The Impact of Explanations on AI Competency Prediction in VQA

Kamran Alipour*, Arijit Ray†, Xiao Lin†, Jurgen P. Schulze*, Yi Yao†, and Giedrius T. Burachas†
*UC San Diego, La Jolla, CA.
{kalipour, jschulze}@eng.ucsd.edu
†SRI International, Princeton, NJ.
{arijit.ray, xiao.lin, yi.yao, giedrius.burachas}@sri.com

***Abstract*—Explainability is one of the key elements for building trust in AI systems. Among numerous attempts to make AI explainable, quantifying the effect of explanations remains a challenge in conducting human-AI collaborative tasks. Aside from the ability to predict the overall behavior of AI, in many applications, users need to understand an AI agents competency in different aspects of the task domain. In this paper, we evaluate the impact of explanations on the users mental model of AI agent competency within the task of visual question answering (VQA). We quantify users understanding of competency, based on the correlation between the actual system performance and user rankings. We introduce an explainable VQA system that uses spatial and object features and is powered by the BERT language model. Each group of users sees only one kind of explanation to rank the competencies of the VQA model. The proposed model is evaluated through between-subject experiments to probe explanations' impact on the users perception of competency. The comparison between two VQA models shows BERT based explanations and the use of object features improve the users prediction of the models competencies.**

I. Introduction

Recent developments in the field of AI and specifically deep neural networks (DNN) have brought them into a broad range of applications. DNNs have automated a wide range of human activities resulting in reduced complexity of many tasks. Users of AI systems, though, need to maintain at least a minimal level of understanding and trust in the system, i.e., they need a proper mental model of the systems internal operations for anticipating success and failure modes.

While accuracy is well-known as the primary metric for AI efficiency, it cannot guarantee a collaborative human-machine interaction in the absence of trust. If the users do not trust a model or a prediction, they will not use it [1]. This mistrust escalates in the presence of adversarial attacks where imperceptible changes to the input lead to wrong outputs and also the susceptibility of DNNs to non-intuitive errors.

Explainable AI aims to gain user's trust on two major steps of interpretability and explainability. Interpretable models provide a basic comprehension of their inner-processes through visual or textual cues. On a higher level, explainable models attempt to provide reason and causality behind their decisions[2].

The appearance of various methods of explanations call for a parallel effort to evaluate and quantify their efficiency. While previous works introduce nominal visualizations and textual justifications on the inner features of DNN models; yet it does not evaluate the impact of explanations on various aspects of users understanding and trust.

Evaluation techniques for explanations include automatic and human methods. Automatic approaches provide quantifiable measures over relevant benchmarks e. g. alignment with human attention datasets[3], however they still cannot propose a straight-forward metric for trust in actual human-machine task.

Furthermore, human-based approaches attempt to quantify explanation effectiveness through collecting user ratings [4], [5]. Despite their insightful results, these methods do not measure user's perception of AI competency in the whole domain.

Users can benefit from AI systems more efficiently if they are familiar with the AI agents competency in the operational domain. The competency of AI can be impacted by the biases in the training data or limited representation of crucial features. An explanation system that provides case-by-case reasoning for AI behavior does not automatically produce a higher view of competency. Particularly, deep learning models are notoriously opaque and difficult to interpret and often have unexpected failure modes, making it hard to build trust.

As prior research shows explanations improve users prediction of system accuracy [6]. Herein, we focus on the users mental model of an AI system in terms of competency understanding. Specifically, we evaluate the importance of explanations for helping users interpret how a VQA system performs on different types of questions. We model users learning process under two different explanation systems to identify the role of the attention-based explanations in users prediction of competency. For this purpose, we evaluate the impact of explanations on user learning rate and also their ultimate score on the task of competency prediction.

II. Related work

Visual question answering (VQA). Originally introduced by [7], the VQA problem involves the task of answering questions about the visual content of an image. The VQA task is specifically challenging due to the complex interplay between the language and visual modalities[8]. Limited

