VQA-MHUG: A Gaze Dataset to Study Multimodal Neural Attention in Visual Question Answering

Ekta Sood[1], **Fabian Kögel**[1], **Florian Strohm**[1], **Prajit Dhar**[2], **Andreas Bulling**[1]
[1]University of Stuttgart, Institute for Visualization and Interactive Systems (VIS), Germany
[2]University of Groningen, Center for Language and Cognition (CLCG), the Netherlands
{ekta.sood,fabian.koegel,florian.strohm,andreas.bulling}@vis.uni-stuttgart.de
p.dhar@rug.nl

Abstract

We present VQA-MHUG – a novel 49-participant dataset of multimodal human gaze on both images and questions during visual question answering (VQA) collected using a high-speed eye tracker. We use our dataset to analyze the similarity between human and neural attentive strategies learned by five state-of-the-art VQA models: Modular Co-Attention Network (MCAN) with either grid or region features, Pythia, Bilinear Attention Network (BAN), and the Multimodal Factorized Bilinear Pooling Network (MFB). While prior work has focused on studying the image modality, our analyses show – for the first time – that for all models, higher correlation with human attention on text is a significant predictor of VQA performance. This finding points at a potential for improving VQA performance and, at the same time, calls for further research on neural text attention mechanisms and their integration into architectures for vision and language tasks, including but potentially also beyond VQA.

1 Introduction

Visual question answering (VQA) has gained popularity as a practically-useful and challenging task at the intersection of natural language processing (NLP) and computer vision (CV) (Antol et al., 2015). The key challenge in VQA is to develop computational models that are able to reason over questions and images in order to generate answers that are well-grounded in both modalities (P. Zhang et al., 2015; Agrawal et al., 2016; Goyal et al., 2017a; Kafle et al., 2019). Attention mechanisms originally introduced in NLP for monomodal language tasks have been successfully applied to multimodal tasks (like VQA) and established a new state of the art (Correia and Colombini, 2021; Kim et al., 2018; Yu et al., 2019b).

These advances have, in turn, triggered research into understanding the reasons for these improvements. A body of work has studied similarities between neural and human attention (Qiuxia et al., 2020; Yun et al., 2013; Das et al., 2016). Models seem to learn very different attention strategies and similarity to human attention might only improve performance for specific model types (Sood et al., 2020a). However, although VQA is an inherently multimodal task, all of these analyses have only focused on image attention. The most likely reason for this is that existing datasets only offer monomodal attention on the image (Das et al., 2016; Fosco et al., 2020; Chen et al., 2020). In addition, due to the challenges involved in recording human gaze data at scale, prior works have instead used mouse data as a proxy to attention (Jiang et al., 2015). However, mouse data was shown to overestimate some image areas (Tavakoli et al., 2017b; Das et al., 2016) or to miss relevant background information altogether (Sugano and Bulling, 2016; Tavakoli et al., 2017a). As of now, there is no publicly available dataset that offers human gaze data on both the images and questions. This severely impedes further progress in this emerging area of research.

Our work fills this gap by introducing VQA-MHUG – the first dataset of multimodal human gaze on both images and questions in VQA. To collect our dataset, we conducted a 49-participant eye tracking study. We used a commercial, high-speed eye tracker to record gaze data on images and corresponding questions of the VQAv2 validation set. VQA-MHUG contains 11,970 gaze samples for 3,990 question-image pairs, tagged and balanced by reasoning type and difficulty. We ensured a large overlap in question-image pairs with nine other VQA datasets to maximize the usefulness of VQA-MHUG for future multimodal studies on human and neural attention mechanisms. Using our dataset, we conduct detailed analyses of the similarity between human and neural attentive strategies, the latter of which we obtained from five top-

performing models in the VQA challenges 2017-2020: Modulated Co-Attention Network (MCAN) with grid or region features, Pythia, Bilinear Attention Network (BAN), and the Multimodal Factorized Bilinear Pooling Network (MFB). These analyses show, for the first time, that correlation with human attention on text is a significant predictor of accuracy for all the studied state-of-the-art VQA models. This suggests a potential for significant performance improvements in VQA by guiding models to "read the questions" more similarly to humans. In summary, our work contributes:

1. VQA-MHUG, a novel 49-participant dataset of multimodal human gaze on both *images* and *questions* during visual question answering collected using a high-speed eye tracker.
2. Detailed analysis of the similarity between human and neural attentive strategies indicating that human-like attention to text could yield significant performance improvements.

2 Related Work

Our work is related to previous work on 1) neural machine attention, 2) attention in VQA, and 3) comparison of neural and human attention.

Neural Machine Attention. Inspired by the human visual system, neural machine attention allows neural networks to selectively focus on particular parts of the input, resulting in significant improvements in performance and interpretability (Correia and Colombini, 2021). Single-modal attention (Bahdanau et al., 2014) as well as approaches that build on it, such as self attention (Xu et al., 2015; Vaswani et al., 2017) or stacked attention (Yang et al., 2016a,b; Zhang et al., 2018; Anderson et al., 2018), have been shown to be particularly helpful for sequence learning tasks in NLP and CV. Initially, attention mechanisms were often combined with recurrent and convolutional architectures to encode the input features (Bahdanau et al., 2014; Yu et al., 2017; Tavakoli et al., 2017b; Kim et al., 2016; Lu et al., 2016; Jabri et al., 2016; Agrawal et al., 2016). More recently, Transformer-based architectures have been introduced that solely rely on attention (Vaswani et al., 2017; Yu et al., 2019b; Khan et al., 2020). Large-scale, pre-trained language models are a key application of Transformers that enabled their current performance lead in both NLP and multimodal vision-language tasks (Devlin et al., 2018; Yang et al., 2019b; Yu et al., 2019b; Lu et al., 2019).

Attention in VQA. Increased interest into capturing multimodal relationships with attention mechanisms have put focus on VQA as a benchmark task (Malinowski and Fritz, 2014; Malinowski et al., 2015; Lu et al., 2016; Yu et al., 2017; Nguyen and Okatani, 2018; Yang et al., 2019a; Li et al., 2019). In fact, attention mechanisms have been extensively explored in VQA and have repeatedly dominated the important VQAv2 challenge (Anderson et al., 2018; Yu et al., 2019b; Jiang et al., 2020). Alhough attention-based models have achieved remarkable success, it often remains unclear how and why different attention mechanisms actually work (Jain and Wallace, 2019; Serrano and Smith, 2019).

Comparing Neural and Human Attention. Several prior works have proposed datasets of human attention on images to study the differences between neural and human attention in VQA (Das et al., 2016; Fosco et al., 2020; Chen et al., 2020). In particular, free-viewing and task-specific mouse tracking from SALICON (Jiang et al., 2015) and VQA-HAT (Das et al., 2016), as well as free-viewing and task-specific gaze data from SBU Gaze (Yun et al., 2015) and AiR-D (Chen et al., 2020) have been compared to neural attention. All of these works were limited to images only and found mouse tracking to overestimate relevant areas and miss scene context (Sugano and Bulling, 2016; Tavakoli et al., 2017b,a; He et al., 2019). Furthermore, while integrating human attention over the image showed performance improvements in VQA (Park et al., 2018; Qiao et al., 2018; Chen et al., 2020), the influence of integrating human text attention remains unclear.

