem, we hypothesize that more region proposals bring in more task-relevant context from the image, thus increasing the rank-correlation of the model’s attention to that of humans and, in turn, increasing the answering accuracy. We show the rank-correlation of ViLBERT’s [Lu et al., 2019] attention maps with human attention maps across successive co-attention layers in **Fig. 3** for varying numbers of region proposals. To put results in perspective, we compare the results against an upper bound given by the rank-correlation for inter-human comparisons and a lower bound given by random attention allocation.

Increasing the number of region proposals led layers 3-6 of the model to attend to regions more similar to those attended by humans. The increased context due to more region proposals also improved the model’s VQA accuracy (**Table 1** and examples in **Fig. 4**). The region proposals are generated using Faster RCNN [Ren et al., 2016], an object detection architecture. Therefore, even in the first co-attention layer, which has little interaction with the language stream, the rank-correlation of the

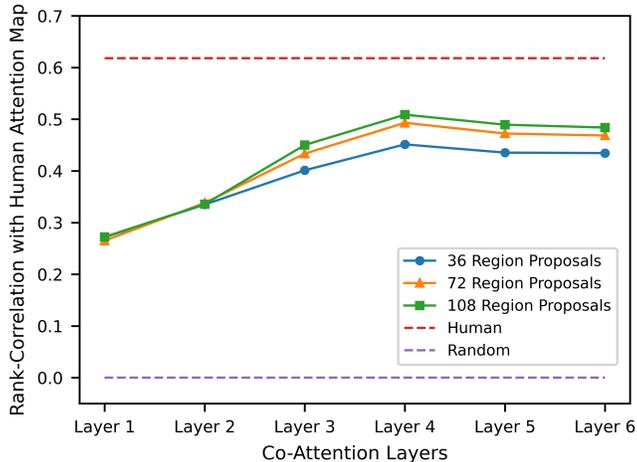


Figure 3: **The similarity between ViLBERT and human attention benefits from more region proposals.** The rank-correlation of ViLBERT’s [Lu et al., 2019] attention with human attention increases monotonically up to layer 4 (see section 4.1 for details). Error bars showing standard error of means are smaller than the symbol size in this plot.

Table 1: VQA accuracy of ViLBERT [Lu et al., 2019] with different number of region proposals. Accuracies are computed over all the question/image pairs in the VQA-HAT [Das et al., 2016] validation set.

Method	VQA Accuracy
ViLBERT [Lu et al., 2019] (36 Region Proposals)	76.57
ViLBERT [Lu et al., 2019] (72 Region Proposals)	79.39
ViLBERT [Lu et al., 2019] (108 Region Proposals)	80.83

model’s visual attention with human attention is well above chance. The correlation in the lower layers is likely due to the observation that the majority of the questions in the VQA dataset [Antol et al., 2015] focus either on object categories or object attributes that are salient in terms of basic visual features.

Given a fixed number of region proposals, the rank-correlation increases monotonically until layer 4 and then stays approximately constant. This initial increase validates the crucial role of co-attention layers in guiding visual attention in the model. Additionally, increasing the number of region proposals captures objects’ features using multiple aspect ratios and scales, often helping the model to better attend to the object in question, as depicted in **Fig. 4** (row 2).

4.2 Words matter more than grammar or semantics

Next, we evaluated the influence of question semantics in driving the visual attention mechanism. Given a question/image pair, we randomly shuffled the order of words in the question and then forward propagated the question and the image through the ViLBERT model [Lu et al., 2019]. For instance, a question like ‘What color is the floor?’ could become ‘Is color floor what the?’. The new question makes no semantic or grammatical sense. The shuffling procedure was done only at test time, while the model was trained with the words in the original order.

We expected that the rank-correlation of the model’s attention with human attention for these modified questions should drop along with the VQA accuracy. However, the results did not match our expectations (**Fig. 5**, and visualization examples in **Fig. 6**). There was only a minimal drop in the degree of similarity of the attention maps upon shuffling the word order. For example, in **Fig. 6** row 1, ‘‘What color is the floor?’’ led to the correct answer (brown) and $\rho = 0.548$ and the shuffled version