labeled data and the complex feature space complicate the process of developing VQA models. These challenges result in models with inconsistent outputs and serious logical contradictions[9]. In such an environment, the choice of hyper-parameters and architectural designs can have drastic impacts on the performance of VQA models[10].
A common approach to VQA is to use DNNs with attention layers that select specific regions of the image, guided by the question for inferring an answer[11], [12], [13], [14]. Herein, we also study two attention based VQA models with different attention structures. As a baseline, we use a model based on Kazemi and Elqursh [15] and Teney *et al.*[11] approaches. We propose a new VQA architecture by replacing the attention mechanism with a BERT model[16] in the baseline VQA model.
The previous work in VQA includes various attempts to optimize the attention mechanism. To improve the attention on the question, Lu *et al.*[17] utilize a co-attention model to jointly reason about image and question on hierarchical levels. Anderson *et al.*[18] propose a combined bottom-up and top-down attention mechanism to calculate attention at the level of objects. The model is further upgraded and fine-tuned to win the VQA Challenge 2018[19].
Despite all the advancement in the overall accuracy of VQA models, their unbalanced performance in different aspects of the task is overtly noticeable. Some prior approaches address this issue by focusing on certain tasks such as reading text in images[20] or counting objects[21]. Other works introduce new datasets to reduce bias [22] or to enforce the logical consistency of model through *visual commonsense reasoning (VCR)* for challenging questions[23].

Explainable AI (XAI). The ever increasing complexity of the modern AI machine demands a trustable source of explanation for all the AI users. Generating automated reasoning and explanations dates back to very early work in the AI field with direct applications from medicine [24]and education [25], [26], to robotics [27]. In the field of computer vision, several explanation systems focus on the importance of image features in the decision-making process [28], [29], [30], [19].
AI explanations for the task of visual question answering usually include image and language attentions [4], [15]. Besides saliency/attention maps, other efforts investigated different explanation modes like layered attentions [31], bounding boxes around important regions [32], textual justifications [24], [33] or a combination of these modes [6].
We propose an explainable VQA system which produces justifications for system answer in the form of an attention map. Unlike previous post-hoc saliency approaches such as GradCAM[34], our method seeks causal explanations by providing attentions as an inherent step of answer inference. Our proposed model uses visual features on both spatial and object level. For better performance in VQA task, the proposed model utilizes BERT language model to process question features along with the visual features.

Explanation evaluation. As the AI machines enter the daily life of people, a new interest has surged among the AI community to make AI algorithms more understandable to the lay users without the technical background[35]. In this work, we choose the subjects for explanation evaluation from a group of individuals with minimum knowledge about AI and deep neural networks.
Evaluating the impact of explanations on user mental model and human-machine performance is widely discussed in the XAI literature. Some of the earlier works take on quantifying the efficacy of explanations through user studies to assess the role of explanations in building a better mental model of AI systems for their human users.
Some of the previous studies introduced metrics to measure trust with users [36], [1], or the role of explanations to achieve a goal [37], [38], [39]. Dodge *et al.* investigated the fairness aspect of explanations through emperical studies[40]. Lai and Tan [41] assessed the role of explanations in user success within a spectrum from human agency to full machine agency. Lage *et al.* proposed a method to evaluate and optimize human-interpretability of explanations based on measures such as size and repeated terms in explanations[42]. Other approaches measured the effectiveness of explanations in improving the predictability of a VQA model [43], [6].
In this work, we conduct a user study to investigate the impact of explanations on the users mental model of system competency. Within the study, subjects attempt to rank system performance among different types of input questions. The results indicate a positive influence on the accuracy of the users mental model in the presence of explanations. We detail the overall and temporal effect of explanations on the users interpretation in two explainable VQA models.

III. Methods

Our approach aims at evaluating the role of attentional explanations in the user's mental model of AI competency. To accomplish this task, we compare two explainable VQA models and test them through user studies.
In this section, we cover the architecture details for these VQA models and the differences in their attention mechanisms. The section later follows with sample cases from both explanation models and the differences between them.

A. Explainable VQA (XVQA) models

Our work compares two VQA agents: spatial attention VQA (SVQA) and spatial-object attention BERT VQA (SOBERT). Both agents are trained on VQA 2.0 dataset. SVQA is based on a 2017 SOTA VQA model with a ResNet [44] image encoder (figure 2). The agent uses an attention mechanism to select visual features generated by an image encoder and an answer classifier that predicts an answer from 3000 candidates.
As shown in figure 2, SVQA takes as input a 224×224 RGB image and question with at most 15 words. A ResNet subnet

