Can You Explain That? Lucid Explanations Help Human-AI Collaborative Image Retrieval

Arijit Ray, Yi Yao, Ajay Divakaran, Giedrius Burachas
SRI International, Princeton, NJ, USA

Abstract

While there have been many proposals on making AI algorithms transparent and explainable, few have attempted to evaluate the impact of AI-generated explanations on human performance in conducting human-AI collaborative tasks. To bridge the gap, we propose a Twenty-Questions style collaborative image retrieval game, Explanation-assisted Guess Which (ExAG), as a method of evaluating the efficacy of explanations (visual evidence or textual justification) in the context of Visual Question Answering (VQA). In our proposed ExAG, a human user needs to guess a secretly picked image by the VQA agent by asking natural language questions. We show that when AI explains its answers, users succeed more often in guessing the secret image correctly. Furthermore, we show that while good explanations improve human performance, incorrect explanations can degrade game performance as compared to no-explanation games. Notably, a few correct explanations can readily improve human performance in game rounds where the AI system's answers are mostly incorrect as compared to no-explanation games. Our experiments, therefore, show that ExAG is an effective means to evaluate the efficacy of AI-generated explanation on human-AI collaborative tasks.

Introduction

Deep networks, as black-box models, often suffer from low reliability and trustworthiness due to their lack of interpretability. In the context of Visual Question Answering (VQA) (Antol et al. 2015), the task of answering natural language questions on images, various methods for shedding light on the inner workings of these networks have been proposed — pointing to evidence in the image and/or question for the answer (Lu et al. 2016; Kazemi and Elqursh 2017; Xu and Saenko 2016; Selvaraju et al. 2016) and human-readable text-based justifications (Park et al. 2018), to mention a few. However, the empirical evidence that such explanations can actually be meaningful and useful for a human-machine collaborative task is lacking.

To this end, we propose a Twenty-Questions (Lewis 2008) style human-machine collaborative game, Explanation-Assisted GuessWhich (ExAG), using VQA (Antol et al. 2015) as the backbone task to evaluate the efficacy of explanations. An explanation, in this context, refers to additional information that sheds light on the reasoning of a VQA agent for generating an answer given a question-image pair. For example, if the answer to a question "what is in the image?" is "car", an explanation could be pointing to salient components of cars such as the wheels and wind-shields. In ExAG, human users and VQA agents collaborate to retrieve a secret image, selected by the agent out of a set of visually similar images. The role of the agent is to help the human identify the secret image by answering questions asked by

Figure 1: We propose the ExAG game, where a human user needs to guess the secretly picked image by a VQA agent via asking natural language questions. The VQA agent answers "5" to the question "How many people?". Without any explanations. As a result, the user finds it difficult to judge the correctness of the agent's answer and hence, the first image seems an obvious choice as the secret image. The explanations (visualizations in bottom row) point out the critical evidence that the agent probably mistakenly sees 5 people in the second image as well. The user can take this into consideration while asking follow-up questions.

the human. Since the VQA is noisy in its answer predictions, finding the correct secret image requires humans to build a proper mental model of the VQA agent in order to decide which answer to trust. This makes ExAG a promising candidate framework for evaluating the efficacy of explanations. Our hypothesis is that humans will succeed more often (i.e., winning rate) and quicker (i.e., using fewer questions) in finding the secret image when the machine explains its reasoning.

We conducted two sets of ExAG games and collected user performance as a function of their usage of explanations. The first set (at-will setting) allowed users to choose the use of explanations at will. This set of experiments provided preliminary evidence that human users spontaneously and increasingly prefer explanations even when their usage is penalized in the final score. The second set was conducted with a more controlled design to study the efficacy of each mode of explanations using a tighter metric (controlled setting). We collected subjective ratings of explanation helpfulness perceived by the users before the outcome of the game was revealed (to avoid the influence of game win or loss on the perceived helpfulness). We also independently collected subjective ratings of explanation correctness for better understanding of the relationship between explanation efficacy and quality.

We show evidence that the ExAG game performance correlates to perceived explanation helpfulness and correctness ratings, making ExAG a suitable tool for evaluating explanation efficacy and quality. Compared to no-explanation games, while helpful explanations increase game performance significantly, unhelpful explanations do not. Moreover, incorrect explanations can degrade game performance. Interestingly, we also note that having a few correct explanations can help performance significantly in game rounds where answers are mostly incorrect.

Practical applications of our proposed ExAG can include image retrieval using free-form queries. For example, assisting disaster personnel, where a rescuer may have to rely on audio answers from a VQA machine because he/she is too busy to look at a video/image feed. It can also help medical professionals, where a doctor may use visual explanations to judge the confidence of a certain diagnosis among others.

Related Work

Explainable AI Early work on explainable models involves template-based systems that spanned from medical systems (Shortliffe and Buchanan 1984) to educational settings (Lane et al. 2005; Van Lent, Fisher, and Mancuso 2004). Recent interest in explaining the inferences of a deep networks for computer vision applications includes introspective explanations that show the intermediate features of importance in making a decision (Lu et al. 2016; Park et al. 2018; Xu and Saenko 2016; Fong and Vedaldi 2018; 2017; Selvaraju et al. 2016; Zeiler and Fergus 2014), as well as post-hoc rationalization techniques such as justifying textual explanations (Park et al. 2018) and generating visual explanations (Hendricks et al. 2016). We focus on using attention-based visualization of important image regions (Lu et al. 2016; Xu and Saenko 2016; Kazemi and Elqursh 2017; Teney et al. 2017), object/scene detection (Szegedy et al. 2017; He et al. 2017) and related question answers (Ray et al.) for our game. (Das et al. 2017a) shows that humans and machines look differently at images when answering questions. Hence, it is not clear whether such explanations are indeed helpful to humans. In this paper, we try to quantify how much these explanation modes help in human-machine collaboration performance.

