REX: Reasoning-aware and Grounded Explanation

Shi Chen Qi Zhao
Department of Computer Science and Engineering,
University of Minnesota
{chen4595, qzhao}@umn.edu

Abstract

Effectiveness and interpretability are two essential properties for trustworthy AI systems. Most recent studies in visual reasoning are dedicated to improving the accuracy of predicted answers, and less attention is paid to explaining the rationales behind the decisions. As a result, they commonly take advantage of spurious biases instead of actually reasoning on the visual-textual data, and have yet developed the capability to explain their decision making by considering key information from both modalities. This paper aims to close the gap from three distinct perspectives: first, we define a new type of multi-modal explanations that explain the decisions by progressively traversing the reasoning process and grounding keywords in the images. We develop a functional program to sequentially execute different reasoning steps and construct a new dataset with 1,040,830 multi-modal explanations. Second, we identify the critical need to tightly couple important components across the visual and textual modalities for explaining the decisions, and propose a novel explanation generation method that explicitly models the pairwise correspondence between words and regions of interest. It improves the visual grounding capability by a considerable margin, resulting in enhanced interpretability and reasoning performance. Finally, with our new data and method, we perform extensive analyses to study the effectiveness of our explanation under different settings, including multi-task learning and transfer learning. Our code and data are available at https://github.com/szzexpoi/rex.

1. Introduction

One of the fundamental goals in artificial intelligence is to develop intelligent systems that are able to reason and explain with the complexity of real-world data to make decisions. While explaining decisions is an integral part of human communication, understanding and reasoning, existing visual reasoning models typically answer questions without explaining the rationales behind their answers. As a result, despite the significantly increased accuracy achieved by powerful deep neural networks [2, 16, 21, 23, 26, 35], existing methods commonly take advantage of spurious data biases [27] and it is difficult to understand if they make decisions by truly understanding the causal relationships between multi-modal inputs and the answers.

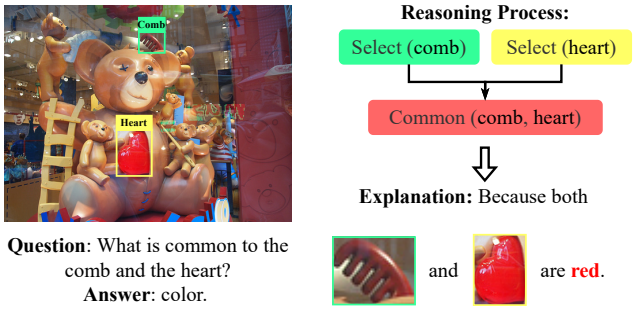

Figure 1. Illustration of our explanation that is derived from the reasoning process (with different reasoning steps color coded) and explicitly grounds key objects in the image.

An important line of research to tackle the issues is to improve the interpretability of visual reasoning models with multi-modal explanations [7, 22, 28, 31, 39, 40, 43]. While showing usefulness in highlighting important visual regions and providing user-friendly textual descriptions, these approaches suffer from two major limitations: (1) Existing explanations are typically defined in the forms of attention maps or free-formed natural language. Attention maps capture the salient regions for generating the answers but fall short of explaining how different regions contribute to the decision-making process. On the other hand, unconstrained textual explanations could be highly diverse and often inconsistent when explaining the same decision. Both of them lack the capability to illustrate the reasoning process behind a decision. (2) The explanations of different modalities are loosely connected and modeled with separate processes [22, 31, 40]. This not only undermines the capability of explaining models' rationales with multiple modalities, but can also result in contradictory explanations [40]. For

instance, textual explanations "The apple is above the pear" and "The pear is above the apple" have opposite meanings but could share the same attention map. We address the aforementioned challenges from two distinct perspectives (*i.e.,* data and model), and propose an integrated framework that consists of a new type of explanations, a functional program, and a novel explanation generation method.

From the data perspective, instead of independently modeling explanations of a single modality without considering the reasoning process, we introduce a new Reasoning-aware and grounded EXplanation (REX) that is derived by traversing the reasoning process and tightly coupling key components across the visual and textual modalities. As shown in Figure 1, it is constructed based on the consecutive reasoning steps (*e.g.*, select, common) for decision making, and explicitly grounds key objects (*e.g.*, comb, heart) with visual regions to elaborate how they contribute to the answer. The structured reasoning process also naturally alleviates the variance of natural language, and enables models to pay focused attention to important information for reasoning instead of the language structure. To automatically construct our explanations, we develop a functional program to progressively execute the reasoning steps and query key information from scene graphs [14, 19], and collect a new dataset with 1,040,830 multi-modal explanations.

