

CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations

Leonard Salewski¹[0000–0001–8531–3011], A. Sophia Koepke¹[0000–0002–5807–0576],
Hendrik P. A. Lensch¹[0000–0003–3616–8668], and
Zeynep Akata^{1,2,3}[0000–0002–1432–7747]

¹ University of Tübingen, Germany

² MPI for Informatics, Saarbrücken, Germany

³ MPI for Intelligent Systems, Tübingen, Germany

{leonard.salewski, a-sophia.koepke, hendrik.lensch,
zeynep.akata}@uni-tuebingen.de

Abstract. Providing explanations in the context of Visual Question Answering (VQA) presents a fundamental problem in machine learning. To obtain detailed insights into the process of generating natural language explanations for VQA, we introduce the large-scale CLEVR-X dataset that extends the CLEVR dataset with natural language explanations. For each image-question pair in the CLEVR dataset, CLEVR-X contains multiple structured textual explanations which are derived from the original scene graphs. By construction, the CLEVR-X explanations are correct and describe the reasoning and visual information that is necessary to answer a given question. We conducted a user study to confirm that the ground-truth explanations in our proposed dataset are indeed complete and relevant. We present baseline results for generating natural language explanations in the context of VQA using two state-of-the-art frameworks on the CLEVR-X dataset. Furthermore, we provide a detailed analysis of the explanation generation quality for different question and answer types. Additionally, we study the influence of using different numbers of ground-truth explanations on the convergence of natural language generation (NLG) metrics. The CLEVR-X dataset is publicly available at <https://github.com/ExplainableML/CLEVR-X>.

Keywords: Visual Question Answering · Natural Language Explanations.

1 Introduction

Explanations for automatic decisions form a crucial step towards increasing transparency and human trust in deep learning systems. In this work, we focus on natural language explanations in the context of vision-language tasks.

In particular, we consider the vision-language task of Visual Question Answering (VQA) which consists of answering a question about an image. This requires multiple skills, such as visual perception, text understanding, and cross-modal reasoning in the visual and language domains. A natural language explanation for a given answer allows a better understanding of the reasoning process for answering the question and adds transparency. However, it is challenging to formulate what comprises a good textual explanation in the context of VQA involving natural images.

VQA-X

Question: Does this scene look like it could be from the early 1950s?



Answer | Explanation:

Yes | The photo is in black and white and the cars are all classic designs from the 1950s

e-SNLI-VE

Hypothesis: A woman is holding a child.

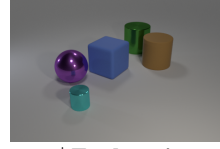


Answer | Explanation:

Entailment | If a woman holds a child she is holding a child.

CLEVR-X

Question: There is a purple metallic ball; what number of cyan objects are right of it?



Answer | Explanation:

1 | There is a cyan cylinder which is on the right side of the purple metallic ball.

Fig. 1: Comparing examples from the VQA-X (left), e-SNLI-VE (middle), and CLEVR-X (right) datasets. The explanation in VQA-X requires prior knowledge (about cars from the 1950s), e-SNLI-VE argues with a tautology, and our CLEVR-X only uses abstract visual reasoning.

Explanation datasets commonly used in the context of VQA, such as the VQA-X dataset [26] or the e-SNLI-VE dataset [12,29] for visual entailment, contain explanations of widely varying quality since they are generated by humans. The ground-truth explanations in VQA-X and e-SNLI-VE can range from statements that merely describe an image to explaining the reasoning about the question and image involving prior information, such as common knowledge. One example for a ground-truth explanation in VQA-X that requires prior knowledge about car designs from the 1950s can be seen in Fig. 1. The e-SNLI-VE dataset contains numerous explanation samples which consist of repeated statements (“x because x”). Since existing explanation datasets for vision-language tasks contain immensely varied explanations, it is challenging to perform a structured analysis of strengths and weaknesses of existing explanation generation methods.