Visual Question Answering We use Visual Question Answering (VQA) (Antol et al. 2015) as the core underlying technology for our human-AI collaborative game. VQA was introduced by (Antol et al. 2015) as an AI complete task for image and text understanding. Most of the effective approaches to VQA consist of works with attention on image features (Yu Jiang* et al. 2018; Lu et al. 2016; Teney et al. 2017; Xu and Saenko 2016; Kazemi and Elqursh 2017) guided by the question in order to answer it. We implement a custom model that attends to both objects and free-form spatial regions in the image to answer the question similar to (Yu Jiang* et al. 2018).

20 Questions Game Our choice of the image-guessing game is the visual version of the popular 20-questions game, which is more formally, a specific version of the classic Lewis Signaling Game (Lewis 2008). There have also been efforts at training AI agents to play such an image-guessing game with humans/AI's (de Vries et al. 2016) using reinforcement learning- (Das et al. 2017b). (Chattopadhyay et al. 2017) used this game to evaluate visual conversational agent performance. However, to our knowledge, we are the first ones to use such a game to evaluate the effect of VQA explanations on human-machine team performance.

Mental Model of an AI System Along the lines of quantifying explanation efficacy, (Chandrasekaran et al. 2017; 2018) quantifies whether attention-based explanation improves human prediction of VQA system answer correctness and accuracy. While they (Chandrasekaran et al. 2017; 2018) show no significant increase in the ability to predict model outcome using attention-based explanations, we show that only good combined attention + related question-answer explanations are helpful in a game setting where multiple rounds of question-answering are involved. We also show that erroneous explanations are severely hurtful to game performance. (Park et al. 2018) show that showing both attention-based and textual explanations is helpful for predicting model performance. We use related question-answers as a form of textual explanation and also see similar trends for such a collaborative question-answering task.

Game Outline

In our game setting, there are two agents: a near state-of-the-art VQA deep learning model trained to answer questions about images (the "VQA agent"), and a human volunteer (the "player") who has to guess a secret image picked by the machine. A secret image is randomly picked from a pool of 1500 images. We select another $N-1$ images from the same pool using a difficulty measure based on the VGG16 (Simonyan and Zisserman 2014) FC7 distance. The difficulty level is adjusted so that the N image set is challenging enough to where it requires multiple rounds of VQA to identify the secret image. The player starts with P_o points and is allowed to ask free-form questions to the VQA agent in order to guess the secret image. Each question costs one point. The final score is $P = P_o - \sum_{i=1}^{Q} p_i$ if the correct image is guessed, where Q is the number of questions asked and p_i is the point deduction for each question. Wrong image guess gets a score of $P = 0$. A success is defined as the player correctly selecting the secret image while keeping $P > 0$. Players are encouraged to keep P as high as possible.

The VQA Model

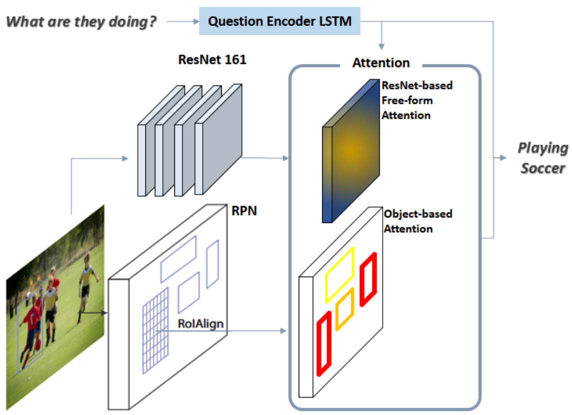

Figure 2: Mask-RCNN-based VQA Model. This network has attention on both global ResNet features (i.e., spatial attention) and region proposal network (RPN) features (i.e., object attention). This model is based on (Kazemi and Elqursh 2017) and (Teney et al. 2017)

As shown in Figure 2, we used a near state-of-the-art VQA model that comprises both ResNet- (Szegedy et al. 2017) and Mask-RCNN-based (He et al. 2017) image encoders and an attention mechanism to weigh the visual features depending on the question asked. The weighted features are fed into an answer classifier that predicts an answer from 3000 candidates (Kazemi and Elqursh 2017; Teney et al. 2017; Yu Jiang* et al. 2018). The at-will setting uses a model with only the ResNet-based image encoder. More details about the architecture and training are included in supplementary.

Modes of Explanations

We define explanations as information that provides insight into why the VQA predicted a certain answer. The insight can be visualizations of the evidence used to infer the answer, such as weights applied to visual features (i.e., attention). It can also be rationalizations, such as stating the semantic beliefs about a fact that led to the answer. Below, we outline three modes of explanations and illustrate how they are used in our ExAG game.

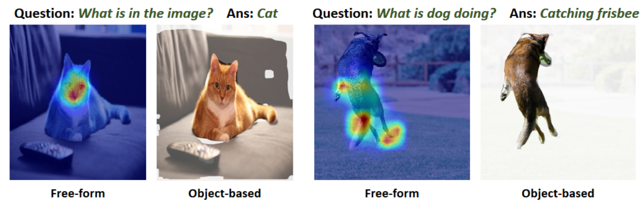

Figure 3: Explanation mode based on attentions that highlight the relevant regions and objects in image to support the machine-generated answers for the given questions.

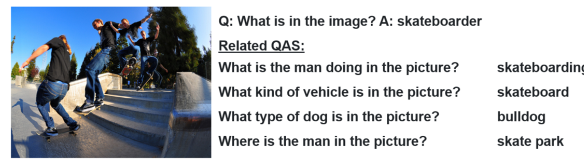

Figure 4: Explanation mode based on related questions and answers. The given question is “What is in the image?” and the machine-generated answer is “skateboarder”.

Attention. We employ attention masks computed based on the question asked to highlight spatial locations/objects in the image that are weighted more heavily in the inference process of answer prediction. We employ two types of attention layers - free-form attention that weighs visual features in the pixel-space and object-based attention that weighs object proposals in the image (Figure 3). A player can check if the attention masks correspond to the relevant part of the image given the question to determine if the machine generated answer is trustworthy.