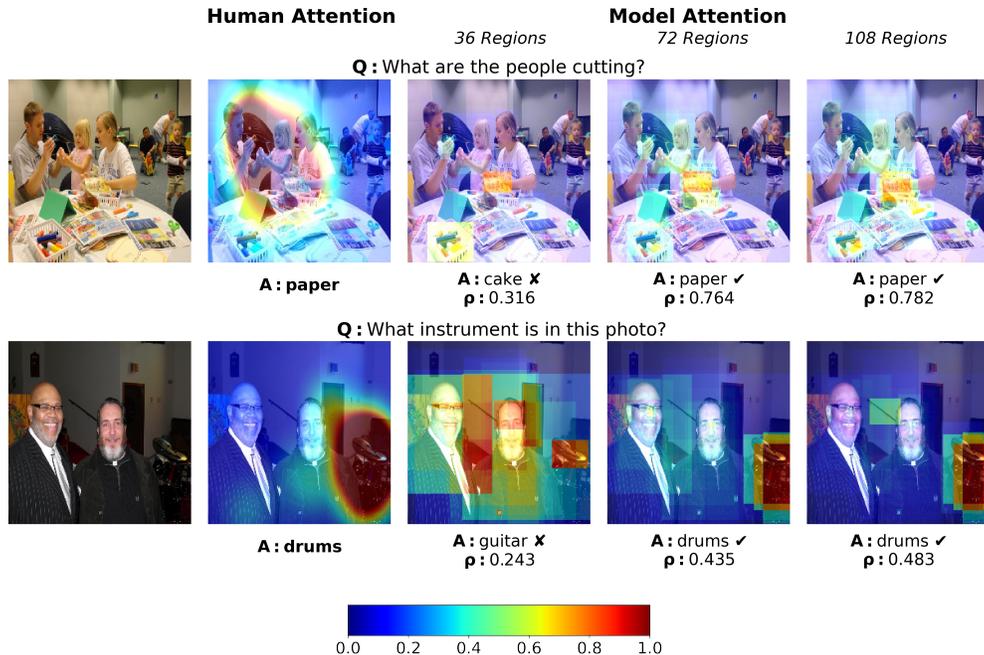


Figure 4: **Visualization for cases where increasing number of region proposals increases accuracy as well as rank-correlation with human attention.** The question and answers are shown above and below the images. Column 1: input image, Column 2: human attention map. Columns 3, 4, 5: ViLBERT’s [Lu et al., 2019] attention map for 36, 72, and 108 region proposals. The bottom colormap describes the intensity of the attention maps. Additional visualizations are provided in **Appendix A.1**.

“Is color floor what the?” also led to the correct answer and $\rho = 0.556$. These results suggest that the question grammar and semantics play little to no role in modulating visual attention. Instead, the presence of specific keywords in the question is responsible for driving attention. Most of the visual grounding here is based on object-centric concepts rather than the overall semantics of the question.

The model’s VQA accuracy dropped considerably after shuffling the words (**Table 2**). Thus, while attention seems to be largely independent of grammar and semantics, the ability to answer the questions correctly does require some notion of grammar and/or semantic information.

Table 2: VQA accuracy of ViLBERT [Lu et al., 2019] in different controls. Note that the reported accuracy is over question/image pairs in VQA-HAT [Das et al., 2016] validation set. Refer section 4.2 for more details.

Method	VQA Accuracy
ViLBERT [Lu et al., 2019] (Normal)	76.57
ViLBERT [Lu et al., 2019] (Shuffled Words)	60.2
ViLBERT [Lu et al., 2019] (Unrelated Question/Image Pair)	10.8

Given that attention was not dependent on the semantic content, we wondered whether it is possible that the model was focusing exclusively on visual information and simply ignoring the language part to drive attention allocation. To assess this possibility, we paired images with another randomly chosen question and compared the human attention maps with a given image/question pair and the model attention maps with the same image but a random question (**Fig. 5**). The rank-correlation in the case of Unrelated Question/Image Pair was largely driven by the visual input, any contribution from language in this case would be spurious.