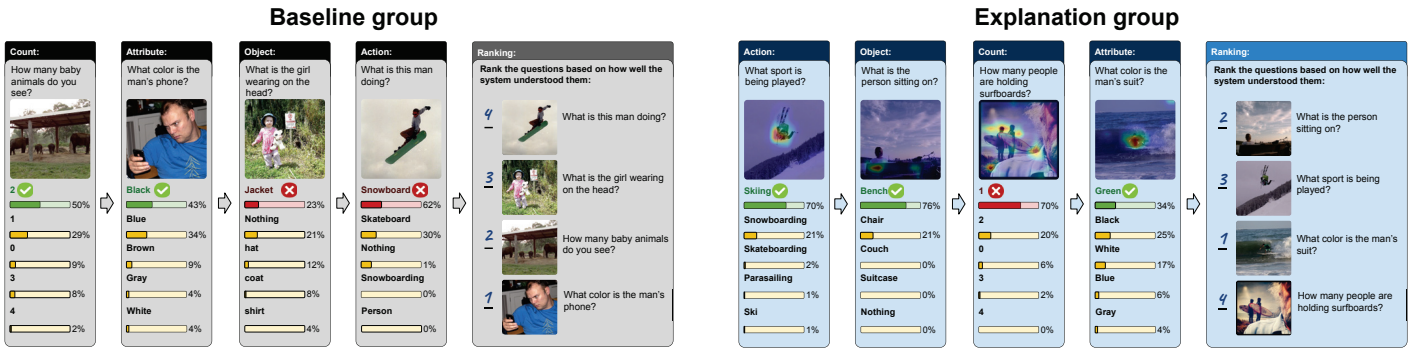

Fig. 1. The workflow for user study groups: Left shows the baseline group where the users only view the top five answers from the model along with the probability of the answers. As shown on the right, users inside the explanation group, also view the attention maps generated by the model. Each group views blocks of trials. At the end of each block, users are asked to rank the question-images based on how well they seem to be *understood* by the model.

encodes the image into a $14\times14\times2048$ feature representation. An LSTM model (GloVe [45]) encodes the input question word embeddings into a feature vector of 512 dimensions.
The attention layer in the SVQA model transfers the question and image features to a set of attention weights on the image features. The model convolves the concatenation of weighted image features and question features to produce the attention layer with $14 \times 14 \times 1024$ dimensions. The model predicts the probability of the final answer from a set of 3000 answer choices using a multilayer perceptron (MLP). The attention layer also goes through a convolution block to generate the spatial attention map.

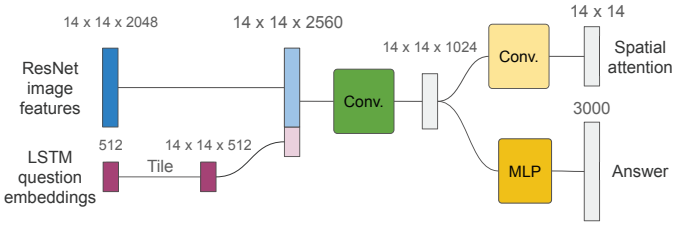

Fig. 2. The architecture of explainable SVQA model.

On the other hand, the SOBERT agent uses a combination of visual embeddings of the image from ResNet and Faster RCNN (FRCNN)[46] alongside question embeddings (figure 3). SOBERT accepts questions with a maximum length of 30 words and the input question embeddings contain the location and token information of words. The location features are encoded in both ResNet and question embeddings.
SOBERT agent uses a BERT model with 4 layers and 12 attention heads. BERT transfers the hidden features (115×768) into spatial attention heads ($12 \times 7 \times 7$) and output layer. An MLP maps the output layer to the final answer prediction out of 3129 candidates.
Based on their training process and their characteristics, VQA agents can reach certain levels of accuracy in each type of question. For our tests, we limit the cases into a subset of VQA 2.0 validation set with questions about action, attribute,

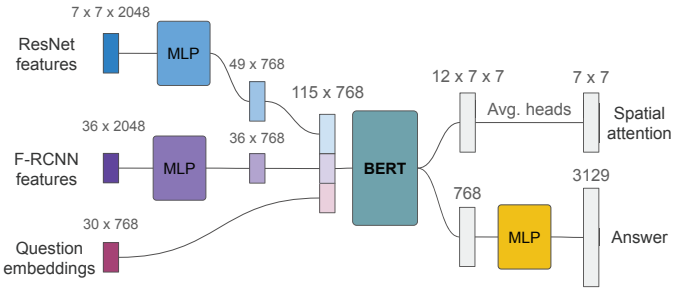

Fig. 3. The architecture of the explainable SOBERT model. This model passes the combination of visual features from ResNet and FRCNN and question embeddings into a BERT model to produce answers and spatial attention.

object, and count. We classified the question using a set of automated methods including word matching in questions and also their answers.
Questions about activity inside an image are labeled as "Action". Questions about objects inside the image are labeled as "Object". Questions that are specific about attributes of entities in the image (e.g. color) are labeled as "Attribute". Finally, questions about counting entities on the image are categorized as "Count". Table I shows the accuracy of SVQA and SOBERT agents in these four categories. The accuracy of models are computed over the four categories within VQA validation dataset.
As numbers in table I show, the two models pose a similar ranking between the four categories of questions, while the SOBERT model can reach a higher accuracy in all of them compared to the SVQA model.