Object and Scene Predictions (ObjScene). We display a list of the most relevant object and scene predictions observed in the image. The relevance of an object/scene word O is measured by $S(O) = \frac{dist(O,A)}{p(O|I)}$ where I and A denote the image and machine generated answer word, respectively, and $dist(O, A)$ denotes the Word2Vec (Mikolov et al. 2013) distance of object and answer words. $p(O|I)$ is calculated using the image encoder in VQA. When object-based attention is used, ObjScene explanation mode is skipped since objects are highlighted by attention masks already. Under this explanation mode, the player needs to judge if the listed relevant objects are consistent with the visual contents in the image and machine-generated answer.

Related questions and answers (RelQAs). Five questions that are semantically close to the given question are retrieved from the VQA2.0 Validation Dataset. The closeness is measured via a semantic similarity based on the averaged Word2Vec distance over all the words in the pair of question and answer. Furthermore, questions with a high word overlap are rejected to avoid paraphrases (eg, ‘what is in the image?’ vs. ‘what is in the picture?’). The VQA agent generates answers to these related questions as part of the explanation.The player needs to judge the trustworthiness of machine-generated answer based on the correctness and co-

Table 1: Human performances of ExAG in the controlled settings with different explanation modes. Overall, explanations improve game performance in terms of both win rate and averaged score. Explanations can help human players identify the secret image not only correctly but also with fewer questions (as reflected by higher scores).

	With Expl		No Expl		Group Baseline		Overall Improv		Stat Sig	
	Score	**Win Rate**	**Score**	**Win Rate**	**Score**	**Win Rate**	**Score**	**Win Rate**	**p**	**conf**
Attention	6.23	66.67	6	64.92	5.66	62.1	0.66	5.52	0.1	none
Rel QAs	**6.8**	**71.48**	**6.03**	**64.54**	**6.02**	**65.45**	**1.23**	**10.33**	**0.0019**	**99%**
Both	6.44	69.03	5.83	63.25	5.68	63.75	0.87	7.88	0.03	90%
Overall	6.52	69.29	5.97	64.3	5.81	63.85	0.95	8.14	0.0015	99%
Overall inc no	**6.52**	**69.29**	**5.74**	**61.85**	**5.57**	**61.15**	**0.95**	**8.14**	**0.0015**	**99%**

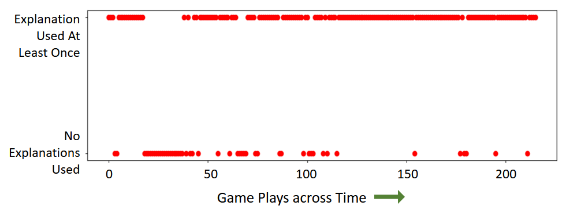

Figure 5: When given a choice, players increasingly opt for explanations even when explanation usage is penalized - extra 2 points per question-answering round.

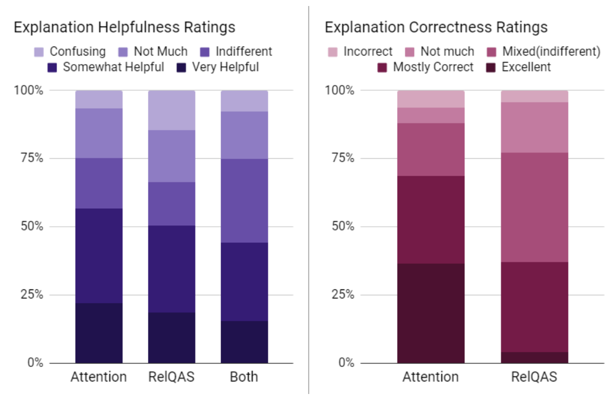

Figure 6: Histogram of ratings of how "helpful" and "correct" explanations were while playing the game. "Helpfulness" ratings were given by workers while playing games before knowing the game outcome. "Correctness" was rated by three independent workers for explanations of the secret image.

herence of these pairs of related questions and answers (Figure 4).

Game Settings

At-will Setting. In this setting, we show $N = 20$ images and the player has the option of receiving explanations for the answer or not. Each question asked costs one point and explanation, if requested, costs an additional two points. All explanation modes are shown once the player chooses to receive explanations. This includes spatial attention for all images, ObjScene for the secret image, and RelQAs for the secret image. To be helpful of game wins, explanations need to be not only correct/coherent with the given answer but also sufficiently subtle/complete against distractor images.

Controlled Setting. In this setting, we show spatial attention, object-based attention and RelQAs for all the images. Since extra information are given for all the images, the coherence between the explanation and given answer plays more important role in assisting game performance. In order to reduce the cognitive load on the players, we reduce image set size from $N = 20$ in the at-will setting to $N = 5$ and make the images more similar to each other to maintain the same level of difficulty. Since we use object-based attention, we omit the ObjScene explanations to avoid repetition. We randomly assign each explanation type to a group of AMT workers. As users play the game, they are also asked to rate how helpful the explanations are after each round of question answering. Note that at the time of rating, players do not yet know the outcome of the game or the secret image. So, their rating is not confounded by whether they succeeded or not, but reflects how helpful the explanations seemed in narrowing down the secret image and identifying the proper question to ask next.

Results

At-will Setting

In this setting, the ExAG game was played in a competitive setting (with cash rewards for the team that won the most) by about 60 people grouped into 6 teams. The players were free to choose explanations or forgo them.

Of 206 total games played, the average win rate was 43%. We divided the game plays into games where explanations were never used ($N = 49$) and those where explanations were used at least once ($N = 157$). The win rate with explanations was 47% and 28% when explanations were never used. The z-test for proportions indicates that this is a statistically significant difference at 95% confidence level (p=0.019).