Following the example in **Fig. 6**, row 1, the same image but using the question “Is this singles or doubles?” (instead of “What color is the floor?”), led to the erroneous answer “singles” and

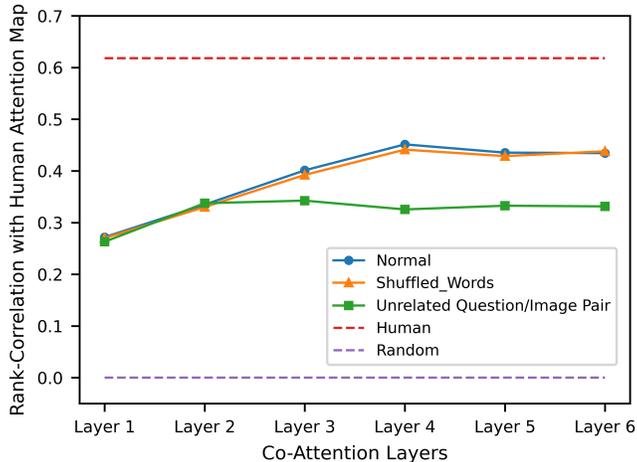


Figure 5: **The semantics of the question plays little role in driving the model’s attention map.** Similarity between model and human attention maps (ρ , using 36 region proposals) for each of the 6 co-attention layers for the default (normal) model (blue), for the shuffled words condition (orange), and a condition where the image is paired with a random question (green). The format is similar to **Fig. 3**, showing the between-human upper bound and the random levels. There is minimal change in ρ after shuffling the words, indicating that semantics has little influence on ViLBERT’s [Lu et al., 2019] attention.

$\rho = 0.02$ (cf. $\rho = 0.548$ for the correct question/image pair). The similarity with human attention was largely independent of the layer number but remained well above chance levels in the case of Unrelated Question/Image Pair (**Fig. 5**). Visual attention alone is sufficient to drive the rank-correlation with humans. Interestingly, even the unrelated question case shows higher similarity than previous benchmarks that combined visual and correct language information (**Table 3**). For layers 3-6, the similarity with human attention dropped considerably with respect to the correct question condition. Thus, attention is largely dictated by visual information, combined with focused co-attention driven by the presence of specific key words irrespective of their ordering.

4.3 Nouns drive attention

We quantified the importance of different parts of speech (POS) in guiding the model’s attention to task-relevant image regions. Given a question and the corresponding image, we dropped words with a certain POS tag. For example, the question “what is the girl holding?” would become “what is the holding?” upon removing nouns. Then, we forward propagated the image and the modified question through the network and generated the corresponding attention maps, and computed the rank-correlation with the human attention maps. Similar to [Goyal et al., 2016], we group POS tags into the following categories: Noun, Pronoun, Verb, Adjective, Preposition, Determiner, and Wh-Words. The Wh-Words category includes WP, WDT, and WRB tags containing words like who, which, and where respectively. We show the results of this experiment in **Fig. 7**, using 36 region proposals.

Consistent with our findings in Section 4.2 that words are more important than semantics, we noticed that nouns specifically played an important role in driving visual attention, followed by prepositions and pronouns. Given a question, nouns often help the model filter the relevant object categories from all the object region proposals. In addition, prepositions sometimes help guide attention based on spatial relationships between objects (see **Appendix A.3** for visualizations and additional qualitative results).

4.4 Better performing VQA models show higher correlation with human attention maps

In **Table 3**, we show the VQA accuracy and rank-correlation of the model’s attention maps and human attention maps for the following networks: ViLBERT [Lu et al., 2019], Stacked Attention