From the model perspective, unlike existing methods [7, 22, 28, 31, 40] that model key components in different modalities with separate processes, we propose a novel explanation generation method that explicitly models the correspondence between important words and regions of interest. It takes into account the semantic similarity between features of the two modalities, and incorporates an adaptive gate to ground words in the visual scene. Our method improves the visual grounding by a large margin, resulting in enhanced interpretability and reasoning performance.

To summarize, our contributions are as follows:

(1) We present REX, a new type of reasoning-aware and visually grounded explanations. Our explanation differentiates itself with its strong correlation with the reasoning process and the tight coupling between different modalities. We develop a functional program to automatically construct our new dataset with 1,040,830 multi-modal explanations.

(2) We propose a novel explanation generation method that goes beyond the conventional paradigm of independently modeling multi-modal explanations [7, 28, 40], and leverages an explicit mapping to ground words in the visual regions based on their correlation.

(3) We demonstrate the effectiveness of our data and method with extensive experiments under different settings, including multi-task learning and transfer learning. We also analyze different visual skills and their correlation with the reasoning performance.

2. Related Works

This paper is related to previous efforts on visual question answering (VQA), multi-modal explanation datasets for visual reasoning, and explanation generation models.

Visual question answering. Visual reasoning is commonly framed as a VQA task. There is a large body of research on constructing VQA datasets [3, 4, 9, 12, 13, 29, 32, 43] and developing VQA models [2, 8, 10, 11, 16, 17, 21, 26, 33, 35, 42]. Early VQA datasets typically collect human-annotated questions through crowd-sourcing [3, 9, 43]. Several recent studies [12, 13] propose to use functional programs to automatically generate questions based on predefined rules and enable more balanced distributions of question-answer pairs. There is also an increasing interest in investigating different types of visual reasoning, *e.g.,* scene text understanding [4], reasoning on dynamic context [32], and knowledge-based reasoning [29]. These data efforts lead to the development of computational approaches for improving different components of VQA models, including multi-modal fusion [8, 17, 42], attention mechanism [2, 16], and inference process [10, 11, 33]. Vision-and-language pretraining [21, 23, 26, 35] has also shown usefulness in VQA for enhancing the understanding of multi-modality. Our framework is complementary to existing efforts in VQA. It augments existing VQA data with reasoning-aware and visually grounded explanations, and enables the development of VQA models with enhanced interpretability and reasoning performance.

Multi-modal explanations for visual reasoning. There is a dearth of studies that construct multi-modal explanation datasets for visual reasoning. The pioneering work [31] collects 41,817 textual explanations annotated by humans on the VQA datasets [3, 9], and develops a visual pointing task to highlight important regions. To automatically build large-scale explanation datasets, Li *et al.* [22] propose to convert captioning annotations [25] to textual explanations. By estimating the similarity between captions and questions, it generates 269,786 synthetic explanations. Zellers *et al.* [43] propose a multi-choice VQA dataset with 264,720 questions and each question is associated with a correct explanation. While the work has annotated regions of interest for the questions, as their dataset focuses on movie scenes, about 91% of the regions are related to human characters and over 40% of the explanations can not be grounded on them. The key differentiators of our explanation lie in its correlation with the reasoning process for problem-solving and the explicit coupling between the key components in different modalities. Our dataset offers 1,040,830 structured and visually grounded explanations that are aware of the decision-making process and ground various keywords in the images.

Generating multi-modal explanations. Aiming to develop visual reasoning models that are capable of explain-

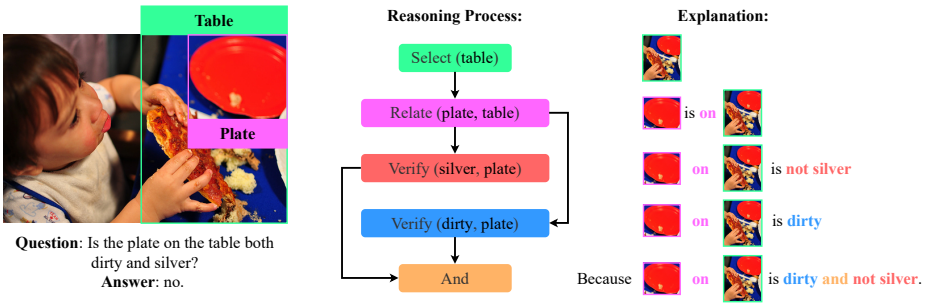

Figure 2. Illustration of the process for sequentially constructing our explanation. Partial explanations are shown to the right of the corresponding reasoning steps, and information collected from different steps is highlighted with their corresponding colors. The final explanation is obtained at the end of the reasoning process.