There is currently no multimodal dataset including real human gaze on VQA questions and images. This represents a major limitation for two different aspects of research, i.e. research aiming to better understand and improve neural attention mechanisms and research focusing on integrating human attention to improve VQA performance.

3 The VQA-MHUG Dataset

We present Visual Question Answering with Multi-Modal Human Gaze (VQA-MHUG)[1]. To the best of our knowledge, this is the first resource containing multimodal human gaze data over a textual

[1]The dataset is publicly available at https://perceptualui.org/publications/sood21_conll/

question and the corresponding image. Our corpus encompasses task-specific gaze on a subset of the benchmark dataset VQAv2 val[2] (Goyal et al., 2017b). We specifically focused on question-image pairs that machines struggle with, but humans answer easily (determined by high inter-agreement and confidence in the VQAv2 annotations). We then balanced the selection by evenly picking questions based on a machine difficulty score and from different reasoning types. Thus, VQA-MHUG covers a wide range of challenging reasoning capabilities and overlaps with many VQAv2-related datasets (see Table 4 in Appendix A).

Reasoning Types. VQAv2 groups question-image pairs based on question words: *what*, *who*, *how*, *when* and *where*. Instead, we binned our pairs into the reasoning capabilities required to answer them. We incorporated the categories proposed by Kafle and Kanan (2017) for their task directed image understanding challenge (TDIUC) and extended them with an additional category, *reading*, for questions that are answered by reading text on the images. This resulted in 12 reasoning types that align better with commonly-diagnosed error cases[3]. We binned VQAv2 val pairs accordingly by training a LSTM-based classifier on 1.6 M TDIUC and 145 K VQAv2 train+val samples which we labelled using regular expressions. The classifier predicted the reasoning type for a given question-answer pair. The final model achieved 99.67% accuracy on a 20% held-out test set.

Machine Difficulty Score. To assess the difficulty for a machine to answer a question-image pair, we ran two popular VQA models – MFB (Yu et al., 2017) for multimodal fusion and MCAN (Yu et al., 2019b) for transformer attention – inspired by Sood et al. (2020a). A difficult question results in low answer accuracy, particularly after rephrasing or asking further control questions. To test this, we evaluated on four datasets and averaged their corresponding normalized metrics: (1) VQAv2 accuracy, (2) VQA-CP accuracy on reduced bias (Agrawal et al., 2018a), (3) VQA-Introspect's consistency with respect to visual perception (Selvaraju et al., 2020a) and (4) VQA-Rephrasings' robustness against linguistic variations (Shah et al., 2019a) (see Appendix C).

[2] https://visualqa.org/download.html

[3] See Appendix B for details on the reasoning type tagging.

Participants and Experimental Setup. We recruited 49 participants at the local university (18 identified female and 31 male) with normal or corrected-to-normal vision, aged between 19 and 35 years ($\mu = 25.8$, $\sigma = 2.8$) and compensated them for their participation[4]. All participants had an English Level of C1 or above (8 were native speakers).[5]

Questions and images were presented one after each other on a 24.5" monitor with resolution 1920x1080 px. They were centered on a white background, and scaled/line-wrapped to fit 26.2x11.5° of visual angle in the center. For the questions, we used a monospace font of size 0.6° and line spacing such that the bounding boxes around each word covered 1.8° vertically. Binocular gaze data was collected with an EyeLink 1000 Plus remote eye tracker at 2 kHz with an average measured tracking error of 0.62° (see Appendix E).

Participants had unlimited viewing time but were instructed to move on as soon as they understood the question, gave an answer, or decided to skip. They completed a set of practice recordings to familiarize themselves with the study procedure. As such, the task was known to the participant, so both the question reading and the subsequent image viewing were conditioned on VQA. They then completed three blocks of 110 recordings in randomized order with 5 minute breaks in-between.

Dataset Statistics. VQA-MHUG contains gaze on 3,990 stimuli from VQAv2 val.For each stimulus, we provide three recordings from different participants over text and image, their corresponding answer, and whether they answered the question correctly (as compared to the VQAv2 annotations). For 3,177 stimuli (79.6%), the majority of participant answers appear in the VQAv2 annotations.

Human Attention Maps. To generate human attention maps, we used the fixation detection algorithm of the EyeLink software with default parameters. We always picked the eye with the lower validation error to prioritize accuracy (Hooge et al., 2019) and represented fixations by Gaussian kernels with $\sigma = 1°$. For our experiments, we assumed that the majority of gaze is valid and averaged the three recordings per stimulus, yielding a single attention map.

[4] The university ethics committee approved our study.

[5] After providing their consent, we collected basic demographic information for each participant. The anonymized data is available with the dataset.

(a) collection study

(b) validation study

Figure 1: Example images for the question "How ripe are the bananas?". Validation images (b) were masked using the attention maps from our VQA-MHUG dataset.

Dataset Validation. To validate that the attention maps indeed contain relevant image regions, we masked 300 stimuli with our recorded VQA-MHUG maps (see Figure 1b). Then, we showed two additional participants these masked stimuli. Comparing their answer accuracy with the participants who saw the full images, validation participants achieved comparable accuracy (62.43% vs. 63.87% in the main study). Therefore our VQA-MHUG maps contain sufficient image areas to answer the questions and mask distracting objects as illustrated in Figure 1.

Comparison to Related Datasets. We further measured the center bias and compared VQA-MHUG to related human attention datasets (Jiang et al., 2015; Das et al., 2016; Chen et al., 2020) on their overlapping samples. All datasets use mouse tracking as a proxy to collect human attention, except for the eye-tracking dataset AiR-D (Chen et al., 2020) which is similar to our recording paradigm, yet has no overlap with VQAv2. Therefore, we showed participants 195 additional stimuli from the AiR-D dataset for comparison. Table 1 shows the mean rank correlation of VQA-MHUG with a synthetic center fixation, inter-participant, and the other datasets. The high correlation between VQA-MHUG and AiR-D indicates that our data is of comparable quality. Our center bias is smaller compared to AiR-D but, as expected from human eye behaviour (Tatler, 2007), larger than in the mouse tracking proxies SALICON and VQA-HAT. We observe that both mouse tracking datasets have significantly lower correlation with VQA-MHUG than the eye-tracking AiR-D corpus.

4 Comparison of Human and Machine Attention

The collected data enabled us to analyze whether models achieve a higher accuracy on VQAv2 val the more their attentive behavior over the text and image correlates with human ground-truth attention. Hence, we investigated the attention weights over text and image features of different SOTA VQA models.

4.1 VQA Models

We selected five top performing VQA models of the VQA challenges 2017 to 2020:

- MFB (Yu et al., 2017) (Runner-up 2017);
- BAN (Kim et al., 2018) (Runner-up 2018);
- Pythia v0.1 (Jiang et al., 2018) (Winner 2018);
- MCAN_R with region image features (Yu et al., 2019b) (Winner 2019);
- MCAN_G with grid image features (Jiang et al., 2020) (Winner 2020).

Instead of using the text and image features directly for classification, these models re-weight the features using linear, bilinear and Transformer (Vaswani et al., 2017) (co-)attention mechanisms, whose attention maps we extracted and compared to human ground-truths from VQA-MHUG.