In order to fill this gap, we propose the novel, diagnostic CLEVR-X dataset for visual reasoning with natural language explanations. It extends the synthetic CLEVR [27] dataset through the addition of structured natural language explanations for each question-image pair. An example for our proposed CLEVR-X dataset is shown in Fig. 1. The synthetic nature of the CLEVR-X dataset results in several advantages over datasets that use human explanations. Since the explanations are synthetically constructed from the underlying scene graph, the explanations are *correct* and do not require auxiliary prior knowledge. The synthetic textual explanations do not suffer from errors that get introduced with human explanations. Nevertheless, the explanations in the CLEVR-X dataset are human parsable as demonstrated in the human user study that we conducted. Furthermore, the explanations contain all the information that is necessary to answer a given question about an image without seeing the image. This means that the explanations are *complete* with respect to the question about the image.

The CLEVR-X dataset allows for detailed diagnostics of natural language explanation generation methods in the context of VQA. For instance, it contains a wider

range of question types than other related datasets. We provide baseline performances on the CLEVR-X dataset using recent frameworks for natural language explanations in the context of VQA. Those frameworks are jointly trained to answer the question and provide a textual explanation. Since the question family, question complexity (number of reasoning steps required), and the answer type (binary, counting, attributes) is known for each question and answer, the results can be analyzed and split according to these groups. In particular, the challenging counting problem [48], which is not well-represented in the VQA-X dataset, can be studied in detail on CLEVR-X. Furthermore, our dataset contains multiple ground-truth explanations for each image-question pair. These capture a large portion of the space of correct explanations which allows for a thorough analysis of the influence of the number of ground-truth explanations used on the evaluation metrics. Our approach of constructing textual explanations from a scene graph yields a great resource which could be extended to other datasets that are based on scene graphs, such as the CLEVR-CoGenT dataset.

To summarize, we make the following four contributions: (1) We introduce the CLEVR-X dataset with natural language explanations for Visual Question Answering; (2) We confirm that the CLEVR-X dataset consists of correct explanations that contain sufficient relevant information to answer a posed question by conducting a user study; (3) We provide baseline performances with two state-of-the-art methods that were proposed for generating textual explanations in the context of VQA; (4) We use the CLEVR-X dataset for a detailed analysis of the explanation generation performance for different subsets of the dataset and to better understand the metrics used for evaluation.

2 Related work

In this section, we discuss several themes in the literature that relate to our work, namely *Visual Question Answering*, *Natural language explanations (for vision-language tasks)*, and the *CLEVR dataset*.

Visual Question Answering (VQA). The VQA [4] task has been addressed by several works that apply attention mechanisms to text and image features [56,55,60,45,15]. However, recent works observed that the question-answer bias in common VQA datasets can be exploited in order to answer questions without leveraging any visual information [1,2,27,59]. This has been further investigated in more controlled dataset settings, such as the CLEVR [27], VQA-CP [2], and GQA [25] datasets. In addition to a controlled dataset setting, our proposed CLEVR-X dataset contains natural language explanations that enable a more detailed analysis of the reasoning in the context of VQA.

Natural language explanations. Decisions made by neural networks can be visually explained with visual attribution that is determined by introspecting trained networks and their features [46,57,43,7,58,7], by using input perturbations [42,13,14], or by training a probabilistic feature attribution model along with a task-specific CNN [30]. Complementary to visual explanations methods that tend to not help users distinguish between correct and incorrect predictions [32], natural language explanations have been investigated for a variety of tasks, such as fine-grained visual object classification [21,20], or self-driving car models [31]. The requirement to ground language explanations in the

input image can prevent shortcuts, such as relying on dataset statistics or referring to instance attributes that are not present in the image. For a comprehensive overview of research on explainability and interpretability, we refer to recent surveys [6,9,17].