ing their answers, several works propose to automatically generate multi-modal explanations. Park *et al.* [31] use a long short-term memory (LSTM) model to generate the textual explanations, and highlight important visual evidence with attention maps. Later on, Wu *et al.* [40] improve the model by correlating the attention for answering questions and generating explanations. Li *et al.* [22] propose a multi-task learning paradigm to simultaneously generate answers and explanations. Instead of predicting explanations from scratch, Zellers *et al.* [43] adopt a multi-choice task setting whose goal is to select the correct explanation out of four candidates. Marasović *et al.* [28] develop an integrated method that incorporates pretrained language models with object recognition models. Dua *et al.* [7] frame both VQA and explanation generation as generative tasks, and sequentially generates words in answers and explanations. There are also studies [39, 41] that leverage generated or ground truth multi-modal explanations to improve the reasoning performance. Different from the aforementioned methods that independently model visual or textual explanations, our proposed method explicitly links words with the corresponding image regions based on their semantic similarity. The enhanced visual grounding capability brought by our method not only improves the interpretability and reasoning performance, but also plays a key role in distilling knowledge from explanations into question answering.

3. Reasoning-aware and Grounded Explanation

Answering visual questions would benefit from capabilities of reasoning on multi-modal contents and explaining the answers. This section presents a principled framework for visual reasoning with enhanced interpretability and effectiveness. It advances the research in visual reasoning from both the data and the model perspectives with: (1) a new type of multi-modal explanations that explain decision making by traversing the reasoning process, together with a functional program to automatically construct the explanations, and (2) an explanation generation method that explicitly models the relationships between words and visual regions, and simultaneously enhances the interpretability as well as reasoning performance.

3.1. Data

The goal of our proposed data is to offer an explanation benchmark that encodes the reasoning process and grounding across the visual-textual modalities. Compared to previous explanations for visual reasoning [22, 31, 43], it has two key advantages: (1) Grounded on the reasoning process, it elaborates how different components in the visual and textual modalities contribute to the decision making, and reduces variance or inconsistency in textual descriptions; and (2) Instead of modeling textual and visual explanations as separate components, our explanation considers evidence from both modalities in an integral manner, and tightly couples words with image regions (*i.e.,* for visual objects, their grounded regions instead of object names are considered in the explanations). It augments visual reasoning models with the capability to explain their decision making by jointly considering both modalities, resulting in enhanced interpretability and reasoning performance.

Figure 2 illustrates the paradigm for constructing our explanation. To answer the question "Is the plate on the table both dirty and silver?", one needs to locate the table, find the plate on top of it based on their relationship, and investigate the cleanliness as well as the color of the plate. We represent each reasoning step with an atomic operation, *e.g., select* and *verify*, and leverage a functional program to sequentially construct the explanation by traversing the reasoning steps and accumulating important information (*e.g.,* visually grounded objects and their attributes). Upon finishing the traversal, our final explanation not only elaborates the decision making with concrete textual description (*i.e.,* the plate is dirty but not silver thus the answer is no), but also supports the explanation with visual evidence (*i.e.,* grounded regions for the plate and the table).

Decomposing the reasoning process with atomic op-

Operation	Semantic
Select	Selecting a specific category of objects.
Exist	Examining the existence of a specific type of objects.
Filter	Selecting the targeted objects by looking for a specific attribute.
Query	Retrieving the value of a attribute from the selected objects.
Verify	Examining if the targeted objects have a given attribute.
Common	Finding the common attributes among a set of objects.
Same	Examining if two groups of objects have the same attribute.
Different	Examining if two groups of objects have different attributes.
Compare	Comparing the values of an attribute between multiple objects.
Relate	Connecting different objects using their relationships.
And/Or	Logical operations that combine results of previous operations.

Table 1. Atomic operations to represent the reasoning process.

erations. We define a vocabulary of atomic operations by characterizing and abstracting functions for question generation in the GQA dataset [12]. Given the 127 different types of operations in GQA, we first follow [5] and represent each operation as a triplet, *i.e.,* <operation, attribute, category>, and then categorize the original operations in GQA programs based on their semantic meanings. As shown in Table 1, we define 12 atomic operations that cover the essential steps for answering various types of visual questions: some require localizing a specific type of objects (*select*, *exist*); some require reasoning on attributes of the objects (*filter*, *query*, *verify*, *common*, *same*, *different*, *compare*, *relate*); and others require logical reasoning (*and*, *or*).

Traversing reasoning process with a functional program. With the defined atomic operations, we develop a functional program to traverse the reasoning process by performing the corresponding operations and sequentially updating the explanation based on the collected information. Inspired by [12, 13], we represent the reasoning process as a directed graph, where nodes denote the reasoning steps and edges represent their dependencies. As shown in Figure 2, starting from the initial reasoning step (*i.e., Select (table)*), we recursively construct the partial explanation for the current node (shown to the right of each node in Figure 2) and pass it to its dependent nodes. Our final explanation is obtained at the last reasoning step (*i.e., And*). To construct the partial explanation for each node, we design a set of templates based on the semantic meanings of the atomic operations (see our supplementary materials for details). The proposed templates dynamically combine the information extracted within the current node and those transited from its dependent nodes in the previous steps. For example, template for the *relate* operation locates a new object based on its relationship with objects selected in the previous nodes.