Pythia and MFB use co-attention: they first use a projected attention map to re-weight text features, then fuse them with the image features using linear (Pythia) and bilinear (MFB) fusion and subsequently re-weight the image features using an attention map projected from the fused features. In this way, the text attention influences the image attention. BAN avoids separating the attention into text and image streams and reduces both input streams simultaneously with a bilinear attention map projected from the fused features. Finally, MCAN as a Transformer model stacks co-attention modules with multi-headed scaled dot-product attention for each modality. After the last Transformer layer in both the text and image stream, another attention map is used to project the feature matrix into a single feature vector.

4.2 Extracting Model Attention

We used an official implementation[6] of the Pythia v0.1 architecture and the OpenVQA[7] implementations (Yu et al., 2019a) for MFB, BAN and MCAN. We re-implemented the grid image feature loader for MCAN_G, since it is not available in OpenVQA.

Following previous work (Sood et al., 2020a), we trained each network architecture twelve times with random seeds on the VQAv2 training set and

[6] https://github.com/zwxalgorithm/pythia

[7] https://github.com/MILVLG/openvqa

		VQA-MHUG	Center Fixation
Dataset	**Method**	$\rho \uparrow$	$\rho \uparrow$
VQA-MHUG	G	0.769 ± 0.079	0.473 ± 0.049
AiR-D	G	0.710 ± 0.060	0.523
VQA-HAT	M	0.612 ± 0.145	0.339 ± 0.107
SALICON	M	0.634 ± 0.063	0.479

G: Gaze, **M**: Mouse-Tracking

Table 1: Spearman's rank correlation (ρ) of VQA-MHUG with itself (inter-participant), related datasets, and a synthetic center fixation – Mean over all samples in the intersection of the datasets and three VQA-MHUG participants. The standard deviation is the mean error over participants. Only VQA-HAT and VQA-MHUG provide multiple attention maps per sample, allowing us to calculate the standard deviation when comparing to the synthetic center fixation.

then chose the top nine models based on the validation accuracy.

For models based on region image features, we used the extracted features provided by Anderson et al. 2018, while we trained MCAN_G with ResNeXt (Xie et al., 2017) grid features as provided by the authors (Jiang et al., 2020)[8].

For MFB and Pythia we extracted the two projected attention maps that re-weight text and image features, while we extracted the single bilinear attention map for BAN. To obtain separate attention maps for text and image from BAN's bilinear attention map, we marginalized over each dimension as suggested by the authors (Kim et al., 2018). MFB, BAN and Pythia generate multiple such attention maps called "glimpses" by using multiple projections. We averaged the glimpses after extraction, yielding a single attention map for each modality. Since it is unclear how the Transformer layer weights relate to the original input features, we instead extracted the attention weights of the final projection layer in text and image streams for MCAN_R and MCAN_G.

The extracted image attention maps contain one weight per feature. To compare them with the human spatial attention maps collected in VQA-MHUG, we mapped the features back to their source region in the image. For region-based features we assigned the attention weights to the corresponding bounding box normalized by region size. Analogously, for grid-based features, we mapped the attention weights to their corresponding grid cells. The text attention vector was directly mapped back to the question token sequence. We excluded 74 samples due to varied tokenization between models.

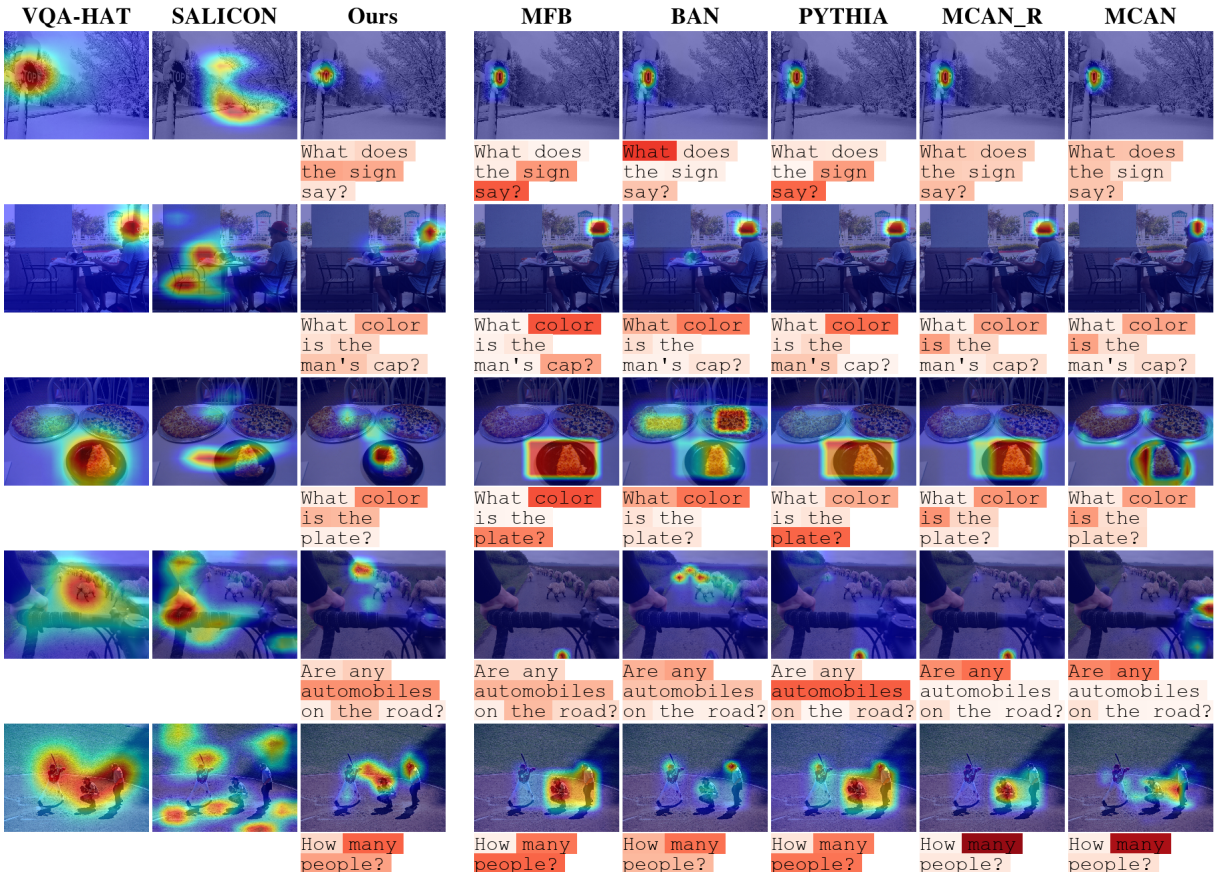
[8] https://github.com/facebookresearch/grid-feats-vqa

4.3 Performance Metrics

We compared the multimodal attention extracted from five models to our human data in VQA-MHUG using three approaches. We used Spearman's rank correlation to compare importance ranking of image regions and words, Jensen-Shannon divergence to compare the distance between the human, and neural attention distributions and a regression model to study the suitability of text and image correlation as predictors of per document model accuracy.