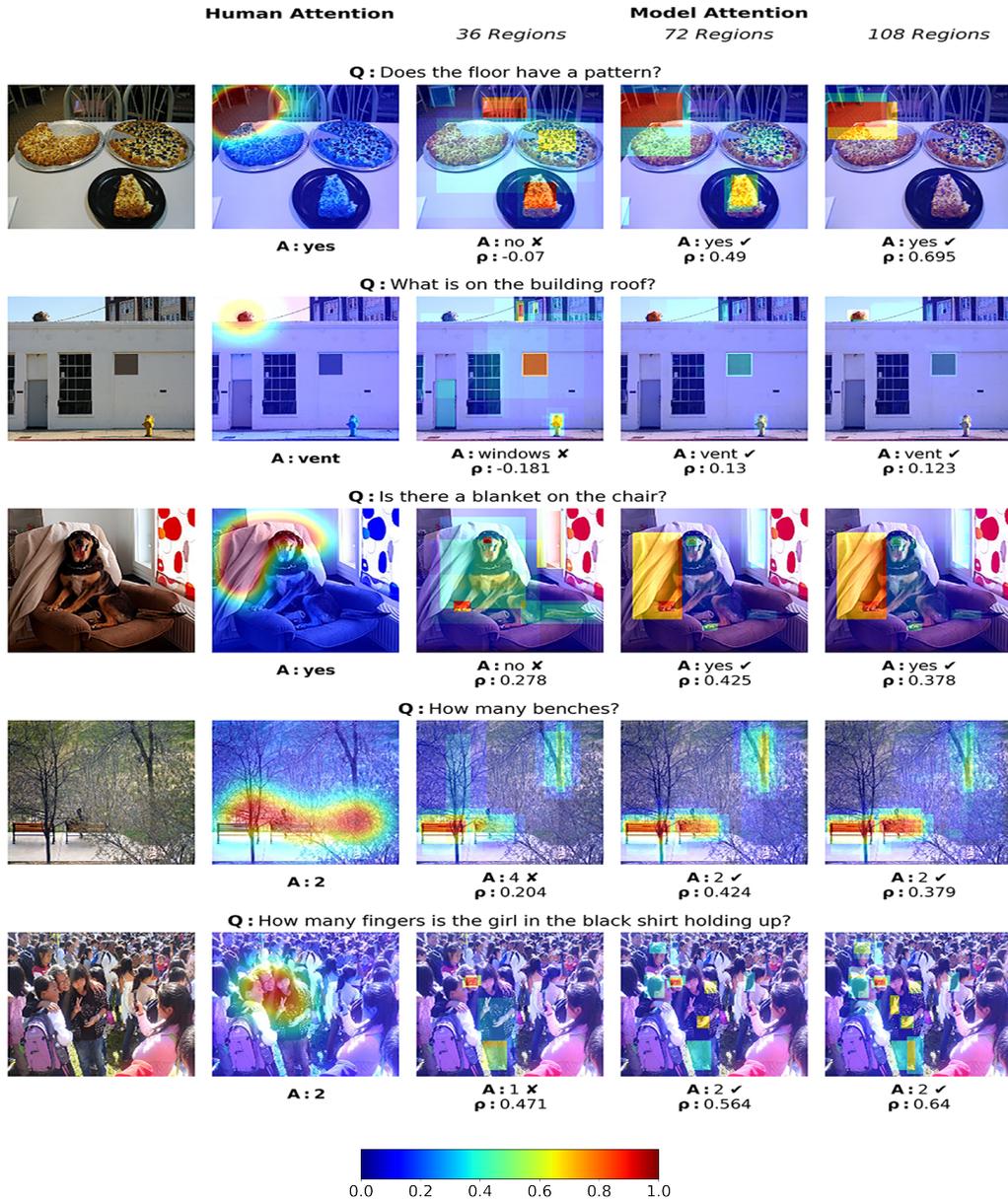


Figure 6: Visualization for different question/image pairs and their corresponding attention maps across multiple controls. Column 1 shows the input image, column 2 contains the human attention maps and Column 3, 4, and 5 show ViLBERT’s [Lu et al., 2019] attention map for **Normal**, **Shuffled_Words**, and **Unrelated Question/Image Pair** conditions, respectively. The answers in bold are ground-truth and the predicted answers are not in bold (see Appendix A.2 for extended analyses).