The aforementioned paradigm allows efficiently traversing the reasoning process and constructing explanation that elaborates how a decision is made based on the visual and textual modalities. It not only enables the construction of our new GQA-REX dataset with 1,040,830 multi-modal explanations (data statistics and qualitative examples provided in the supplementary materials), but also plays a key role in improving the interpretability and accuracy of visual reasoning models, as detailed in the next subsection.

3.2. Explanation Generation Model

Explaining the rationale behind a decision requires reasoning on visual and textual evidence and elaborating their relationships. Existing explanation generation methods [22, 28, 31, 40] model textual and visual explanations with separate processes, and pay little attention to how key components in each modality correlate with each other. As a result, they have limited capability of generating explanations that jointly consider both modalities and ground words in the images. With the overarching goal of improving the interpretability and accuracy of visual reasoning models, we propose a novel explanation generation model that couples related components across the two modalities and generates the explanation based on their relationships.

Figure 3 illustrates an overview of our method. The principal idea behind the method is to explicitly measure the semantic similarity between words and visual regions, and leverage it to generate multi-modal explanation with enhanced visual grounding. Specifically, unlike conventional methods [22, 28, 31, 40] that generate the explanation solely based on textual features $T_i \in \mathbb{R}^{1\times D}$ (*e.g.,* LSTM hidden state for predicting the i^{th} word), we further measure the similarity between the textual features T_i and visual features $V \in \mathbb{R}^{N\times D}$, and compute the probability of linking the current word with different regions $S_i \in \mathbb{R}^{1\times N}$:

$$S_i^n = \frac{e^{T_i \cdot V_n}}{\sum_{j=1}^{N} e^{T_i \cdot V_j}} \tag{1}$$

where N denotes the total number of image regions, D is the dimension of features, and n is the index for an image region. $T \cdot V$ is the dot product between two features and corresponds to their cosine similarity.

To incorporate visual grounding with explanation generation, we leverage a transformation matrix $M \in \mathbb{R}^{N\times K}$ to map grounding results to the prediction of the next word:

$$y_i^g = S^i \cdot M \tag{2}$$

where K is the number of vocabulary, M is a binary matrix, and $M_{ij} = 1$ if the j^{th} token denotes the i^{th} region (*i.e.,* we use the token $\#i$ to represent grounding a word in the i^{th} region). Since not every word in the explanation can be grounded in the image, *e.g.,* words like "is" do not have an associated region, we further develop a gating function to determine if the current word should be grounded:

$$\hat{g}_i = \sigma(W_g \cdot T_i) \tag{3}$$

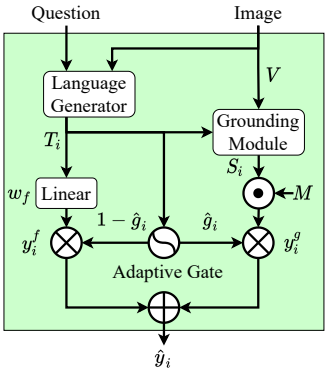

Figure 3. Overview of our explanation generation method.

where $\hat{g}_i$ is the probability of grounding the i^{th} word, $W_g \in \mathbb{R}^{1\times D}$ denotes the trainable weights, and σ is the sigmoid activation function. We use a balanced binary cross-entropy loss to train the gating function:

$$L_g = -\sum_i \frac{C^-}{C} g_i \log \hat{g}_i + \frac{C^+}{C}(1-g_i)\log(1-\hat{g}_i) \quad (4)$$

where g_i is the binary ground truth, C^+ and C^- denote the number of grounded and non-grounded words in the current explanation, and $C = C^+ + C^-$.

Upon obtaining the grounding probability $\hat{g}_i$, we adaptively combine the grounding results y_i^g with the probabilities of different words derived from textual features $y_i^f = softmax(W_f \cdot T_i)$ to determine the next word $\hat{y}_i$:

$$\hat{y}_i = \hat{g}_i y_i^g + (1-\hat{g}_i) y_i^f \quad (5)$$

where $W \in \mathbb{R}^{K\times D}$ represents trainable weights.

We train our model with a linear combination of the balanced binary cross-entropy loss L_g for the gating function and the conventional cross-entropy loss for question answering L_{ans} and explanation generation L_{exp} [22]:

$$L = L_{ans} + L_{exp} + L_g \quad (6)$$

With the aforementioned method that couples key components from both modalities, we significantly improve the model's visual grounding capability, which leads to enhanced interpretability and reasoning performance.