Spearman's rank correlation and Jensen-Shannon divergence. Similar to prior work, we downsampled all attention maps to 14x14 matrices and calculated the mean *Spearman's rank correlation* ρ (Das et al., 2016) and *Jensen-Shannon divergence* (JSD) (Sood et al., 2020a,b) between the neural attention and the corresponding human attention. We computed both metrics for both image and text modalities. We also evaluated the corresponding accuracy scores on the VQAv2 validation set (Agrawal, 2015).

Ordinal Logistic Regression. Averaging correlation over the whole dataset is too coarse and obscures the impact that similarity to human attention has on accuracy. Additionally, rank correlation does not allow to analyze the effect of two independent variables on a dependent variable (Bewick et al., 2003), e.g. image and text attention correlation on accuracy. To account for this and to study on a per document basis which modality factors influence the likelihood of a model to predict the



(a) comparison to other attention datasets

(b) model attention - text, image, and inter-modal comparison

Figure 2: Attention maps visualized across question types. Image attention seems mostly plausible throughout models. Previous datasets lack attention on the questions, but we reveal now that text attention is not always human-like, nor plausible. Mouse tracking datasets, SALICON and VQA-HAT, seem to over-estimate the relevant areas.

answer correctly, we performed an *Ordinal Linear Regression* (OLR).

The official VQAv2 evaluation (Agrawal, 2015) score per document is based on agreement with ten human annotator answers, where each match increases the score by 0.3 (capped at 1.0 or 4 agreed answers). Since our response variable (accuracy score) is not necessarily ordered equidistant, we binned accuracy scores for each document into a likelihood scale (accuracy correctness).

The model predicts the likelihood of accuracy correctness for each document with three different predictors — the text correlation (x), the image correlation (y), and the interaction between the text and image correlation (z). The latter we deem inter-modal correlation predictor, as it allows us to test if the interaction between the correlation of text and image impacts accuracy. Given that the dependant variables are ranked we opted for using ordered logistic regression to predict for each accuracy bin.

5 Results

5.1 Human and Neural Attention Relationship – Averaged Over Documents

Table 2 shows the overall accuracy scores of the five models on the VQAv2 validation set when trained only on the training partition. The models improved over the challenge years – MCAN grid is the current SOTA (Jiang et al., 2020). For each model and modality, we report the Spearman's rank correlation and JSD scores averaged over the entire VQA-MHUG corpus (cf. Section 4.3). All figures were averaged over nine model runs and the standard deviation is given over those instances. Given that one cannot average p-values we used a paired t-test to check if the differences in correlation and JSD per document and between models were statistically significant at $p < 0.05$ (see Appendix D).

Image attention. Models using region features, i.e. excluding MCAN_G, are more correlated with human visual attention on images. MCAN_R achieves the highest correlation, MFB the lowest, and the general trend shows that models with higher correlation had higher overall validation accuracy. Although MCAN_G achieves the highest accuracy, it had the lowest correlation with human image attention. For all model types, the difference between image correlation scores is significant, except between Pythia and BAN (see Appendix D). With respect to the JSD, we observed similar patterns except for the Pythia model, which was more dissimilar to human attention (had a higher overall JSD) compared to BAN. For all model types, the difference between image JSD scores was statistically significant (see Appendix D).

Text attention. Both the correlation and JSD scores indicate that Pythia is the most similar to human text attention, followed by MFB.Models with higher overall accuracy do not have high similarity to human visual attention over text on the JSD and correlation metrics. For both metrics, the difference in text attention between every model pairing is statistically significant, except for the JSD scores between pairings of BAN, MCAN_G, and MCAN_R (see Appendix D).

5.2 Ordinal Logistic Regression

By averaging evaluation metrics (correlation and JSD) across documents, we obscure the impact that similarity has *on each document* with respect to accuracy. The Ordinal Logistic Regression model results uncover the importance of the text and image correlation scores as predictors on per document accuracy.

Text Correlation. We show (cf. Table 3) **for all five different VQA models**, that as the correlation to human text attention decreases, the likelihood that the models will be able to correctly predict the answer significantly decreases/ Our findings show that correlation to human text attention is a significant predictor on accuracy. The MCAN_G, MCAN_R, and MFB model have the strongest relationship ($p < 0.001$) to text correlation being a significant predictor on accuracy. This indicates that for these models in particular, the less the model is correlated with human text attention, the less likely the model will predict the answer correctly.

Image Correlation. Interestingly, we observe the same trend as text correlation, in which image attention correlation is also a significant predictor on accuracy, but **not consistently across all models**.It is a significant predictor for three (MCAN_G, Pythia, and BAN) out of the five total models. Notably, the MCAN_G model has a significantly strong relationship to image correlation. This indicates that when the Pythia, BAN, and in particular MCAN_G learn attention which is less correlated to human image attention, then the model is more less likely to be able to predict the answer correctly.

Inter-Modal Correlation. We paired the text correlation x and the image correlation y together as an inter-modal predictor z. Inter-modal correlation tests whether the interaction between the two correlation scores, as the predictor z, has an effect on accuracy. Interestingly, inter-modal correlation z is a significant predictor on accuracy for the MCAN_G and Pythia models but not for the other 3 model types.

Attention Maps – Qualitative Analysis. Figure 2 visualizes the human as well as neural attention distributions of five VQA models for a selection of examples from the benchmark VQAv2 dataset.[9] As can be seen, all previous datasets only uncover the differences between human and neural image attention, while VQA-MHUG (ours) allows for studying multimodal neural VQA models attention. We also find our attention maps to be highly relevant and confirm that the mouse tracking datasets SALICON and VQA-HAT seem to over-estimate relevant areas. As the AiR-D dataset does not overlap with VQAv2, we separately visualize a selection of examples (see Section 3 overlapping with our VQA-MHUG data (see Appendix F).

5.3 Discussion

When averaging metrics across all documents in VQA-MHUG, our results regarding similarity between machine and human image attention and performance follow insights derived from previous work (Das et al., 2016), where they observed that as the models improved with respect to accuracy they were also more correlated to human attention on the images. However, notably we only observe this trend with the models which use region features. That is, though the MCAN grid is the highest performing model with respect to accuracy, it is also

[9] See Appendix G for additional examples.

		Image		Text	
Model	**Accuracy**	$\rho \uparrow$	**JSD** $\downarrow$	$\rho \uparrow$	**JSD** $\downarrow$
MCAN_G	70.24%	0.509 ± 0.026	0.537 ± 0.003	-0.059 ± 0.012	0.402 ± 0.007
MCAN_R	67.24%	0.602 ± 0.003	0.467 ± 0.002	-0.042 ± 0.018	0.398 ± 0.017
Pythia	66.00%	0.584 ± 0.003	0.479 ± 0.001	0.251 ± 0.016	0.337 ± 0.015
BAN	65.91%	0.582 ± 0.004	0.469 ± 0.002	-0.132 ± 0.030	0.398 ± 0.021
MFB	65.06%	0.530 ± 0.003	0.523 ± 0.004	0.225 ± 0.055	0.352 ± 0.011

Table 2: Accuracy of the five models as well as the Spearman's rank correlation (ρ) and the Jensen–Shannon divergence (*JSD*) between neural and human attention over images (left) and text (right). Standard deviation was calculated over nine model runs and indicates the attention variability between different instances of the same architecture. All correlation and JSD scores between models differ significantly (p<0.05), except for the image correlation between Pythia and BAN as well as the JSD text scores between BAN, MCAN_G and MCAN_R

the model which is least similar to human image attention. Such an observation was also reported in previous work which compared the XLNet transformer to human attention (Sood et al., 2020a).