Moreover, as the players proceeded to play the ExAG game, they tended to opt for using explanations even though that resulted in additional 2 point penalty. Figure 5 shows this spontaneous adoption of explanation with increasing number of plays. A z-test comparing the proportion of

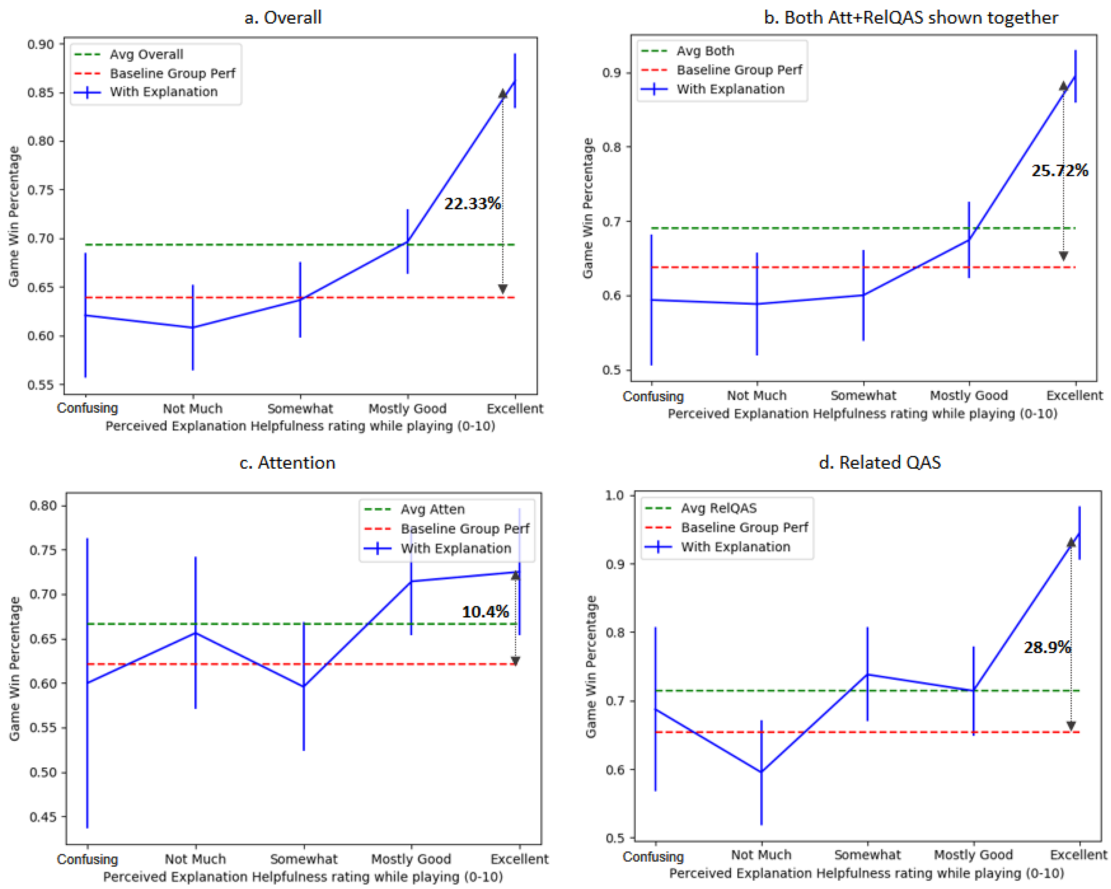

Figure 7: Game win percentage as a function of user perceived helpfulness rating. Baseline is first 5 no-explanation performance for the same group of workers. Helpfulness was self rated by workers before they knew the GT image or the outcome of the game. We see that overall, *excellent* helpful explanations significantly improve performance.

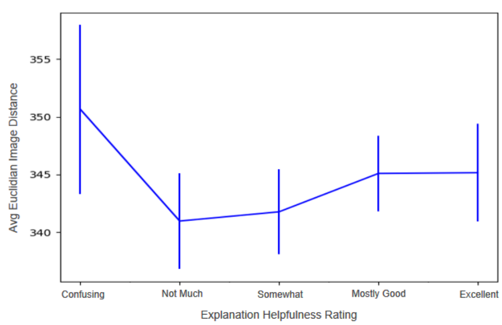

Figure 8: Explanations are rated more helpful for difficult image sets where distractor images are visually more similar to the secret image.

games using the explanations during the first half of plays (61.2%, $N = 103$) vs. during the second half (91.3%, $N = 103$) indicates highly statistically significant increase in explanation use ($p < 10^{-5}$).

Controlled Setting

This setting shows explanations for all images and hence, players have to solely rely on the coherence of explanations to the answer to aid their game (if explanations are good, the secret image's explanation will be the most coherent with answer shown). Extra random information about the secret image cannot be misused as in the at-will setting, hence, making this setting more challenging. All analyses henceforth are performed on this setting.

The games were played by 69 individual AMT workers. The instructions were to try to win the game by guessing the secret image correctly and were warned that a lack of effort would lead to rejection. For AMT worker selection, worker qualification threshold was set to above 98% (number of HITs > 1000) to ensure the quality of game plays. The workers played 1469 games in total covering four explanation modes - attention, RelQAs, both shown together (referred to as 'Both'), and also without any explanations.

Each individual worker always saw only one mode of explanation while playing. Each worker was instructed to play at least four sessions with each session consisting of five games. Sessions with and without explanations were