4. Experiments

In this section, we present the implementation details (Section 4.1), and conduct experiments to analyze the proposed framework. We first experiment with the conventional multi-task learning paradigm [22] (Section 4.2). It demonstrates the effectiveness of our explanation in simultaneously enhancing the models' accuracy and interpretability, and highlights the significance of improving visual grounding with our model. We also conduct experiments under the transfer learning paradigm and analyze different visual skills of the reasoning model, aiming to answer the following research questions:

(1) Is the knowledge learned from the explanations transferable to question answering? (Section 4.3)

(2) How do different visual skills affect answer correctness? (Section 4.4)

4.1. Implementations

Dataset. We experiment with our proposed GQA-REX dataset, which is constructed based on the balanced training and validation sets of GQA [12]. We optimize models on the training set and evaluate their performance of explanation generation on the validation set. To evaluate the reasoning performance, we adopt the balanced validation and test-standard sets of GQA. Since the annotated bounding boxes for visual grounding may not align with the visual inputs (*i.e.*, UpDown regional features [2]), we convert the grounding annotations into a set of tokens by finding the input region that has the highest Intersection of Union (IoU) with the ground truth bounding box (*i.e.*, $\#i$ means the bounding box is aligned with the i^{th} input region). We also experiment with the recently introduced GQA-OOD dataset [15] with out-of-distribution data (*i.e.*, "tail" questions).

Evaluation. We evaluate models from multiple perspectives, including reasoning performance, quality of explanations, visual grounding, and recognition of attributes. We use answer accuracy for evaluating reasoning performance. For the quality of explanations, we follow [22, 31, 40] and adopt five language evaluation metrics, including BLEU-4 [30], METEOR [20], ROUGE-L [24], CIDEr [37], and SPICE [1]. Similar to [12], visual grounding (*i.e.*, Grounding) is evaluated by aggregating grounded regions in the predicted explanations and computing their IoU with the ground truth. We evaluate the recognition of eight unique types of attributes, including color, material, sport, shape, pose, size, activity, and relation. We only consider samples where the attributes do not appear in the questions to avoid trivial solutions, and calculate the recall of predicting the correct attributes in the explanations.

Model specification. We use the state-of-the-art VisualBert [21] with UpDown regional features [2] as the backbone for visual reasoning (more details in the supplementary materials). The model is pretrained on MSCOCO [25] dataset without using annotations for question answering. Therefore, it enables us to study the transferability of knowledge learned from our explanations. For explanation generation, we adopt the language generator developed in [40] as our baseline, and incorporate our method proposed in Section 3.2 to enhance its grounding capability.

Training. We use Adam [18] optimizer to train the models with batch size 128. For multi-task learning paradigm

	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Grounding	GQA-val	GQA-test	OOD-val	OOD-test
VisualBert [21]	-	-	-	-	-	-	64.14	56.41	48.70	47.03
VisualBert-VQAE [22]	42.56	34.51	73.59	358.20	40.39	31.29	65.19	57.24	49.20	46.28
VisualBert-EXP [40]	42.45	34.46	73.51	357.10	40.35	33.52	65.17	56.92	49.43	47.69
VisualBert-REX	**54.59**	**39.22**	**78.56**	**464.20**	**46.80**	**67.95**	**66.16**	**57.77**	**50.26**	**48.26**

Table 2. Comparative results on explanation generation and question answering. GQA- and OOD- denote results on GQA and GQA-OOD. Best results are highlighted in bold.

[22], we train the model to simultaneously predict answers and explanations for 8 epochs. The learning rate is initialized as 10^{-4} and decayed once by 0.25 at the last epoch. For the transfer learning paradigm (Section 4.3), we first train the models on explanation generation for 8 epochs and then fine-tune them under the multi-task learning paradigm for 15 epochs. The learning rate is decayed at the 8^{th} and 12^{th} epoch, respectively.

4.2. Results

We first validate the effectiveness of our framework under the multi-task learning paradigm [22]. We compared our model (*i.e.,* VisualBert-REX) with three approaches with the same backbone, including the VQA baseline (*i.e.,* VisualBert [21]), and two explanation generation methods (*i.e.,* VisualBert-VQAE [22] and VisualBert-EXP [40]).