Analysis from the Ordinal Logistic Regression model shows, for the first time, that correlation to human text attention is a significant predictor across all VQA model types, where dissimilarity between human and neural text attention decreases the likelihood of the models ability to predict the answer correctly. We conclude that striving to enhance neural attention to more similarly emulate human attention on text will improve performance in the five VQA models. As can be observed in Figure 2, text attention is not always human-like, especially for the otherwise high performing MCAN models, suggesting that increased similarity to human text attention might lead to further improvements with respect to accuracy.

Due to the lack of human attention data over text, researchers were not able to uncover the limitations or relevance of high correlation to human text attention on VQA model accuracy. In addition, our analysis on the role of image attention and inter-modal attention as a predictor on accuracy indicates that for certain model types it would be beneficial to improve image and inter-modal correlation. These findings are consistent with Sood et al. (2020b) which found that different model types learn different attention strategies and similarity of machine to human attention does not guarantee best performance. This may be due to factors such as features used (grid versus region), the different learned attention strategies across model types and how the architectures model the interactions between the multimodal input features. For example,

Model	**Text**	**Image**	**Inter-Modal**
MCAN_G	-4.60***	-8.32***	-8.33***
MCAN_R	-5.50***	0.21	0.05
Pythia	-2.83**	-1.83*	-1.81*
BAN	-2.20*	-3.62***	0.53
MFB	-3.32***	-0.05	-0.208

Table 3: Ordinal Logistic regression model t-values such that for every one unit decrease in correlation, each respective model is less likely to predict the answer correctly. Significance is denoted as $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. Correlation to human text attention is a significant predictor of accuracy for **all five models**. Correlation to human image attention is an important for the accuracy of MCAN_G, Pythia, and BAN while inter-modal correlationis a significant predictor of accuracy for both MCAN_G and Pythia.

the MCAN grid model applies self- and guided attention to model the interplay between grid-based image and text feature representations. On the other hand, the Pythia model uses both bottom up attention (image features extracted on the region level) and top down attention (text attention applied over the images), where the text attention weights are not learned by the image feature representations.

6 Conclusion and Future Work

In this work we have presented VQA-MHUG – a new, fully annotated 49-participant dataset for visual question answering that includes nearly 4,000 question-answer pairs. Our dataset is unique in that it is the first to provide real human gaze data on both images and corresponding questions and, as such, allows researchers to jointly study human and machine attention. Revealed through a detailed

comparison of multiple leading VQA models, we showed that higher correlation between neural and human text attention is a significant predictor of high VQA performance. This novel finding highlights the potential to improve VQA performance with human-like attention biases and simultaneously calls for further investigation of neural text attention mechanisms, as we find these are an indicator for success on language and vision tasks, including VQA.

Ethical Statement

We identified a number of potential benefits and risks of our approach.

Potential benefits By leveraging human behavioral data, our method could be used to guide intelligent user interfaces using human attentive abilities within the context of reading behaviors. We see significant potential of approach to interpret text attention to enable a new generation of attentive text interfaces, particularly when jointly modelling with user task specific eye movement behaviors during comprehension tasks. We see potential for e-learning multimodal applications approach could be used to qualify reader actions and provide feedback to encourage improvement in comprehension. By bridging the gap between human and neural attention, we see a potential positive impact in improving attention strategies in users.

Potential risks Though we see the aforementioned potential benefits, we also identified a some risks and ethical concerns. By aiming to interpret the gap between human and machine attention, we open the door for potentially exploiting user biases. In addition, one can conceive that there is potential for using the findings of our work to develop tool which discriminate against specific users given their eye movement behaviors. This leads to the discussionabout the behavioral data collection, it is conceivable that one could generate a system which might predicts cognitive impairments in order to filter out individuals from some program or opportunity.

Dataset Curation To protect the privacy of our participants we saved all data anonymized and collected only directly relevant data and demographic information in compliance with our university's code of ethics and the General Data Protection Regulation (GDPR) of the European Union (EU). Our study was approved by the ethics committee (institutional review board) of the university. Additional measures for safety during the COVID-19 pandemic were taken with disinfection of the material, obligatory masks and breaks between scheduled recording sessions. All participants signed a consent form that included details about the purpose, goal, procedure, risks, benefits and privacy measures of our research. For the 45-60 minute study an above average compensation of 20€ was paid. At any point the participant could abort the study without penalty. The study took place in a standard university lab and the participant's head was not fixed. Every 15 minutes a 5 minute break was scheduled.

Acknowledgements

E. Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016; F. Strohm and A. Bulling were funded by the European Research Council (ERC; grant agreement 801708);
We would like to especially thank Simon Tannert for his valuable insights and support, as well as Dr. Philipp Müller and Dr. Paul Bürkner for their helpful suggestions. Lastly, we would like to thank the anonymous reviewers for their useful feedback.

References

Aishwarya Agrawal. 2015. Visualqa official evaluation code. https://github.com/GT-Vision-Lab/VQA.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018a. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018b. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for

image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Viv Bewick, Liz Cheek, and J. Ball. 2003. Statistics review 7: Correlation and regression. *Critical Care*, 7:451 – 459.

Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. 2020. Air: Attention with reasoning capability. In *European Conference on Computer Vision*, pages 91–107. Springer.

A. S. Correia and E. Colombini. 2021. Attention, please! a survey of neural attention models in deep learning. *ArXiv*, abs/2103.16775.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937, Austin, Texas. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. 2020. How Much Time Do You Have? Modeling Multi-Duration Saliency. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 4473–4482.

Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. 2017. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1811–1820.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. 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.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. 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.

Sen He, Hamed Rezazadegan Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human attention in image captioning: Dataset and analysis. In *2019 International Conference on Computer Vision*, pages 8528–8537, Piscataway, NJ. IEEE.

Ignace T. C. Hooge, Gijs A. Holleman, Nina C. Haukes, and Roy S. Hessels. 2019. Gaze tracking accuracy in humans: One eye is sometimes better than two. *Behavior Research Methods*, 51(6):2712–2721.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *Computer Vision - ECCV 2016*, Lecture Notes in Computer Science, pages 727–739, Cham and s.l. Springer International Publishing.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. 2020. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276.

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Kushal Kafle, Robik Shrestha, and Christopher Kanan. 2019. Challenges and prospects in vision and language research. *Frontiers in Artificial Intelligence*, 2:28.

Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1571–1581.

Jin-Hwa Kim, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2016. Multimodal residual learning for visual qa. *arXiv preprint arXiv:1606.01455*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in Neural Information Processing Systems*, 27:1682–1690.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096.

P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. 2015. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering: Hlat. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

LAI Qiuxia, Salman Khan, Yongwei Nie, Sun Hanqiu, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020a. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011.

Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020b. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019a. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019b. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6649–6658.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proc. ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yusuke Sugano and Andreas Bulling. 2016. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*.