B. Explanations

The VQA agents can produce a spatial attention map to visualize the areas of focus while producing the answer. SVQA model convolves the attention tensor into a 14×14 spatial map. In the SOBERT model, the attention tensor is averaged over the 12 attention heads into a 7×7 spatial attention map.
The attention maps generated by the VQA agents provide a causal explanation to the users as they illustrate AI spatial/object attentions as an inherent step in answer inference.

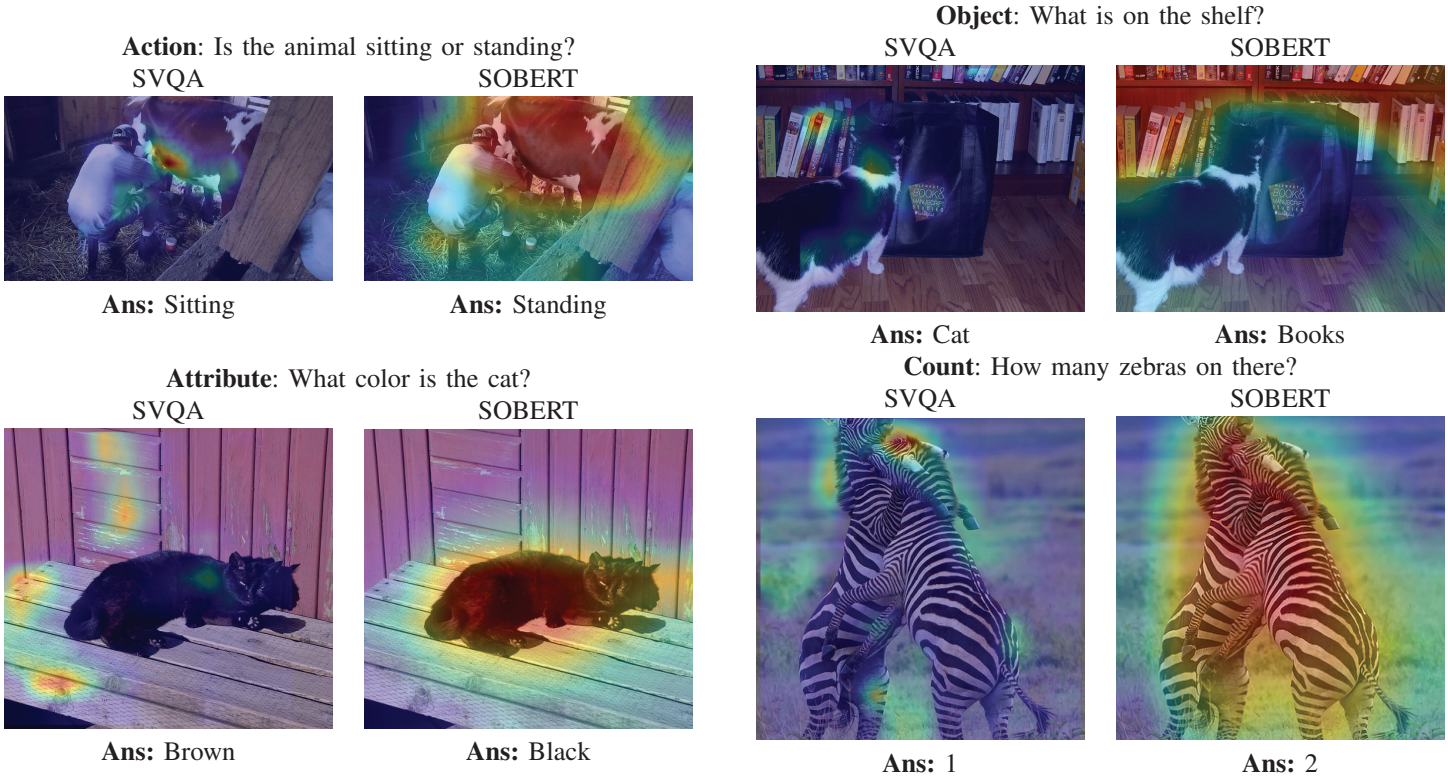

Fig. 4. Attention maps generated by the AI agents for questions in different question type categories. As illustrated in the results, the SOBERT model produces attention maps with more focus on the areas related to the question.

	Action	Attribute	Object	Count
SVQA	81.21%	70.83%	64.46%	45.78%
SOBERT	**88.35%**	**86.63%**	**71.84%**	**60.14%**

TABLE I
THE ACCURACY OF VQA AGENTS IN FOUR SELECTED CATEGORIES OF QUESTION.

Both models use spatial features from the images while gaining a general representation of image content. SOBERT model also incorporates object-level F-RCNN features into the process.