As shown in Table 2, learning with both answers and explanations (*i.e.,* VisualBert-VQAE and VisualBert-EXP) leads to a reasonable improvement over the counterpart using answers alone (*i.e.,* VisualBert), and can further provide illustration of the decision-making process. It shows that our proposed explanations can complement the answer annotations and simultaneously increase the accuracy and interpretability of visual reasoning models. However, it is important to note that, the existing approaches lack the capability of correlating words with their corresponding regions of interest, and thus have low visual grounding scores. Differently, by explicitly modeling the correspondence between key components across the visual and textual modalities, our VisualBert-REX method significantly increases the visual grounding score, leading to further improvements in the quality of the generated explanations and reasoning performance. These observations validate the usefulness of our explanations, and highlight the advantages of our explanation generation method in augmenting visual reasoning models with enhanced visual grounding capability.

In addition to the quantitative evaluation, we also perform qualitative analyses on the predicted answers and explanations. As shown in Figure 4, our VisualBert-REX explains the rationales behind decisions with high-quality explanations, which leads to more accurate answers. Unlike VisualBert-VQAE and VisualBert-EXP that have difficulties localizing the important regions (*e.g.,* horse and car in the 1^{st} example), it accurately captures the key objects with enhanced visual grounding capability and compares their attributes to answer correctly. It also avoids hard-negative objects (*e.g.,* the large carriage on the road in the 2^{nd} sample) by investigating the relationships between objects (*e.g.,* carriage on the grass). Moreover, while conventional methods generate answers without actually reasoning on the visual observations (*e.g.,* without focusing on key objects in the 3^{rd} sample), our method faithfully answers the questions by reasoning on all regions of interest.

4.3. Is the knowledge learned from the explanations transferable to question answering?

Previous methods either simultaneously answer questions and generate explanations [22] or generate explanations for fixed answers [31,39,40], and pay little attention to how transferable is knowledge learned from explanations. Inspired by the recent study [36] that shows knowledge learned from text corpus can enable few-shot visual question answering, in this section, we evaluate the transferability of the proposed reasoning-aware and grounded explanation and analyze its usefulness in distilling knowledge from the reasoning process into question answering.

Specifically, we consider a transfer learning paradigm: first training the models on explanation generation with our complete training set, and then fine-tuning them on both explanation generation and question answering with a subset of 1%, 5%, and 10% of training data. The subsets are created by randomly sampling the specific proportions of questions from each reasoning type, so that the overall statistics about different reasoning tasks are well preserved. We evaluate models on the complete validation set regardless of the amount of training data. To demonstrate the transferability of the knowledge learned from our explanations, we compare the aforementioned method with three alternatives: We first consider (1) a VQA-only and (2) a multi-task learning baseline. They are identical to those discussed in Section 4.2 but trained only on the respective subsets, and thus do not benefit from the knowledge transferred from explanations. (3) To validate that the improvements achieved by transfer learning come from the explanations instead of access to additional questions, we further compare with a self-supervised learning method that pretrains the model on all training questions under the BERT [6] paradigm and then fine-tunes it on the subsets for question answering. Two ob-