Natural language explanations for vision-language tasks. Multiple datasets for natural language explanations in the context of vision-language tasks have been proposed, such as the VQA-X [26], VQA-E [35], and e-SNLI-VE datasets [29]. VQA-X [26] augments a small subset of the VQA v2 [18] dataset for the Visual Question Answering task with human explanations. Similarly, the VQA-E dataset [35] extends the VQA v2 dataset by sourcing explanations from image captions. However, the VQA-E explanations resemble image descriptions and do not provide satisfactory justifications whenever prior knowledge is required [35]. The e-SNLI-VE [29,12] dataset combines human explanations from e-SNLI [10] and the image-sentence pairs for the Visual Entailment task from SNLI-VE [54]. In contrast to the VQA-E, VQA-X, and e-SNLI-VE datasets which consist of human explanations or image captions, our proposed dataset contains systematically constructed explanations derived from the associated scene graphs. Recently, several works have aimed at generating natural language explanations for vision-language tasks [26,53,52,38,40,29]. In particular, we use the PJ-X [26] and FM [53] frameworks to obtain baseline results on our proposed CLEVR-X dataset.

The CLEVR dataset. The CLEVR dataset [27] was proposed as a diagnostic dataset to inspect the visual reasoning of VQA models. Multiple frameworks have been proposed to address the CLEVR task [24,41,23,28,47,44]. To add explainability, the XNM model [44] adopts the scene graph as an inductive bias which enables the visualization of the reasoning based on the attention on the nodes of the graph. There have been numerous dataset extensions for the CLEVR dataset, for instance to measure the generalization capabilities of models pre-trained on CLEVR (CLOSURE [51]), to evaluate object detection and segmentation (CLEVR-Ref+ [37]), or to benchmark visual dialog models (CLEVR dialog [34]). The Compositional Reasoning Under Uncertainty (CURI) benchmark uses the CLEVR renderer to construct a test bed for compositional and relational learning under uncertainty [49]. [22] provide an extensive survey of further experimental diagnostic benchmarks for analyzing explainable machine learning frameworks along with proposing the KandinskyPATTERNS benchmark that contains synthetic images with simple 2-dimensional objects. It can be used for testing the quality of explanations and concept learning. Additionally, [5] proposed the CLEVR-XAI-simple and CLEVR-XAI-complex datasets which provide ground-truth segmentation information for heatmap-based visual explanations. Our CLEVR-X augments the existing CLEVR dataset with explanations, but in contrast to (heatmap-based) visual explanations, we focus on natural language explanations.

3 The CLEVR-X dataset

In this section, we introduce the CLEVR-X dataset that consists of natural language explanations in the context of VQA. The CLEVR-X dataset extends the CLEVR dataset with 3.6 million natural language explanations for 850k question-image pairs. In Section 3.1, we briefly describe the CLEVR dataset, which forms the base for our proposed

dataset. Next, we present an overview of the CLEVR-X dataset by describing how the natural language explanations were obtained in Section 3.2, and by providing a comprehensive analysis of the CLEVR-X dataset in Section 3.3. Finally, in Section 3.4, we present results for a user study on the CLEVR-X dataset.

3.1 The CLEVR dataset

The CLEVR dataset consists of images with corresponding full scene graph annotations which contain information about all objects in a given scene (as nodes in the graph) along with spatial relationships for all object pairs. The synthetic images in the CLEVR dataset contain three to ten (at least partially visible) objects in each scene, where each object has the four distinct properties size, color, material, and shape. There are three shapes (box, sphere, cylinder), eight colors (gray, red, blue, green, brown, purple, cyan, yellow), two sizes (large, small), and two materials (rubber, metallic). This allows for up to 96 different combinations of properties.

There are a total of 90 different question families in the dataset which are grouped into 9 different question types. Each type contains questions from between 5 and 28 question families. In the following, we describe the 9 question types in more detail.