One major impact of including object-level attention emerges in the attention map outputs of the model. As can be seen in figure 4, the attentions from the SOBERT model cover broader areas that are associated with objects on the scene. Also, the averaging layer that generates attention produces smooth attention distributions in the SOBERT model compared to more localized and scattered attention in SVQA.

IV. EXPERIMENTS

We designed an interface for an in-person user study to evaluate the impact of explanations on the users understanding of AI agent competency among different question types. At the introductory section of each study session, subjects are reminded that the model competency and accuracy of the AI model is unknown to minimize their prior knowledge and judgment of the AI agent competency.

In this user study, subjects go through a set of trial blocks where the AI agent answers questions about images. Each block consists of four trials with one image-question of each type: object, attribute, action, and count. On each trial, subjects first see the input image and question and then they proceed to see the outputs of the AI agents.

For each model, the study is divided into two groups of baseline and explanation. Each study group contains 10 subjects and each subject goes through 100 trials (25 blocks). In all groups, users see the agents top five answers, their probabilities, and agents Shannon's confidence in each trial. In the explanation group, subjects first view the attention map from the model and then see the top answers and confidence value. Subjects are asked to rank the helpfulness of attention maps on understanding AI's performance on that trial.

At the end of each block, subjects rank the trials within the block based on system performance in each question type. Comparing question type rankings from subjects between baseline and explanations measures the explanation on subjects opinions of system competency (figure 5).

In each block of trials, four question-images show up in random order. The AI agent's success ratio in each block is also random. Among the baseline group, users can rely

Model	Condition	Final ranking corr.	Max. user learning rate (corr. / blocks)
SVQA	Baseline	0.757	0.0105
	Explanation	**0.805**	**0.0769**
SOBERT	Baseline	0.611	0.0253
	Explanation	**0.921**	**0.0468**

TABLE II
THE MAXIMUM LEARNING RATE OF USERS AND THE FINAL VALUE OF CORRELATION IN COMPETENCY RANKING TASK. BOTH EXPLANATION MODELS SHOW AN IMPROVEMENT IN EARLY LEARNING RATES. WHILE EXPLANATION FROM THE SOBERT MODEL DOES INCREASE THE LEARNING RATE AS MUCH AS SVQA, HOWEVER, SOBERT REACHES A RELATIVELY HIGHER FINAL LEARNING RATE.

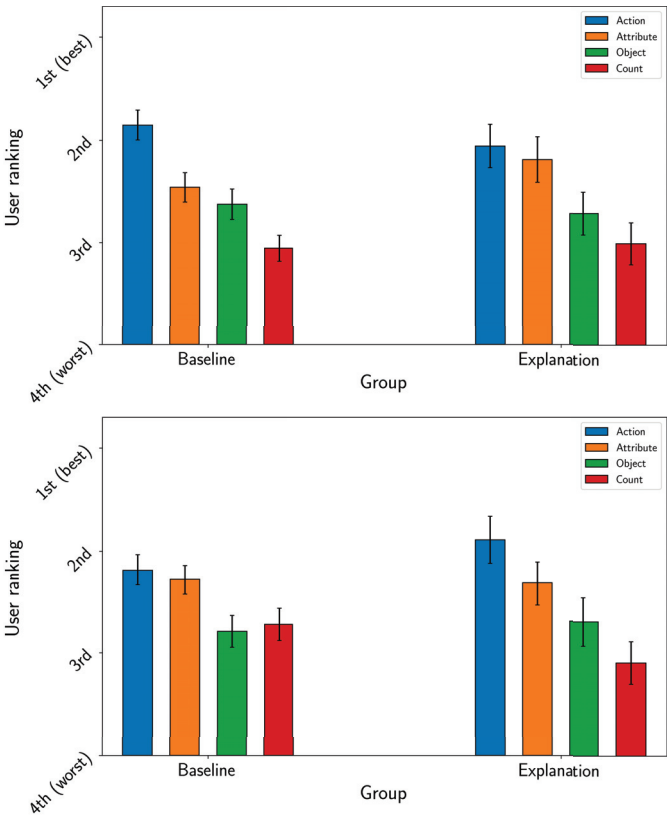

Fig. 5. The average of all rankings entered by the subjects at the end of every block of trials (Top: SVQA model, Bottom: SOBERT model).

on the top answers and their probabilities to understand system performance on that question and image. On the other hand, subjects from the explanation groups have the extra information provided by the attention maps (figure 1).