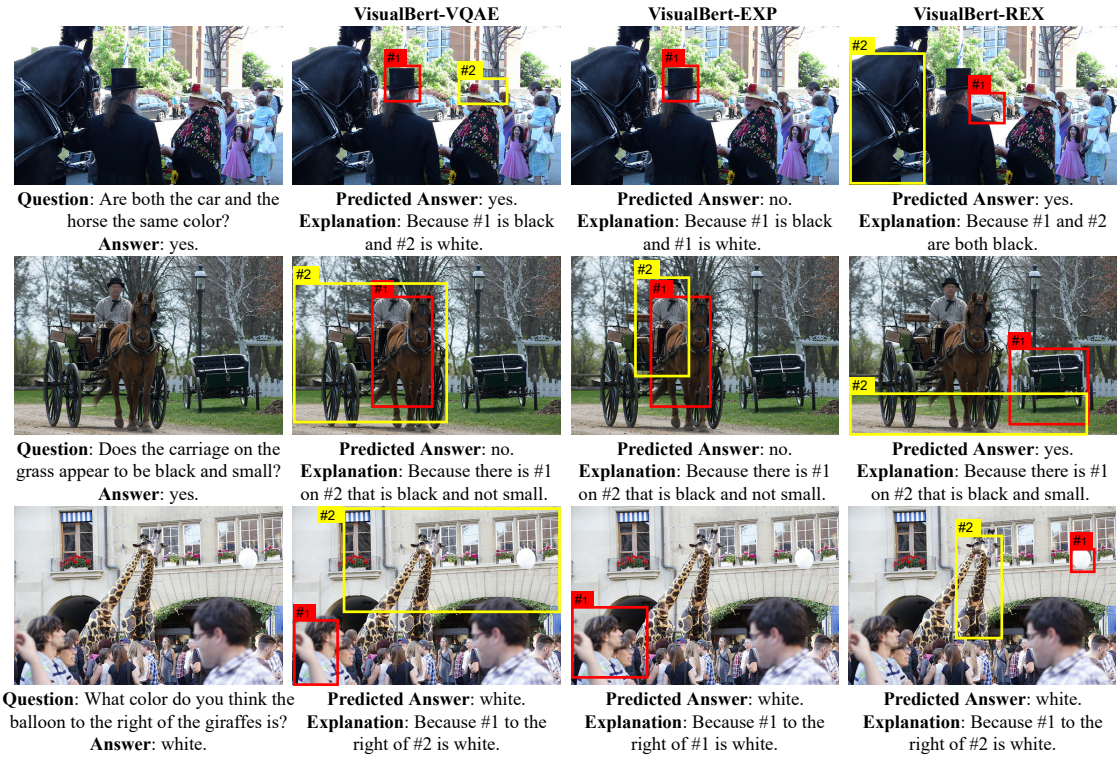

Figure 4. Qualitative results for explaining models' decision-making process. Visual grounding is represented with the token #.

		1%		5%		10%	
		GQA	OOD	GQA	OOD	GQA	OOD
VQA-only	VisualBert	41.41	27.11	48.53	33.78	51.79	37.83
Multi-task learning	VisualBert-EXP	41.70	27.09	49.36	34.50	52.83	38.33
	VisualBert-REX	40.42	23.95	50.30	35.69	53.90	40.08
Self-supervised learning	VisualBert	45.06	30.62	52.12	38.68	54.74	40.12
Transfer learning	VisualBert-EXP	51.32	35.26	56.34	41.20	57.65	43.15
	VisualBert-REX	**57.07**	**40.03**	**61.28**	**45.02**	**61.90**	**45.98**

Table 3. Comparative results for models trained using different proportions of answer annotations. Results are reported on the balanced validation set of GQA and the validation set of GQA-OOD. Best results are highlighted in bold.

servations can be made on the results reported in Table 3:

Knowledge transferred from the explanations plays a key role in question answering. By incorporating our explanations, models trained under the multi-task learning paradigm outperform the VQA-only baseline despite the scarcity of data. Moreover, with more abundant knowledge transferred from the explanations, the transfer learning method improves the performance by a large margin and achieves the best results regardless of the amount of answer annotations. It is notable that VisualBert-REX with only 10% of answers achieves comparable performance to VisualBert trained on the complete training set. On the contrary, while self-supervised learning also increases the reasoning performance, it is not as effective as transfer learning. These observations demonstrate the transferability of knowledge learned from our explanations, and highlight its role in establishing understanding of the reasoning process for more efficient learning on visual reasoning.

Visual grounding is important for transferring knowledge. Compared to VisualBert-EXP, VisualBert-REX with enhanced visual grounding capability achieves much better results under the transfer learning paradigm, demonstrating the advantages of our method under various training paradigms. More importantly, it highlights the significance of visual grounding for developing better understanding of the reasoning process and transferring knowl-

	Recall Rate	Pearson's Correlation
Color	56.01	0.742
Material	49.27	0.708
Sport	72.77	0.575
Shape	40.64	0.548
Pose	74.80	0.417
Size	65.31	0.574
Activity	46.58	0.666
Relation	29.00	0.182

Table 4. Recall rates for capturing key concepts related to different visual skills, and their correlation with the reasoning performance.

edge from explanations to question answering.

4.4. How do different visual skills affect answer correctness?

Answering visual questions involves performing various visual skills [38], *e.g.,* recognition of objects' attributes such as colors and positional relationships. Existing studies assess these skills by categorizing questions into different groups and analyzing them separately [3, 9, 12, 38]. While showing usefulness in studying the models' capability of tackling different types of questions, they fall short of explaining the relationship between successfully performing a skill and correctly answering the question. In this paper, we use a more explicit approach to analyze how various skills affect the answer correctness. Specifically, we evaluate the recognition of eight common attributes based on the model's capability to derive the corresponding concepts in the explanations, *e.g.,* a successful recognition of color requires the model to explain its decision with the key colors, and leverage recall rates of capturing the concepts for quantitative analyses. In Table 4, we report the evaluation scores of different skills and their Pearson's correlation with the predicted probabilities on the correct answers. We make two observations on the results:

Recognition of attributes is important for answering correctly. Our results show that the recall rates for all attributes have reasonable correlation with the reasoning performance, which validates the importance of capturing key attributes in the images for answering visual questions.

Attributes do not contribute equally to answer correctness. It is notable that the model has diverse performance on recognizing different attributes, and the differences in the correlation between skills and answer correctness are also significant (*e.g.*, recognition of colors v.s. recognition of relations). The results indicate that, while it is important for humans to capture different key attributes in order to answer correctly, the attributes do not contribute equally to the decision making of a computational model.