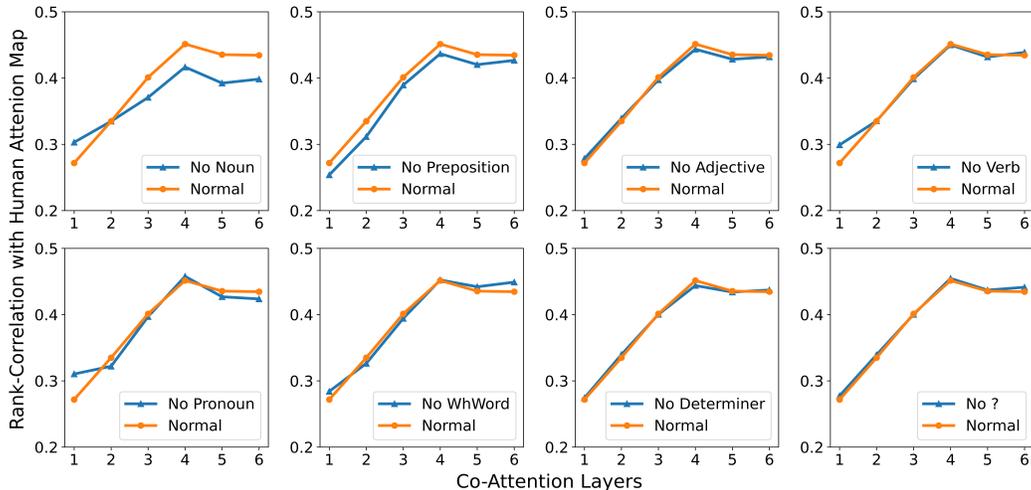


Figure 7: **Removing nouns, and to a lesser degree prepositions, led to a drop in similarity of attention maps.** Rank-correlation with human attention map (ρ) for each of the 6 co-attention layers upon removing different parts of speech (blue). The reduction in rank-correlation was maximal in the case of nouns, followed by prepositions and pronouns. Other parts of speech had little effect on the rank-correlation. Rank-correlation values shown here were averaged over question/image pairs containing words from the corresponding category (see Section 4.3 for details). Error bars showing standard error of means are smaller than the symbol size in this plot.

Network [Yang et al., 2016] with 2 attention layers (SAN-2), Hierarchical Co-Attention Network [Lu et al., 2016] with Word-Level (HieCoAtt-W), Phrase-Level (HieCoAtt-P), and Question-Level (HieCoAtt-Q). ViLBERT [Lu et al., 2019] uses a multi-modal transformer architecture while SAN-2 [Yang et al., 2016] and HieCoAtt [Lu et al., 2016] are based on CNN and LSTM architectures. The rank-correlation for the CNN/LSTM based models is considerably lower than the transformer-based model indicating a superior co-attention mechanism and better fusion of vision and language information in multi-modal transformers. Finally, it’s interesting also to note that an increase in the VQA accuracy is accompanied by a better correlation with human attention.

Table 3: Accuracy for different VQA models on the VQA test-std set as reported in [Yang et al., 2016, Lu et al., 2016, 2019]. Error bars in rank-correlation here show standard error of means.

Method	Rank-Correlation	VQA Accuracy
Random	0.000 ± 0.001	-
SAN-2 [Yang et al., 2016]	0.249 ± 0.004	58.9
HieCoAtt-W [Lu et al., 2016]	0.246 ± 0.004	
HieCoAtt-P [Lu et al., 2016]	0.256 ± 0.004	62.1
HieCoAtt-Q [Lu et al., 2016]	0.264 ± 0.004	
ViLBERT [Lu et al., 2019]	0.434 ± 0.006	70.92
Human	0.618 ± 0.006	-

5 Conclusion & Discussion

We conducted a series of experiments to interpret and study co-attention transformer layers and their role in aiding rich cross-modal interactions. We probed the modulation from language to vision in these co-attention layers and compared them with human attention maps. Transformer models lead to a substantial improvement in the similarity of attention maps with humans. In addition, the attention maps of VQA models with higher accuracy are better correlated with human attention maps

Interestingly, the overall question semantics play a minimal role in guiding visual attention. Attention is governed by the visual inputs and by the presence of key nouns in the question.

The interpretability of multi-modal transformers has received little attention, despite their notable success in terms of performance metrics. While we are enthusiastic about recent advancements in Vision-Language models, it is also critical and instructive to examine transformer layers carefully. We illustrate through visualizations the observation that the object-based region proposals often act as a bottleneck and prevent the network from looking at task-relevant regions. There remains a large gap in accuracy between state-of-the-art VQA models and human performance. At the same time, even though our results demonstrate that co-attention transformer layers yield a large boost to the congruency of attentional modulation in models and humans with respect to previous baselines, there is also a gap in the similarity of attention maps. We argue that this two gaps are related: building models that better capture human attention maps, perhaps by emphasizing the role of word combinations and semantics, can bring fundamental improvements in future VQA networks.

References

- Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *arXiv preprint arXiv:2103.15679*, 2021.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.
- Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-shan Lee. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *arXiv preprint arXiv:1910.11559*, 2019.
- Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision (ECCV)*, volume 5. Springer, 2020.
- Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*, 2016.
- 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.
- Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.
- Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, 2019a.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019b.

- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*, 2016.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.
- Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.
- Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.
- Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa, 2019.