A. Explanation helpfulness

In the explanation group, subjects view the attention explanations before they see the final answers and accuracy of AI. At this stage, subjects rate the explanations based on their helpfulness towards understanding AI's performance.
The helpfulness rankings are specifically interesting for action and count question types within which the VQA agents show their highest and lowest competencies. The helpfulness rankings of within these categories on SOBERT explanations show an increase compared to SVQA (figure 6). While subjects rank 17% of SVQA explanations are ranked as not helpful in count questions, this number is reduced to 7% by SOBERT explanations. In action questions, SOBERT also reduces the unhelpful explanations from 8% to 3%.

B. Competency ranking

We assess the accuracy of subjects ranking by measuring the correlation between that and the ground truth competency ranking of AI agents (figure I) and the collected rankings at the end of each block. Figure 6 illustrates this correlation in the starting and finishing blocks of each study group. The start and finish values of correlation are the average of 1-5 and 20-25 blocks respectively.

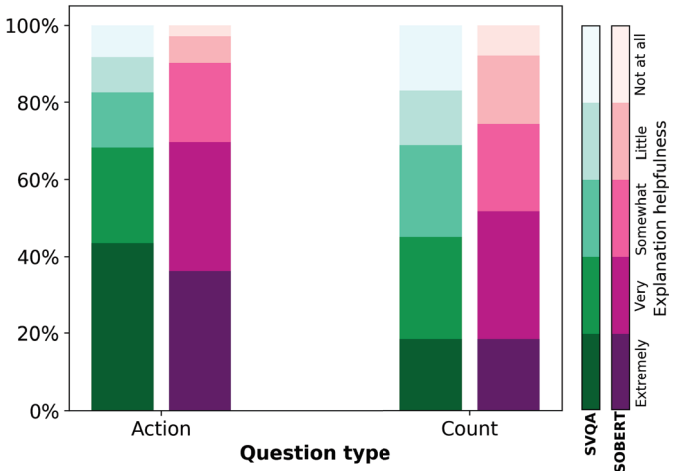

Fig. 6. Histogram of ratings of how helpful explanations are for the subjects. These helpfulness ratings are given by the subjects as they view the explanations and before they see the system top 5 answers. So these ratings are not confounded by the accuracy of the AI.

Overall, the ranking correlation shows an increase in both models with a slightly higher slope in the presence of explanations (table 8). To better picture the temporal impact of explanations on the users mental model, figure 7 presents the progress of ranking correlation throughout the study. In the early blocks for both models, the explanation groups increase their ranking correlation with a higher rate than baseline.

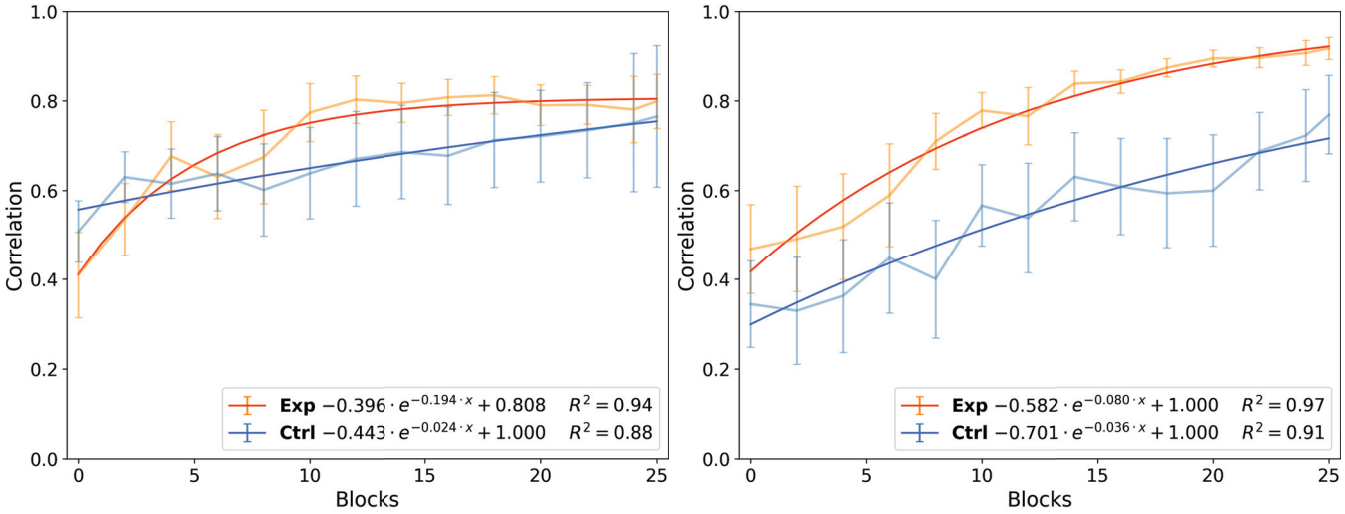

Fig. 7. Temporal impact of attention maps on user rankings. Left: the growth of correlation in baseline and explanation groups is compared between baseline (blue) and explanation (orange) groups for two models SVQA (left) and SOBERT (right). T-test p-values for SVQA and SOBERT data are 0.07 and $3.7e-8$ respectively.