Our results shed light on the underlying decision making of visual reasoning models, and reveal the influences of various visual skills on answer correctness.

5. Discussion

We introduce REX, a principled framework with a new type of reasoning-aware and grounded explanations, a functional program for automatically constructing explanations, and a novel explanation generation method that explicitly couples key components in different modalities. Experimental results demonstrate the usefulness of our framework in explaining the models' decision-making process and improving the visual reasoning performance. They also highlight the critical need to increase models' visual grounding capability for understanding the reasoning process.

Limitations. Despite the aforementioned advantages of our data and model, we believe there is still a large room for interpretable visual reasoning. While the proposed data offers multi-modal explanations derived from a diverse set of images and vocabulary, it may still fail to cover all types of real-world problems. For example, some questions may require external knowledge that is unavailable in the given visual-textual data [29]. One possible direction to address the challenges is to incorporate explanations with vision-and-language pretraining on external knowledge base (*e.g.,* [34]), as our experiments show the effectiveness of transferring knowledge with explanations.

6. Broader Impact

Endowing AI systems the ability to elaborate their decision-making process with high quality multi-modal explanations is an important step toward trustworthy AI. It could fundamentally address the critical needs in opening the black-box of AI algorithms. We therefore envision this work to offer new opportunities to a wide range of domains especially those taking interpretability and transparency as a high priority, such as healthcare, finance, and legislation. The new paradigm highlighting multi-modal understanding, and the large-scale dataset with diverse high-quality explanations may spur innovations and developments in these areas. The explanation generation model elucidates the reasoning process and key components in decision making, alleviating safety or fairness risks of decision-critical applications. We hope that this work would be a useful resource and open a new avenue for the community to develop interpretable and transparent AI systems.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016. 5

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 1, 2, 5

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, pages 2425–2433, 2015. 2, 8

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4290–4300, 2019. 2

[5] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *ECCV*, pages 91–107, 2020. 4

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 6

[7] Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. Beyond vqa: Generating multi-word answers and rationales to visual questions. In *CVPR Workshop*, pages 1623–1632, 2021. 1, 2, 3

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468, 2016. 2

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, pages 6325–6334, 2017. 2, 8

[10] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *ECCV*, pages 55–71, 2018. 2

[11] Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *NeurIPS*, volume 32, 2019. 2

[12] Drew A. Hudson and Christopher D. Manning. GQA: a new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6693–6702, 2019. 2, 4, 5, 8

[13] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 1988–1997, 2017. 2, 4

[14] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015. 2

[15] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Roses are red, violets are blue... but should vqa expect them to? In *CVPR*, pages 2776–2785, 2021. 5

[16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, pages 1571–1581, 2018. 1, 2

[17] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2017. 2

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *IJCV*, volume 123, pages 32–73, 2017. 2

[20] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, 2007. 5

[21] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *ACL*, pages 5265–5275, 2020. 1, 2, 5, 6

[22] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. *ECCV*, pages 570–586, 2018. 1, 2, 3, 4, 5, 6

[23] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 121–137, 2020. 1, 2

[24] Chin-Yew Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, page 74–81, 2004. 5

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 5

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, volume 32, pages 13–23, 2019. 1, 2

[27] Varun Manjunatha, Nirat Saini, and Larry S. Davis. Explicit bias discovery in visual question answering models. In *CVPR*, pages 9554–9563, 2019. 1

[28] Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *EMNLP*, pages 2810–2829, 2020. 1, 2, 3, 4

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3190–3199, 2019. 2, 8

[30] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 5

[31] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *CVPR*, pages 8779–8788, 2018. 1, 2, 3, 4, 5, 6

[32] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *ECCV*, pages 508–524, 2020. 2

[33] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8368–8376, 2019. 2

[34] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, page 4444–4451, 2017. 8

[35] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5100–5111, 2019. 1, 2

[36] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Arxiv*, 2021. 6

[37] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 5

[38] Spencer Whitehead, Hui Wu, Heng Ji, Rogerio Feris, and Kate Saenko. Separating skills and concepts for novel visual question answering. In *CVPR*, pages 5628–5637, 2021. 8

[39] Jialin Wu, Liyan Chen, and Raymond J. Mooney. Improving vqa and its explanations by comparing competing explanations. In *AAAI Workshop*, 2021. 1, 3, 6

[40] Jialin Wu and Raymond Mooney. Faithful multimodal explanation for visual question answering. In *ACL*, pages 103–112, 2019. 1, 2, 3, 4, 5, 6

[41] Jialin Wu and Raymond J. Mooney. Self-critical reasoning for robust visual question answering. In *NeurIPS*, 2019. 3

[42] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, pages 1839–1848, 2017. 2

[43] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6713–6724, 2019. 1, 2, 3