Hop questions: The *zero hop*, *one hop*, *two hop*, and *three hop* question types contain up to three relational reasoning steps, e.g. “What color is the cube to the left of the ball?” is a *one hop* question.

Compare and relate questions: The *compare integer*, *same relate*, and *comparison* question types require the understanding and comparison of multiple objects in a scene. Questions of the *compare integer* type compare counts corresponding to two independent clauses (e.g. “Are there more cubes than red balls?”). *Same relate* questions reason about objects that have the same attribute as another previously specified object (e.g. “What is the color of the cube that has the same size as the ball?”). In contrast, *comparison* question types compare the attributes of two objects (e.g. “Is the color of the cube the same as the ball?”).

Single and/or questions: *Single or* questions identify objects that satisfy an exclusive disjunction condition (e.g. “How many objects are either red or blue?”). Similarly, *single and* questions apply multiple relations and filters to find an object that satisfies all conditions (e.g. “How many objects are red and to the left of the cube.”).

Each CLEVR question can be represented by a corresponding functional program and its natural language realization. A functional program is composed of basic functions that resemble elementary visual reasoning operations, such as *filtering* objects by one or more properties, *relating* objects to each other, or *querying* object properties. Furthermore, logical operations like *and* and *or*, as well as counting operations like *count*, *less*, *more*, and *equal* are used to build complex questions. Executing the functional program associated with the question against the scene graph yields the correct answer to the question. We can distinguish between three different answer types: Binary answers (*yes* or *no*), counting answers (integers from 0 to 10), and attribute answers (any of the possible values of shape, color, size, or material).

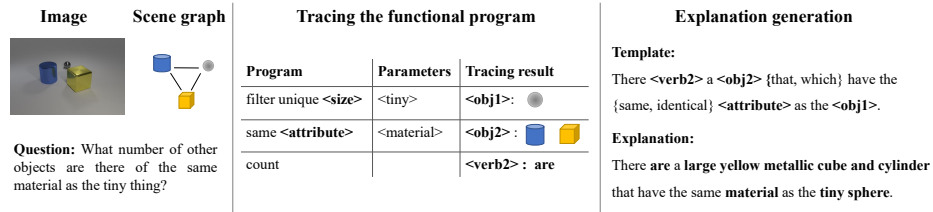


Fig. 2: CLEVR-X dataset generation: Generating a natural language explanation for a sample from the CLEVR dataset. Based on the question, the functional program for answering the question is executed on the scene graph and traced. A language template is used to cast the gathered information into a natural language explanation.

3.2 Dataset generation

Here, we describe the process for generating natural language explanations for the CLEVR-X dataset. In contrast to image captions, the CLEVR-X explanations only describe image elements that are relevant to a specific input question. The explanation generation process for a given question-image pair is illustrated in Fig. 2. It consists of three steps: Tracing the functional program, relevance filtering (not shown in the figure), and explanation generation. In the following, we will describe those steps in detail.

Tracing the functional program. Given a question-image pair from the CLEVR dataset, we trace the execution of the functional program (that corresponds to the question) on the scene graph (which is associated with the image). The generation of the CLEVR dataset uses the same step to obtain a question-answer pair. When executing the basic functions that comprise the functional program, we record their outputs in order to collect all the information required for explaining a ground-truth answer.

In particular, we trace the *filter*, *relate* and *same-property* functions and record the returned objects and their properties, such as *shape*, *size* etc. As a result, the tracing omits objects in the scene that are not relevant for the question. As we are aiming for complete explanations for all question types, each explanation has to mention all the objects that were needed to answer the question, i.e. all the evidence that was obtained during tracing. For example, for *counting* questions, all objects that match the *filter* function preceding the *counting* step are recorded during tracing. For *and* questions, we merge the tracing results of the preceding functions which results in short and readable explanations. In summary, the tracing produces a *complete* and *correct* understanding of the objects and relevant properties which contributed to an answer.