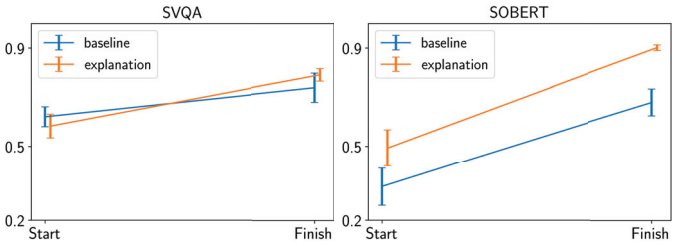

Fig. 8. The overall correlation between the users rankings and the systems actual competencies. Comparing the results from the SVQA (left) model and our SOBERT (right) model suggests a better improvement of correlations in the presence of SOBERT attention maps.

C. Competency learning curves

We also investigate the temporal pattern of temporal ranking correlation by fitting curves into the data in baseline and explanation groups. This problem, in general, can be viewed as modeling user learning curve for a certain task.

The modeling user learning curve is widely discussed in cognitive science. In previous works, researchers analytically derived exponential learning equations to describe user improvement in the task[47], [48]. The assumption of a monotonically decreasing improvement is the main foundation beneath the exponential learning curves.

Here in the context of learning AI competency rankings, subjects start the study with no prior knowledge of the AI agent's rankings. Also, the correlation metric cannot exceed the value of 1.0.

Considering these similarities to the general learning model, we also considered an exponential curve with an upper bound as blocks grow to infinity. With this analogy, we considered the following curve to fit the ranking correlation trends:

$$c = \alpha \cdot e^{-\beta \cdot b} + \delta$$

where b and c are the block count and ranking correlation respectively. In this setting, the ranking correlation approaches δ as the subjects continue the study. The value of δ is penalized for curves fitting to satisfy the condition $\delta \leq 1.0$.

The slope of the fit curves in figure 7 represents the growth rate of ranking correlations with respect to the number of blocks. Higher rates of correlation growth show faster learning by the subjects. To compare the learning rates, we consider the maximum slope of each curve (table II).

The results indicate a higher rate of learning for users in the presence of an explanation. The explanation from the SVQA agent causes a higher increase in the learning rate compared to SOBERT. However, the ultimate value of ranking correlation in the SVQA model is bound to $\delta = 0.808$ while the SOBERT model approaches the maximum correlation at $\delta = 1.0$ (figure 7).

V. Discussion

In these user studies, the overall progress of ranking correlations is measured as a metric to evaluate the users mental model of system competency. We test the user's mental model after they only see 100 instances (trials) of the AI agent's performance. However, the results strongly suggest that even with this limited view of system performance, the subjects learn the overall competency of AI agents throughout these tests.

Adding the attentional explanations for both models results in a significant improvement over competency rankings. Comparing the early learning rates between baseline and explanation groups suggests a significant improvement by attention map explanations especially for the SVQA model. However, the SVQA learning curve suggests an upper bound to the correlation in the presence of explanations. On the other hand, the SOBERT model shows a higher learning rate with explanations compared to the baseline while still

reaching the maximum value of correlation.
These results highlight the effect of input features on the information that the explanations can carry. The SOBERT model uses object and spatial features vs. the spatial features in the SVQA model. The SOBERT model also uses BERT to transfer the features into attention maps. These changes w. r. t. the SVQA has raised the upper bound on the maximum reachable competency prediction by the subjects.

VI. Conclusion

In this paper, we evaluate the role of attention map explanations on the users mental model of AI competency. We designed an experiment where subjects rank the performance of the VQA model among four different types of questions. To quantify the subjects mental model, we compute the correlation between user rankings and AIs actual ranking among the question types.
We propose a new XVQA model that produces answers and attention maps from spatial and object features of the image. This explainable model uses a BERT language module to better process the visual and textual embeddings of the input. The proposed model is compared with a baseline model to show the effect of input object features and also the BERT attention module.
Overall results from the experiments suggest an improvement in the user mental model when exposed to the attention map explanations. The progress of the user mental model (ranking correlations) throughout the experiments indicates a higher learning rate in the presence of explanations. Furthermore, the subject group interacting with the newly proposed model shows a higher rate of ranking correlation compared to the baseline model. This improvement suggests a positive impact on the explanations by including the object feature and the BERT language model.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[2] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[3] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[4] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[5] A. Chandrasekaran, D. Yadav, P. Chattopadhyay, V. Prabhu, and D. Parikh, "It Takes Two to Tango: Towards Theory of AI's Mind," *arXiv preprint arXiv:1704.00717*, 2017.