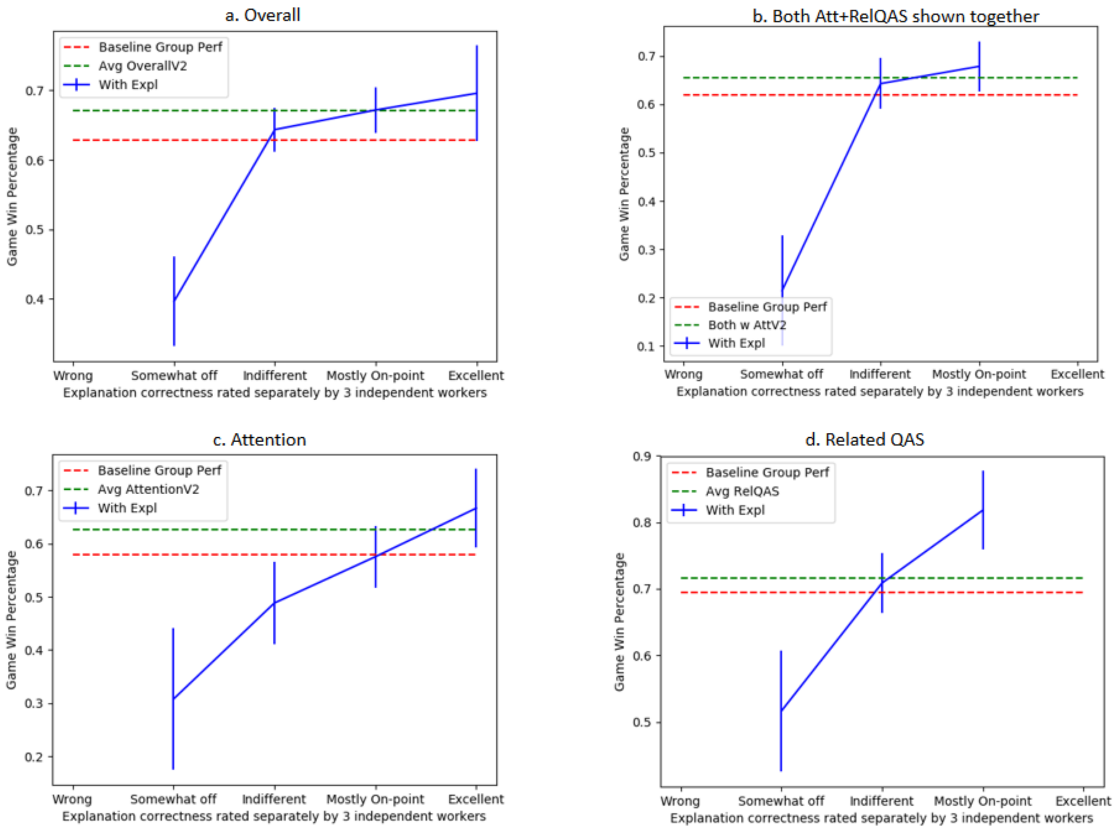

Figure 9: Game win percentage as a function of independently rated explanation accuracy. Baseline performance was adjusted to reflect the baseline of only that subset of games where correctness ratings were collected. Overall, we see that incorrect explanations hurt performance, while correct explanations are not sufficient to improve performance compared to no-explanation games.

alternated. For instance, the first session does not provide explanations, followed by a session with assigned choice of explanations. The first block is used as the baseline no-explanation performance for the assigned explanation worker group.

Overall impact of explanations. The overall impact of explanations on the performance of the ExAG game is summarized in Table 1. For each of the modes, 'With Expl.'and 'No Expl.' list the performance with and without explanations, respectively. 'Group baseline' is the baseline performance without explanations (the first 5 rounds) of that worker group. When reporting the 'Overall inc no' results, the 'No Expl' performance includes performance of workers that never saw any explanations. 'Overall Improv' shows the improvement with respect to 'Overall inc no'. The game score starts from 10 ($P_o = 10$) with each question asked costing 1 point ($p_i = 1$). We observe that the game win rate is statistically significantly ($p = 0.0015$, 99% confidence) improved by explanations overall and fewer questions were required to guess correctly (as suggested by higher scores). Consistent with prior reports (Chandrasekaran et al. 2017; 2018), we don't see a very significant effect of attention-based explanations when presented in isolation. However, showing 'RelQAs' explanations improves game performance the most ($p = 0.0019$, 99% confidence), followed by 'Both' ($p = 0.03$, 90% confidence).

We also analyze how the difficulty level of the selected image set affects the game performance. We use the L_2 norm on the VGG16 FC7 feature (Simonyan and Zisserman 2014) to evaluate the distance between each images and the secret image and the averaged distance as the overall difficulty level of the set. We observe that for game rounds with explanations, the average difficulty level is similar between winning and losing game rounds. However, for game rounds without explanations, the difficulty level of winning games is much lower than that of the losing games. This shows that explanations help players to identify subtle cues which may not be visually salient but critical to differentiate the secret image.

To understand the effect of explanation quality on performance, we collect two types of ratings for the explanations 1) While playing the game, workers are asked to self rate their "perceived helpfulness" of the explanation for zeroing down on an image after each question asked. Note that the workers do not know the secret image or the outcome of the game while rating. 2) We separately collect the correctness

Figure 10: Explanations help when VQA answers are wrong. Without explanations (red line), if the answer from the VQA is wrong, user performance drops dramatically. However, at least a few good explanations (black line) help reveal VQA answer correctness so that it can be taken into account. Hence, game performance without explanations is much lower when answers are wrong than with explanations. The black line is defined as games where there was at least one explanation above *indifferent*.

of the explanations and answers from 3 independent AMT workers by showing them the answer and explanation for the asked question and the secret image. Below, we use these ratings to analyze how "perceived helpfulness" and "correctness" of explanations and answers affect game performance.

Impact of explanations as a function of perceived helpfulness ratings. We collect self-reported helpfulness ratings of explanations as workers receive answers and explanations for their questions (Figure 6), while the play outcome is not yet known in order to avoid circularity in the ratings. We analyze game performance as a function of these ratings to see if explanations perceived as "helpful" do help game performance. Since workers didn't know the secret image or the game outcome while rating, their decision was not affected by the (future) play outcome and likely reflected the helpfulness of explanations in narrowing down their image choices. The workers were asked to rate the explanations according to the following options: "Helping a lot" (refered to as *Excellent* here-on), "Mostly helping" (*Mostly Good*), "Somewhat Helpful" (*Somewhat*), "Not helping much" (*Not Much*) and "Completely Confusing" (*Confusing*). The histogram of the ratings across all the games is shown in Figure 6 and the explanations are mostly perceived as helpful.

We see that overall (Figure 7a), explanations perceived as *Excellent* significantly increase game performance. When explanations are rated less helpful, performance is similar to playing without explanations supporting the notion that workers seem to ignore the explanations, or somewhat degraded as the workers were confused by them. Notably, we calculate the average difficulty levels of the candidate image sets for each game round and further observe that explanations were rated more helpful when image sets were difficult (i.e., images are more visually similar to the secret image, Figure 8).