Relevance filtering. To keep the explanation at a reasonable length, we filter the object attributes that are mentioned in the explanation according to their relevance. For example, the *color* of an object is not relevant for a given question that asks about the *material* of said object. We deem all properties that were listed in the question to be relevant. This makes it easier to recognize the same referenced object in both the question and explanation. As the *shape* property also serves as a noun in CLEVR, our explanations always mention the *shape* to avoid using generic shape descriptions like “object” or “thing”. We distinguish between objects which are used to build the question

(e.g. “[...] that is left of the *cube*?”) and those that are the subject of the posed question (e.g. “What color is the *sphere* that is left of the cube?”). For the former, we do not mention any additional properties, and for the latter, we mention the queried property (e.g. `color`) for question types yielding attribute answers.

Explanation generation. To obtain the final natural language explanations, each question type is equipped with one or more natural language templates with variations in terms of the wording used. Each template contains placeholders which are filled with the output of the previous steps, i.e. the tracing of the functional program and subsequent filtering for relevance. As mentioned above, our explanations use the same property descriptions that appeared in the question. This is done to ensure that the wording of the explanation is consistent with the given question, e.g. for the question “Is there a small object?” we generate the explanation “Yes there is a small cube.”⁴. We randomly sample synonyms for describing the properties of objects that do not appear in the question. If multiple objects are mentioned in the explanation, we randomize their order. If the tracing step returned an empty set, e.g. if no object exists that matches the given filtering function for an *existence* or *counting* question, we state that no relevant object is contained in the scene (e.g. “There is no red cube.”).

In order to decrease the overall sentence length and to increase the readability, we aggregate repetitive descriptions (e.g. “There is a red cube and a red cube”) using numerals (e.g. “There are two red cubes.”). In addition, if a function of the functional program merely restricts the output set of a preceding function, we only mention the outputs of the later function. For instance, if a `same-color` function yields a large and a small cube, and a subsequent `filter-large` function restricts the output to only the large cube, we do not mention the output of `same-color`, as the output of the following `filter-large` causes natural language redundancies⁵.

The selection of different language templates, random sampling of synonyms and randomization of the object order (if possible) results in multiple different explanations. We uniformly sample up to 10 different explanations per question for our dataset.

Dataset split. We provide explanations for the CLEVR training and validation sets, skipping only a negligible subset (less than 0.04%) of questions due to malformed question programs from the CLEVR dataset, e.g. due to disjoint parts of their abstract syntax trees. In total, this affected 25 CLEVR training and 4 validation questions.

As the scene graphs and question functional programs are not publicly available for the CLEVR test set, we use the original CLEVR validation subset as the CLEVR-X test set. 20% of the CLEVR training set serve as the CLEVR-X validation set. We perform this split on the image-level to avoid any overlap between images in the CLEVR-X training and validation sets. Furthermore, we verified that the relative proportion of

⁴ The explanation could have used the synonym “box” instead of “cube”. In contrast, “tiny” and “small” are also synonyms in CLEVR, but the explanation would not have been consistent with the question which used “small”.

⁵ E.g. for the question: “How many large objects have the same color as the cube?”, we do not generate the explanation “There are a small and a large cube that have the same color as the red cylinder of which only the large cube is large.” but instead only write “There is a large cube that has the same color as the red cylinder.”

Table 1: Statistics of the CLEVR-X dataset compared to the VQA-X, and e-SNLI-VE datasets. We show the total number of images, questions, and explanations, vocabulary size, and the average number of explanations per question, the average number of words per explanation, and the average number of words per question. Note that subsets do not necessarily add up to the Total since some subsets have overlaps (e.g. for the vocabulary).