A Appendix

A.1 Additional qualitative results

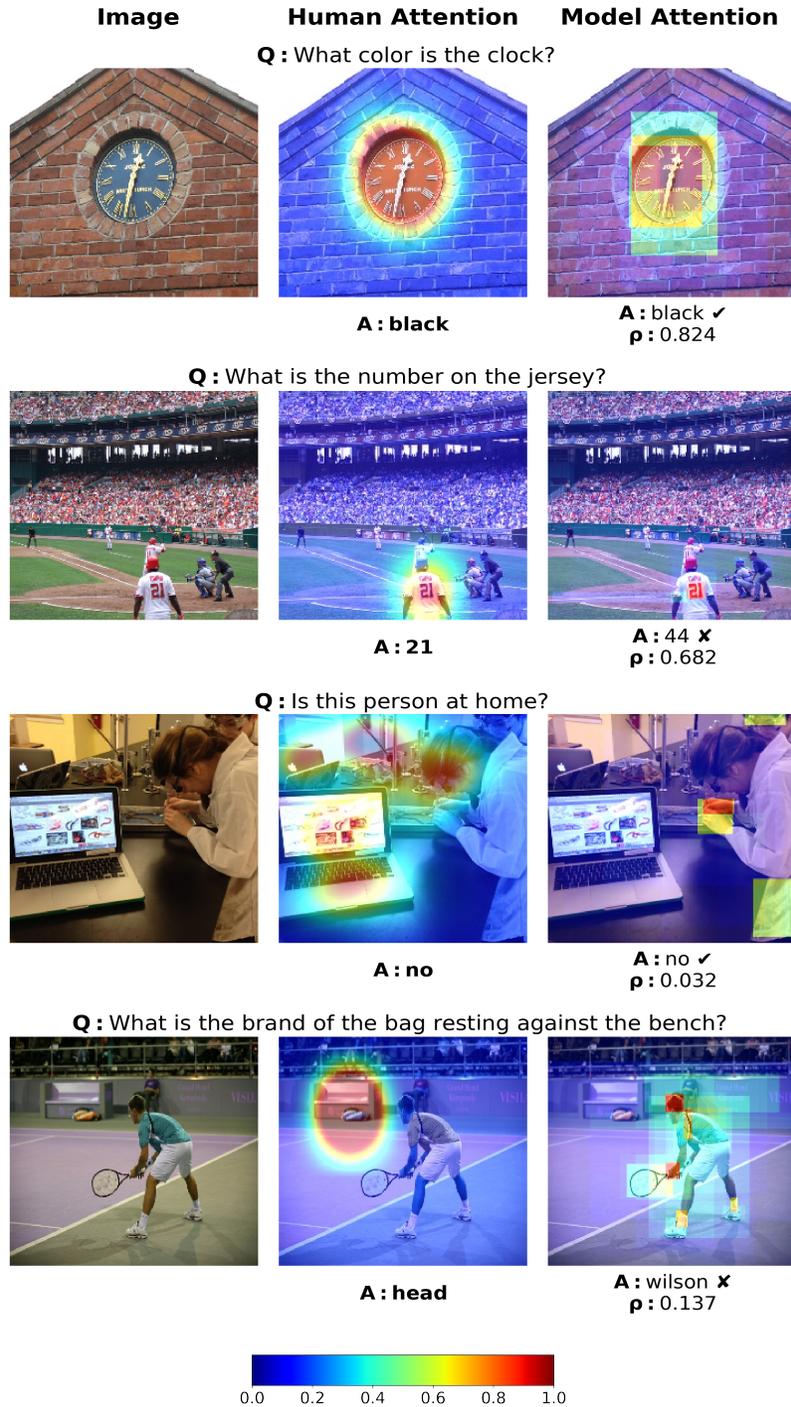


Figure 8: Row 1: high rank-correlation with 100% accuracy, Row 2: high rank-correlation with 0% accuracy, Row 3: low rank-correlation with 100% accuracy, Row 4: low rank-correlation with 0% accuracy. Column 1 shows the input image, column 2 contains the human attention maps, and column 3 shows ViLBERT’s attention map. The answers in bold are ground-truth and the predicted answers are not in bold.