B. Tatler. 2007. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of vision*, 7 14:4.1–17.

Hamed R. Tavakoli, Fawad Ahmed, Ali Borji, and Jorma Laaksonen. 2017a. Saliency revisited: Analysis of mouse movements versus fixations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1774–1782.

Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017b. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2487–2496.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Chao Yang, Mengqi Jiang, Bin Jiang, Weixin Zhou, and Keqin Li. 2019a. Co-attention network with question type for visual question answering. *IEEE Access*, 7:40771–40781.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016a. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016b. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. 2019a. Openvqa. `https://github.com/MILVLG/openvqa`.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

K. Yun, Y. Peng, D. Samaras, G. J. Zelinsky, and T. L. Berg. 2015. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 739–746.

Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J Zelinsky, and Tamara L Berg. 2013. Studying relationships between human gaze, description, and computer vision. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746.

Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett. 2018. Learning to count objects in natural images for visual question answering. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Appendix

A VQA-MHUG Overlap to Related Datasets

During the selection of stimuli for VQA-MHUG, we maintained large overlaps with other benchmark and attention datasets that also used subsets of VQAv2 questions/images to allow for easy integration and comparison of our data with existing approaches (see Table 4).

Dataset	∩
VQAv2 val (Goyal et al., 2017a)	3,990
VG (Krishna et al., 2017)	2,238
VQA-CP2 (Agrawal et al., 2018b)	1,904
VQA-Rephrasings (Shah et al., 2019b)	1,373
VQA-Introspect (Selvaraju et al., 2020b)	1,213
SALICON (Jiang et al., 2015)	1,134
TDIUC (Kafle and Kanan, 2017)	1,125
VQS (Gan et al., 2017)	695
VQA-X (Park et al., 2018)	491
VQA-HAT (Das et al., 2016)	410

Table 4: Overlap of VQA-MHUG question-image pairs with different established VQA related datasets.

B Reasoning Types

We binned question-image pairs by 12 reasoning types, as they align better with potential error classes than the VQAv2 question types. Figure 4 shows the relationship of reasoning types to question types. The reasoning types incorporate the categories proposed by Kafle and Kanan (2017), except the absurd category and adding a new *reading* category for questions that ask about text on the images.

- Scene Recognition
- Object Presence
- Colour
- Positional Reasoning
- Counting
- Utility Affordance
- Object Recognition
- Activity Recognition
- Attribute
- Reading
- Sentiment Understanding
- Sport Recognition

B.1 Tagger

To label VQA-MHUG with our reasoning types we used a LSTM-based classifier to predict the reasoning type given the question-answer pair. The input text is encoded using 300D glove embeddings (Pennington et al., 2014), which are passed though a single LSTM layer with hidden size 256 and a final softmax classification layer. We labeled 145 K VQAv2 train-val questions and extended the 1.6 M TDIUC questions by the *reading* category using regular expressions and manual work. We trained the network using this data by optimizing cross-entropy loss with the Adam optimizer and a batch size of 128. The final model achieves an accuracy of 99.67% on a held-out set of 20% of the training data. The trained tagger was then used to label the question-image pairs in VQA-MHUG. Figure 3 shows the label distribution.

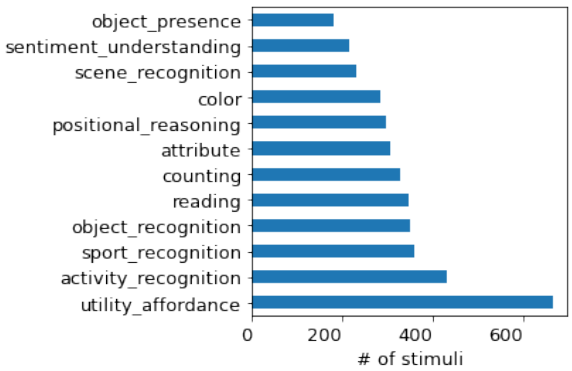


Figure 3: Final distribution of tagged reasoning types in VQA-MHUG. When no other type fit, the tagger assigned *utility affordance*, which had the least training data. This indicates that there could be clusters that do not fit any current type.

C Machine Difficulty Score

For our machine difficulty score we evaluated the Multimodal Factorized Bilinear Pooling Model (MFB) (Yu et al., 2017) for multimodal fusion and the Multimodal Co-Attention Network (MCAN) (Yu et al., 2019b) for transformer attention on four datasets (VQAv2, VQA-CPv2, VQA-Introspect and VQA-Rephrasings). The standard VQAv2 and VQA-CPv2 use simple accuracy, but VQA-CPv2 has intentionally dissimilar answer distributions in the train and validation splits to not allow exploitation of priors. In general, low accuracy on both model architectures indicates a harder ques-

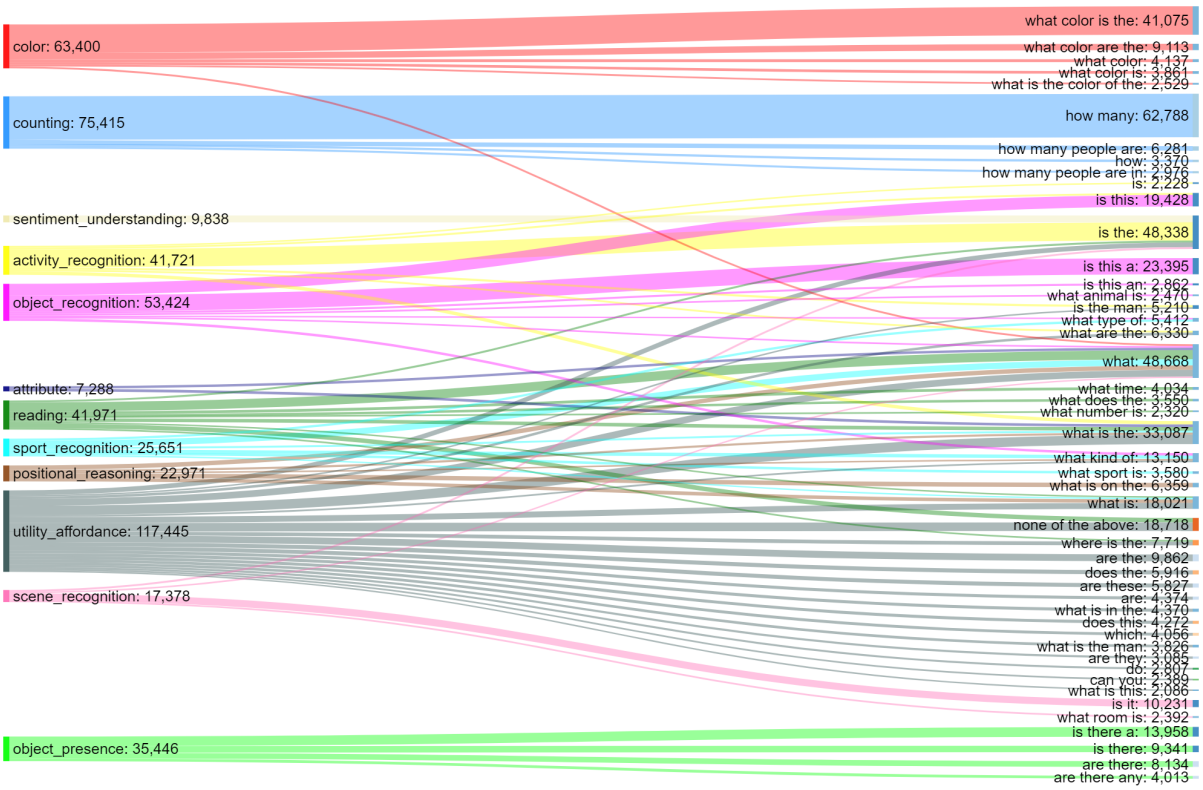


Figure 4: Relationship of our VQA-MHUG reasoning types (left) and VQAv2 question types (right). Question types are no good categories for error case analysis since they mix many reasoning capabilities.