Next we break down explanation helpfulness-dependent performance by explanation modes. Figure 7b shows that combining attention and RelQAs improves performance significantly for explanations rated as *Excellent*, but hurts performance slightly when rated below *Somewhat*. Figure 7c indicates that attention-based explanations don't help much on their own, however, when they seem very helpful for making a choice, they do help game performance slightly. *Excellent* Related QAs were the most helpful for game performance when presented in isolation as in Figure 7d. We reason that this is due to the consistency of related question answers being a slightly better indicator of VQA answer accuracy. We calculate the correlation coefficient of the explanation correctness to answer correctness and observe that it is 0.37 for related question-answers as compared to 0.33 for attention. This combined with the ease of parsing textual attention than heatmaps make related question answers more effective.

Impact of explanations as a function of independent explanation correctness ratings We observe that although helpfulness of Attention is rated better on average than RelQAs (Figure 6), overall, Attention contributes lesser than RelQAs in terms of user performance improvement. To analyze further, we ran a separate AMT task to collect correctness ratings of the explanations for the secret image, the question that was asked and the answer given by the VQA model. Three independent workers were asked to rate the explanations for the given question and image according to the following options: "Exactly on-point" (referred to as *Excellent* hereon), "Mostly on-point" (*Mostly on-point*), "Indifferent" (*Indifferent*), "Somewhat off" (*Somewhat off*) and "Completely Wrong" (*Wrong*). The histogram of ratings is shown in Figure 6.

Figure 9 shows game performance as a function of explanation correctness ratings. We see that for modes with overall (9a), combined (9b) and attention-based (9c), correct explanations are not sufficient for improving game performance, while, incorrect explanations can severely degrade the performance. This suggests that giving incorrect explanations can make players disbelieve a correct answer and hence fail games. We observe that while RelQAs has very few *Excellent* cases, as long as they are *Mostly on-point*, they help game performance substantially (9d).

Impact of explanations as a function of VQA answer correctness. As noted before, we observed that correct explanations, on their own, do not necessarily help game performance. Further examination indicates that game performance is indeed influenced by the combined effect of the correctness of explanations and machine-generated answers. We conduct studies to disentangle the effect of VQA answer correctness. We collect the answer correctness ratings for the ExAG games through an independent AMT experiment. The workers were required to rate the answer for a given question and image pair as either "correct", "somewhat correct" or "completely wrong". We analyzed the game performance as a function of average VQA answer correctness in a game round where at least some of the explanations are rated as

Figure 11: An ExAG game round where explanations help the user in winning the game. Even though the most prominent 'toilet' is in the fourth image, explanations make it clear to the users that the fifth image could also be the secret image. Explanation suggest that the machine probably mis-detected the 'bathtub' or 'sink' for a 'toilet'. This hints the user to ask follow-up questions like "is there a sink?" to finally select the correct image.

correct (average rating above *indifferent*).

As displayed in Figure 10, we see that having at least a few correct explanations (black line), interestingly, helps user performance significantly ($p = 0.013$, 95% confidence) in games where the VQA answers are mostly wrong as compared to no-explanation games (red line). Such improvement was, however, not observed when explanations were mostly "incorrect" (average rating below *indifferent*). This suggests that having a few correct explanations helped workers in identifying incorrect answers which motivates them to ask clarification question and eventually leads to winning of the game. As expected, without explanations (red line of Figure 10), game performance degraded as VQA answers got less accurate, suggesting that a player had no way of telling if an answer was wrong or correct without the help of explanations.

An qualitative run of ExAG is shown in Figure 11. RelQAs explanations suggest that the VQA also understandably sees a 'toilet' in the fifth image. This prevents the user from selecting the obvious fourth image choice straight away and prompts him/her to further ask questions like "is there sink?", eventually resulting in him/her selecting the correct secret image. Examples of workers blindly trusting incorrect answer without explanations are included in the supplementary.

Conclusion

We propose the ExAG game as an evaluation framework for explanations and show that game performance correlates to the explanation helpfulness and correctness.

Our experiments provide empirical evidence that overall explanations help improve performance on a human-machine collaborative image guessing task. When analyzed by user self-rated "helpfulness" and independently-rated "correctness", helpful explanations (rated as *excellent*) significantly improve performance, while *incorrect* explanations degrade performance. Moreover, since the self-rated helpfulness is not influenced by the outcome of the game, this suggests that users can use their insight into explanation helpfulness to decide when to trust and include them in decision making process for choosing the image.

We also note that "correct" explanations, interestingly, help significantly when machine predictions are noisy (Figure 10). Without explanations, users blindly trust incorrect machine predictions, which hurts game performance. However, with explanations, users ask follow-up questions and are able to figure out the correct answer based on at least one correct explanation.

We believe that our ExAG framework can help in designing more accurate and helpful explanations that improve human-AI collaboration outcomes in terms of performance, trust and satisfaction.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; and Zheng, X. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Chandrasekaran, A.; Yadav, D.; Chattopadhyay, P.; Prabhu, V.; and Parikh, D. 2017. It takes two to tango: Towards theory of ai's mind. *arXiv preprint arXiv:1704.00717*.

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

Chattopadhyay, P.; Yadav, D.; Prabhu, V.; Chandrasekaran, A.; Das, A.; Lee, S.; Batra, D.; and Parikh, D. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

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

Das, A.; Kottur, S.; Moura, J. M.; Lee, S.; and Batra, D. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*.

de Vries, H.; Strub, F.; Chandar, S.; Pietquin, O.; Larochelle, H.; and Courville, A. C. 2016. Guesswhat?! visual object discovery through multi-modal dialogue. *CoRR* abs/1611.08481.

Fong, R., and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Fong, R., and Vedaldi, A. 2018. Net2Vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2980–2988. IEEE.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *European Conference on Computer Vision*, 3–19. Springer.

Kazemi, V., and Elqursh, A. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

Lane, H. C.; Core, M. G.; Van Lent, M.; Solomon, S.; and Gomboc, D. 2005. Explainable artificial intelligence for training and tutoring. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES.

Lewis, D. 2008. *Convention: A philosophical study*. John Wiley & Sons.

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

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; and Rohrbach, M. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1802.08129*.