A.2 Object region proposals act as a bottleneck

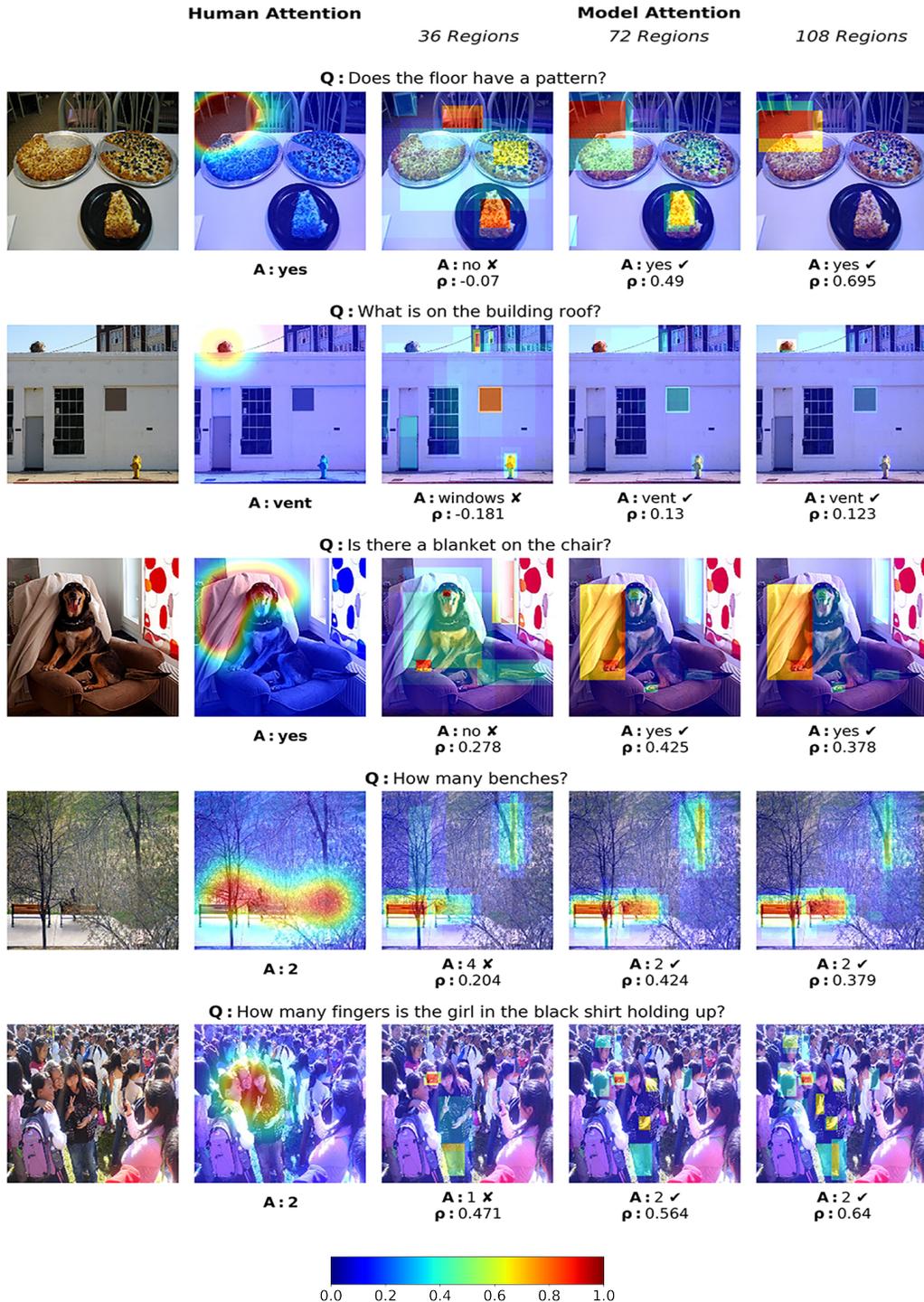


Figure 9: Visualization for cases where number of regions proposals act as a bottleneck and restrict the network from attending to task-relevant regions. Column 1 shows the input image, column 2 contains the human attention maps, and Column 3,4, and 5 show ViLBERT’s attention map for 36, 72, and 108 regions respectively. The answers in bold are ground-truth and the predicted answers are not in bold.

A.3 Question semantics play little role in visual attention



Figure 10: **Additional visualizations for different question/image pairs and their corresponding attention maps across multiple controls.** Column 1 shows the input image, column 2 contains the human attention maps, and Column 3, 4, and 5 show ViLBERT’s attention map for **Normal**, **Shuffled_Words**, and **Unrelated Question/Image Pair** conditions, respectively. The answers in bold are ground-truth and the predicted answers are not in bold.