tion (Equation 1).

$$\begin{aligned} &\text{score}_{\text{VQA}}(\text{ans}) = \\ &\text{score}_{\text{VQA-CP}}(\text{ans}) = \\ &1 - (\min(\frac{\text{\# of annotators that said ans}}{3}, 1)) \end{aligned} \tag{1}$$

VQA-Introspect asks additional perceptual sub-questions to VQAv2 and tests consistency w.r.t. visual grounding. If a model is correct on the main question "Can birds fly?", but fails the perceptual sub-question "Are the birds in the air?", it is inconsistent and the question is potentially too easy, as it can be answered from the question alone. We assign the (binary encoded) four combinations of "main correct/incorrect" and "all sub-questions correct/incorrect" numerical values to combine it with the other metrics. In Equation 2, we purposefully assign a high difficulty to a question where the perceptual sub-question is correctly answered, but the main reasoning question is not (01) and a low difficulty for a question that seems to exploit question bias (10).

$$\text{score}_{\text{Intro}} = \begin{cases} 1.0 & 00 \text{ or } 01 \\ 0.25 & 10 \\ 0.0 & 11 \end{cases} \tag{2}$$

Finally VQA-Rephrasings tests robustness against 3 linguistic variations per question and measures this with a "consensus score" – the share of fully correctly answered subsets of size k of a question and its rephrasings. It is unclear how to interpret different settings of k, so we set $k = 1$, which simplifies to simple accuracy over the rephrasings (Equation 3).

$$\begin{aligned} &\text{score}_{\text{Rep}} = \\ &1 - (\frac{\text{\# of correct k-sized subsets}}{\text{\# of k-sized subset}} \end{aligned} \tag{3}$$

We combined the four resulting scores of each MFB and MCAN in equal parts (Equation 4). Since not all our candidate stimuli are present in all four datasets used in the difficulty score, there is a set S of $|S| \in [1, 4]$ scores per question-image pair. We penalize the cases where only one score is available by normalizing over avg($|S|$) instead of $|S|$ to

counterweight the uncertainty it brings.

$$\text{difficulty} = \frac{1}{max(|S|, \text{avg}(|S|))} \cdot \sum_{s \in S} \left(\frac{\text{score}_s^{\text{MFB}} + \text{score}_s^{\text{MCAN}}}{2} \right) \tag{4}$$

Figure 5 shows the resulting distribution of difficulty scores.

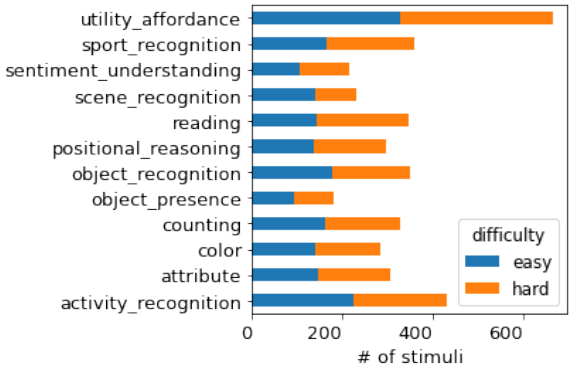


Figure 5: Final distribution of difficulty per tagged reasoning type in VQA-MHUG. Clearly some types like *reading* and *counting* are harder than others.

D Significance between models

Table 5 shows the significance of the differences in Rank Correlation and JSD for pairs of models.

E Experimental Setup

Binocular gaze data was collected with an EyeLink 1000 plus remote eye tracker at 2kHz. To ensure gaze estimation accuracy, participants were asked to use a mounted chin rest (see Figure 6). The stimuli was shown on a 24.5" screen with resolution of 1920×1080 pixels. The monitor was placed 90cm in front of the participants.

F MHUG vs. AiR-D Examples

The AiR-D dataset does not overlap with VQAv2, as such we separately visualize a selection of examples (see Figures 7 and 8) from the overlapping 195 additional stimuli presented to humans during the VQA-MHUG data collection.

G More Model Examples

Figures 9 and 10 show additional visualization examples of VQA-MHUG data in comparison with the extracted model data. We randomly sampled

Figure 6: Setup of the eye tracker in our lab

question-image pairs with high image attention correlation (Figure 9) and high text attention correlation (Figure 10).

Between Model Comparison	Image Correlation	Image JSD	Text Correlation	Text JSD
MCAN_G vs MCAN_R	***	***	***	***
MCAN_G vs PYTHIA	***	***	***	***
MCAN_G vs BAN	***	***	***	p>0.05
MCAN_G vs MFB	***	***	***	***
MCAN_R vs PYTHIA	***	***	***	***
MCAN_R vs BAN	***	***	***	p>0.05
PYTHIA vs BAN	p>0.05	***	***	***
PYTHIA vs MFB	***	***	***	***
BAN vs MFB	***	***	***	***

Table 5: We performed a paired t-test to indicate if the differences between correlation and JSD scores is statistically significant where p<0.05 (*), p<0.01 (**), p<0.001 (***). We show that for all models, the image correlation scores are statistically differ except when comparing the Pythia and BAN models. The image JSD scores and text correlation scores are significantly different for all models. The difference between models text JSD scores are significant, except for between BAN and both MCAN networks.