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

Ray, A.; Burachas, G. T.; Sikka, K.; Roy, A.; Ziskind, A.; Yao, Y.; and Divakaran, A. Make up your mind: Towards consistent answer predictions in vqa models.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *See https://arxiv. org/abs/1610.02391 v3* 7(8).

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

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, 12.

Teney, D.; Anderson, P.; He, X.; and Hengel, A. v. d. 2017. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*.

Van Lent, M.; Fisher, W.; and Mancuso, M. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence*, 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Xu, H., and Saenko, K. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, 451–466. Springer.

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

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Appendix

VQA Architecture Details

We use Tensorflow (Abadi et al. 2015) for all our implementations. The network we use for At-will game setting is shown in Figure 12. Outlined below are the network details:

- Input Image - size 448×448. We center crop all images to the mentioned size during training and reshape dur-

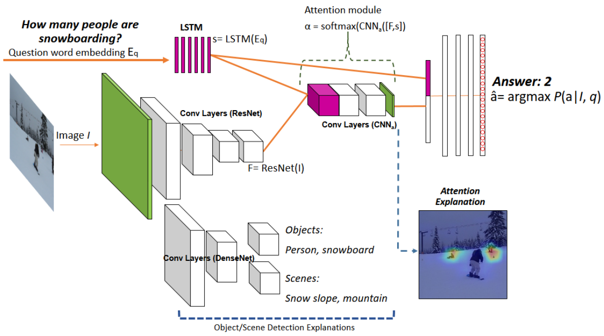

Figure 12: ResNet-based VQA model used in the At-will game setting. This network attends to ResNet 161 features. This model is based on (Kazemi and Elqursh 2017).

ing evaluation. We encode the image using a ResNet 161 (Szegedy et al. 2017) network.

- Question Input - Each question word is encoded using the Glove (Pennington, Socher, and Manning 2014) 300 dimensional embeddings before feeding into an LSTM word by word. We take the final 512 dimensional LSTM features as the question feature. Embeddings are fine-tuned.
- Attention- For Resnet feature attention, We tile the question features (512) into $14 \times 14 \times 512$ and concatenate with $14 \times 14 \times 2048$ image features. Attention predicts a set of weights in the shape of 14×14 using a 2-D convolutional layer.
- Answer classifier- We concatenate weighted flattened ResNet features and question features and pass it through a fully connected layer to get 3000 answer logits. We compute a softmax to get probabilities.

For Controlled Game Setting, we use a VQA network as shown in Figure 2 in main paper. Details are as follows:

- Input Image - For ResNet161 (Szegedy et al. 2017) image features, we center crop images to 448×448 similar to Network 1 to get $14 \times 14 \times 2048$ dimensional features. We also use the Region Proposal Network from Mask RCNN (He et al. 2017) to generate 100 object proposals per image. The input image size to Mask RCNN is 1024×1024 and the images are re-sized without cropping. We pool 1024 dimensional features from each of the 100 proposal boxes.
- Question Input - Each question word is encoded using the Glove (Pennington, Socher, and Manning 2014) 300 dimensional embeddings before feeding into an LSTM similar to network used in the at-will setting. The embeddings are fine-tuned.
- Attention- We have two attention modules - one for attending to the ResNet features (same as Network 1) and one for attending to the 100 object proposals in the image (object attention). Resnet feature attention is same as at-will setting. For object proposal attention, we concatenate the question features (512) to each 1024 image feature for the 100 proposals. Attention predicts a weight of shape 100×1024 using a 1-D convolutional layer.
- Answer classifier- We concatenate weighted flattened ResNet features, averaged weighted object features and question features and pass it through a fully connected layer to get 3000 answer logits. We compute a softmax to get probabilities.

Qualitative Results

Figure 13 and Figure 14 show qualitative examples of plays without and with explanations and how explanations may help in choosing the correct image more often when VQA answers are noisy. The anonymized game is available at https://gwapp.nautilus.optiputer.net/.