[6] K. Alipour, J. P. Schulze, Y. Yao, A. Ziskind, and G. Burachas, "A Study on Multimodal and Interactive Explanations for Visual Question Answering," Tech. Rep. [Online]. Available: https://arxiv.org/abs/2003.00431

[7] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[8] Y. Zhang, J. C. Niebles, and A. Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 349–357.

[9] A. Ray, K. Sikka, A. Divakaran, S. Lee, and G. Burachas, "Sunny and dark outside?! improving answer consistency in vqa through entailed question generation," *arXiv preprint arXiv:1909.04696*, 2019.

[10] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4223–4232.

[11] ——, "Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge," *CoRR*, vol. abs/1708.0, 2017. [Online]. Available: http://arxiv.org/abs/1708.02711

[12] H. Xu and K. Saenko, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," *CoRR*, vol. abs/1511.0, 2015. [Online]. Available: http://arxiv.org/abs/1511.05234

[13] ——, "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 451–466.

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[15] V. Kazemi and A. Elqursh, "Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering," *CoRR*, vol. abs/1704.0, 2017. [Online]. Available: http://arxiv.org/abs/1704.03162

[16] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Tech. Rep. [Online]. Available: https://github.com/tensorflow/tensor2tensor

[17] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," *CoRR*, vol. abs/1606.0, 2016. [Online]. Available: http://arxiv.org/abs/1606.00061

[18] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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*, 2018, pp. 6077–6086.

[19] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0. 1: the winning entry to the vqa challenge 2018," *arXiv preprint arXiv:1807.09956*, 2018.

[20] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] Y. Zhang, J. Hare, and A. Prügel-Bennett, "Learning to count objects in natural images for visual question answering," *arXiv preprint arXiv:1802.05766*, 2018.

[22] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6720–6731.

[24] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Rule-based expert systems*, pp. 233–262, 1984.

[25] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, and D. Gomboc, "Explainable artificial intelligence for training and tutoring," UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE∼…, Tech. Rep., 2005.

[26] M. Van Lent, W. Fisher, and M. Mancuso, "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the national conference on artificial intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 900–907.

[27] M. Lomas, R. Chevalier, E. V. Cross II, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the*

seventh annual ACM/IEEE international conference on Human-Robot Interaction. ACM, 2012, pp. 187–188.

[28] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[29] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[30] Z. Jiang, Y. Wang, L. Davis, W. Andrews, and V. Rozgic, "Learning discriminative features via label consistent neural network," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 207–216.

[31] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[32] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 264–279.

[33] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8779–8788.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[35] J. Drozdal, J. Weisz, D. Wang, G. Dass, B. Yao, C. Zhao, M. Muller, L. Ju, and H. Su, "Trust in automl: Exploring information needs for establishing trust in automated machine learning systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI 20. New York, NY, USA: Association for Computing Machinery, 2020, p. 297307. [Online]. Available: https://doi.org/10.1145/3377325.3377501

[36] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, "Is seeing believing?: how recommender system interfaces affect users' opinions," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2003, pp. 585–592.

[37] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more?: the effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1–10.

[38] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[39] A. Ray, G. Burachas, Y. Yao, and A. Divakaran, "Lucid Explanations Help: Using a Human-AI Image-Guessing Game to Evaluate Machine Explanation Helpfulness," *arXiv preprint arXiv:1904.03285*, 2019.

[40] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ser. IUI 19. New York, NY, USA: Association for Computing Machinery, 2019, p. 275285. [Online]. Available: https://doi.org/10.1145/3301275.3302310

[41] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 29–38.

[42] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez, "An evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1902.00006*, 2019.

[43] A. Chandrasekaran, V. Prabhu, D. Yadav, P. Chattopadhyay, and D. Parikh, "Do explanations make VQA models more predictable to a human?" *arXiv preprint arXiv:1810.12366*, 2018.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[45] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Tech. Rep. [Online]. Available: https://github.com/

[47] A. Heathcote, S. Brown, and D. Mewhort, "The power law repealed: The case for an exponential law of practice," *Psychonomic bulletin & review*, vol. 7, no. 2, pp. 185–207, 2000.

[48] F. E. Ritter and L. J. Schooler, "The learning curve," *International encyclopedia of the social and behavioral sciences*, vol. 13, pp. 8602–8605, 2001.