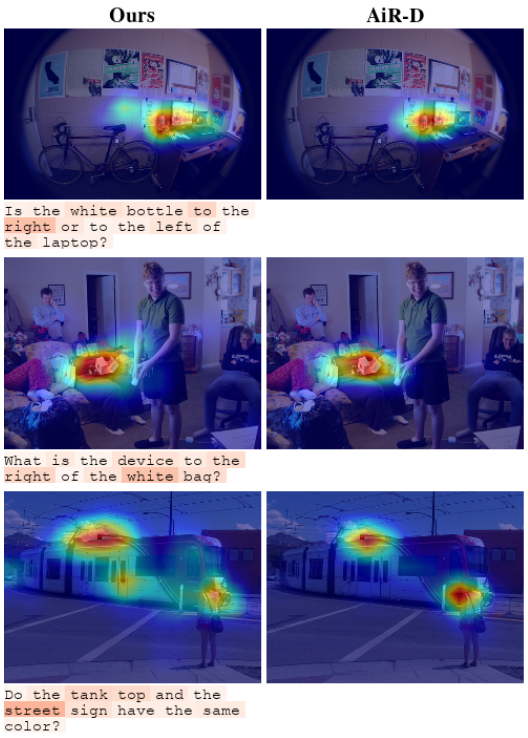


Figure 7: Examples of MHUG gaze vs. AiR-D gaze

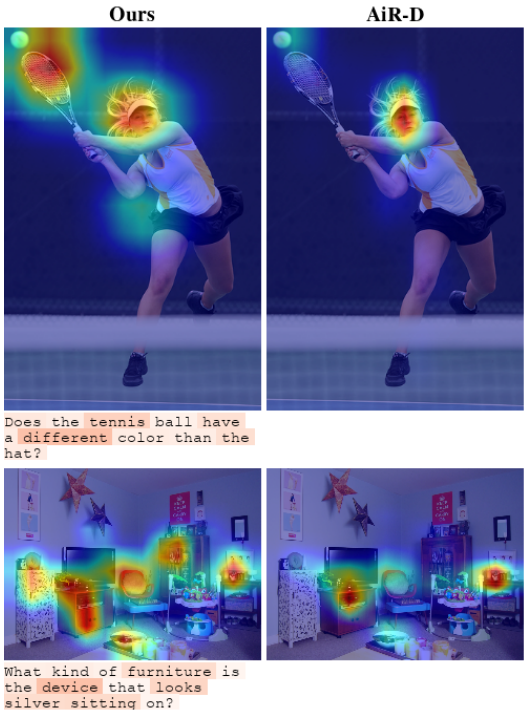


Figure 8: Examples of MHUG gaze vs. AiR-D gaze

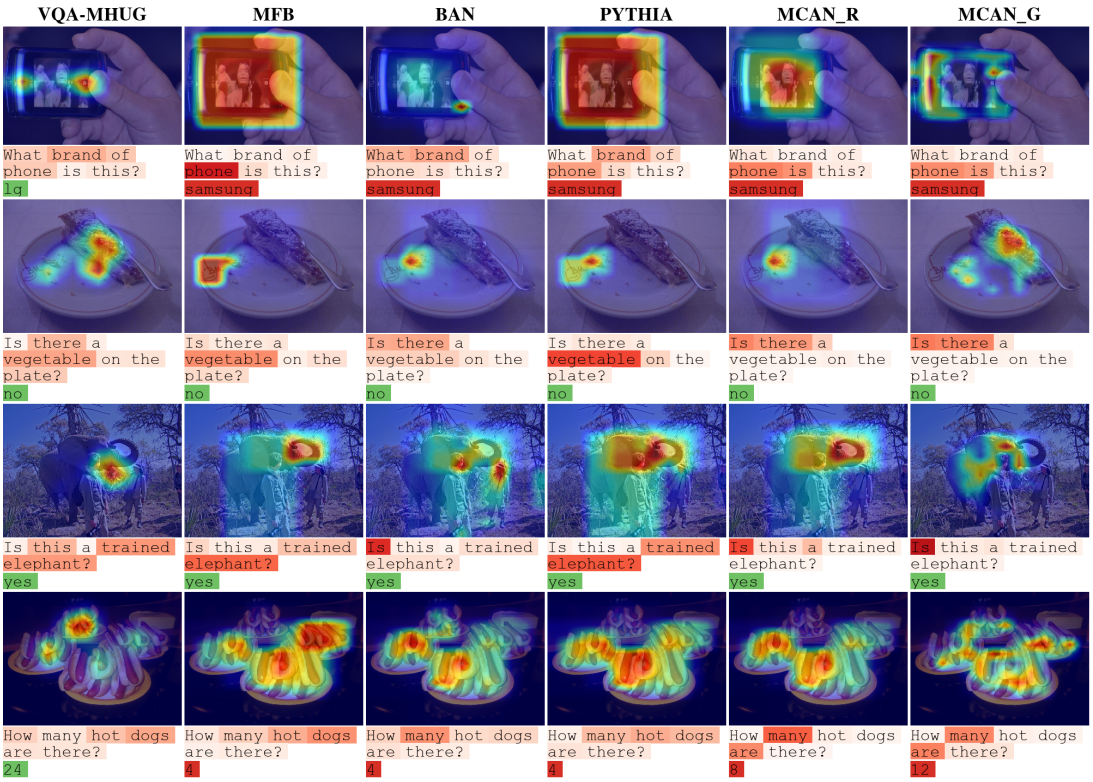


Figure 9: Comparison of VQA-MHUG attention maps and the model extracted attention maps on text and images, **where image attention correlation is high**.

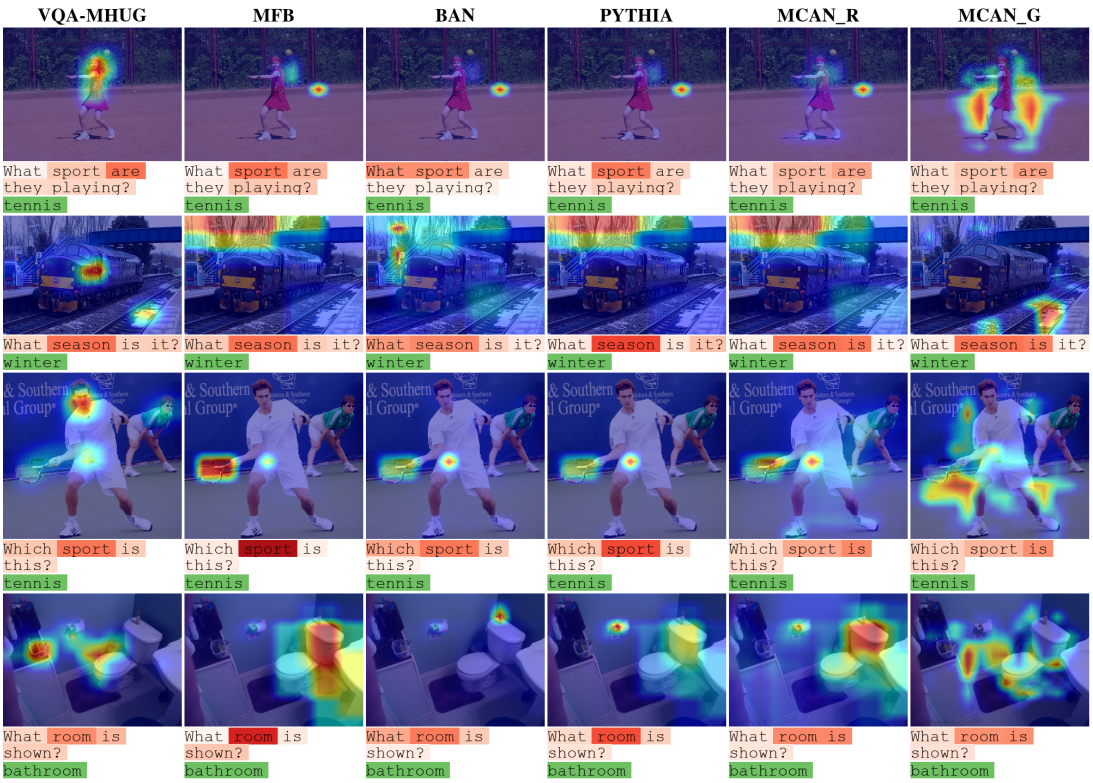


Figure 10: Comparison of VQA-MHUG attention maps and the model extracted attention maps on text and images, **where text attention correlation is high**.