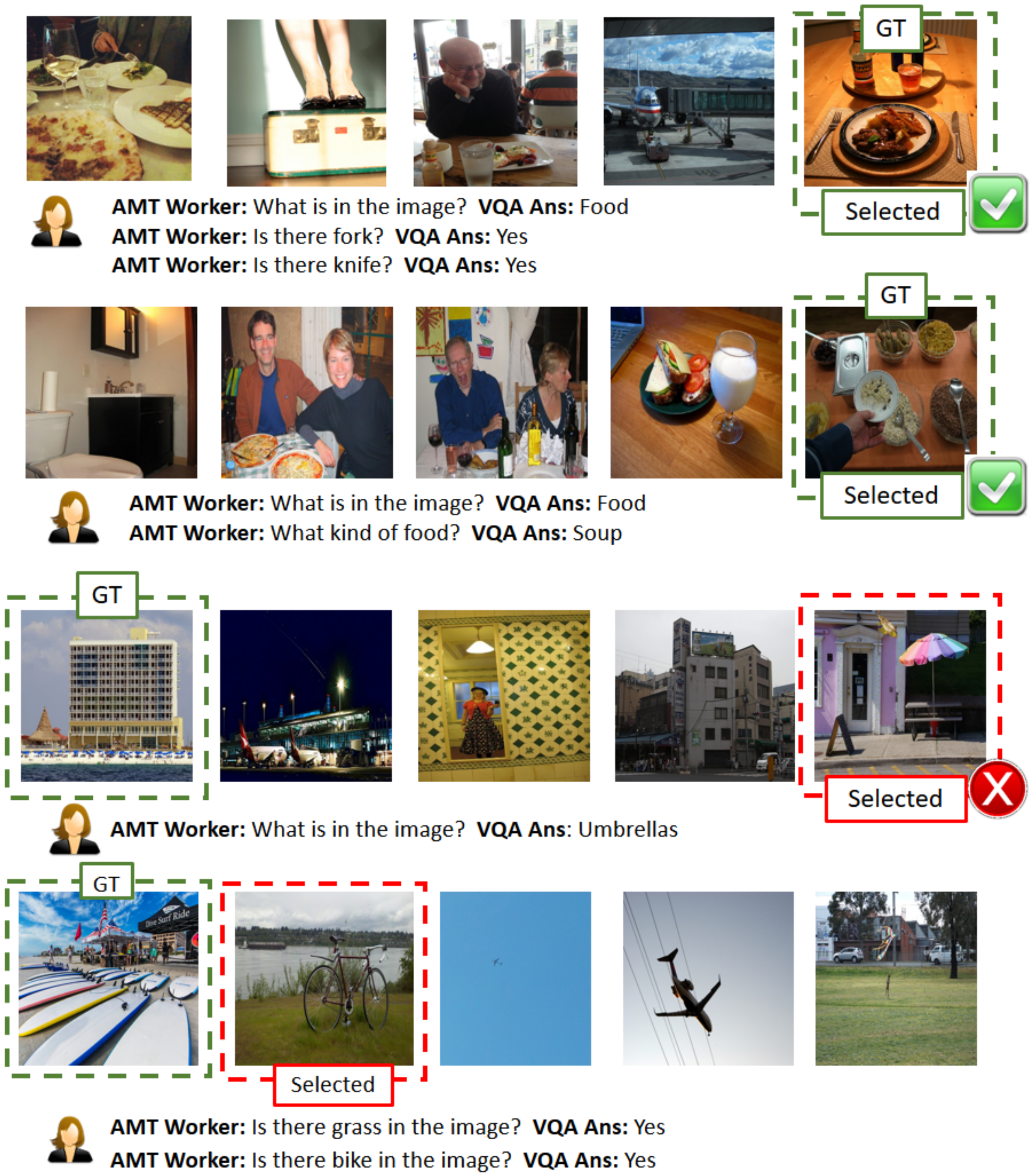

Figure 13: This image shows game plays without explanations (each row is a game-play example). When the VQA is fairly accurate, a user is easily able to pick out the correct image as shown in row 1 and 2. However, when the VQA answer is incorrect, it leads to a reasonably wrong selection by the user since the user has to blindly trust the VQA answer.

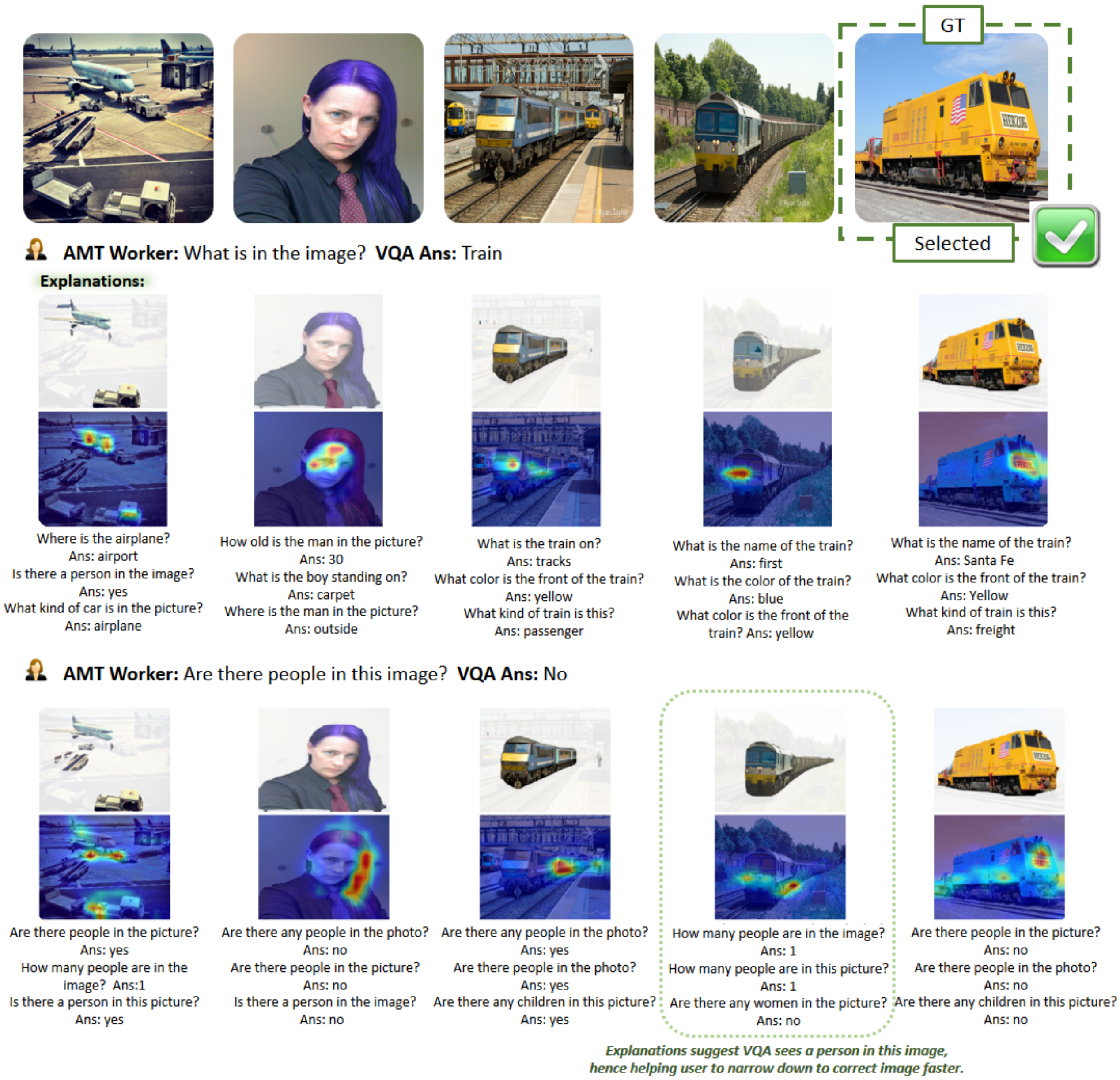

Figure 14: This image shows a game with both explanations. Note how the explanations help suggest that VQA sees one person in the 4th image choice although there is no salient person. This makes the user take that into consideration while guessing the secret image.