A.4 Importance of certain POS tags in guiding model’s attention

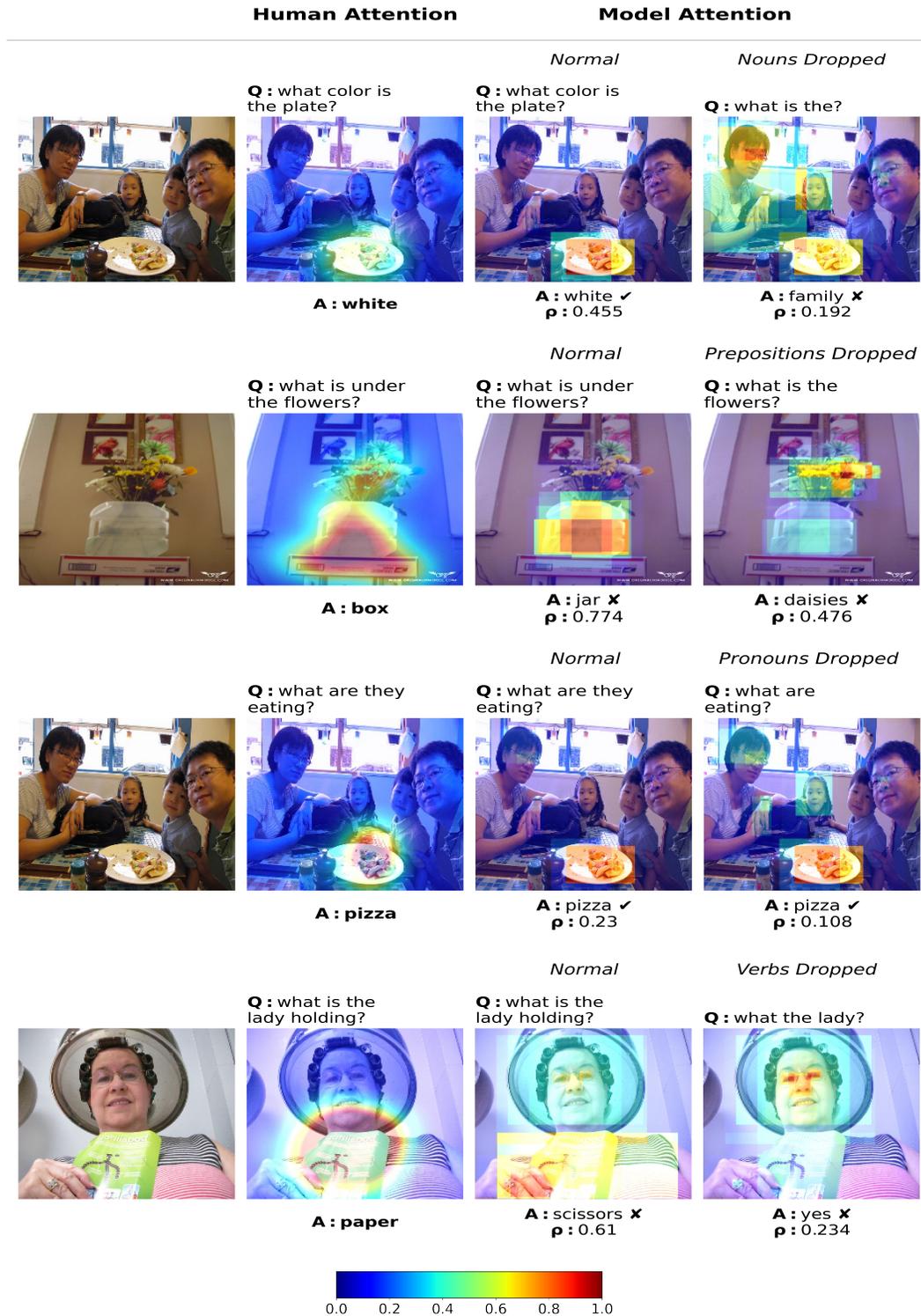


Figure 11: Visualization for different question/image pairs and their corresponding attention maps after dropping words with certain POS tags. Row 1: Nouns dropped, Row 2: Prepositions dropped, Row 3: Pronouns dropped, Row 4: Verbs dropped. The answers in bold are ground-truth and the predicted answers are not in bold.