Dataset	Subset	Total #				Average #		
		Images	Questions	Explanations	Vocabulary	Explanations	Expl. Words	Quest. Words
VQA-X	Train	24,876	29,549	31,536	9,423	1.07	10.55	7.50
	Val	1,431	1,459	4,377	3,373	3.00	10.88	7.56
	Test	1,921	1,921	5,904	3,703	3.07	10.93	7.31
	Total	28,180	32,886	41,817	10,315	1.48	10.64	7.49
e-SNLI-VE	Train	29,779	401,672	401,672	36,778	1.00	13.62	8.23
	Val	1,000	14,339	14,339	8,311	1.00	14.67	8.10
	Test	998	14,712	14,712	8,334	1.00	14.59	8.20
	Total	31,777	430,723	430,723	38,208	1.00	13.69	8.23
CLEVR-X	Train	56,000	559,969	2,401,275	96	4.29	21.52	21.61
	Val	14,000	139,995	599,711	96	4.28	21.54	21.62
	Test	15,000	149,984	644,151	96	4.29	21.54	21.62
	Total	85,000	849,948	3,645,137	96	4.29	21.53	21.61

samples from each question and answer type in the CLEVR-X training and validation sets is similar, such that there are no biases towards specific question or answer types.

Code for generating the CLEVR-X dataset and the dataset itself are publicly available at <https://github.com/ExplainableML/CLEVR-X>.

3.3 Dataset analysis

We compare the CLEVR-X dataset to the related VQA-X and e-SNLI-VE datasets in Table 1. Similar to CLEVR-X, VQA-X contains natural language explanations for the VQA task. However, different to the natural images and human explanations in VQA-X, CLEVR-X consists of synthetic images and explanations. The e-SNLI-VE dataset provides explanations for the visual entailment (VE) task. VE consists of classifying an input image-hypothesis pair into entailment / neutral / contradiction categories.

The CLEVR-X dataset is significantly larger than the VQA-X and e-SNLI-VE datasets in terms of the number of images, questions, and explanations. In contrast to the two other datasets, CLEVR-X provides (on average) multiple explanations for each question-image pair in the train set. Additionally, the average number of words per explanation is also higher. Since the explanations are built so that they explain each component mentioned in the question, long questions require longer explanations than short questions. Nevertheless, by design, there are no unnecessary redundancies. The explanation length in CLEVR-X is very strongly correlated with the length of the corresponding question (Spearman’s correlation coefficient between the number of words in the explanations and questions is 0.89).

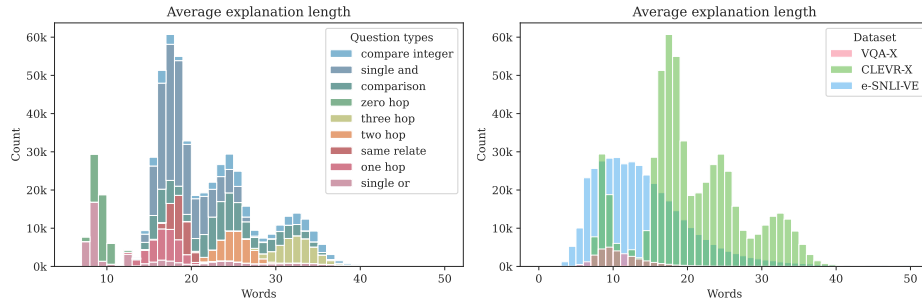


Fig. 3: Stacked histogram of the average explanation lengths measured in words for the nine question types for the CLEVR-X training set (left). Explanation length distribution for the CLEVR-X, VQA-X, and e-SNLI-VE training sets (right). The long tail of the e-SNLI-VE distribution (125 words) was cropped out for better readability.

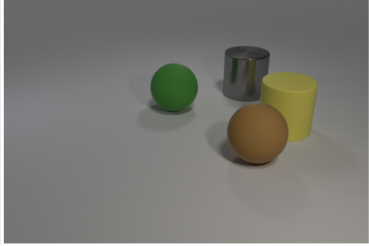
Fig. 3 (left) shows the explanation length distribution in the CLEVR-X dataset for the nine question types. The shortest explanation consists of 7 words, and the longest one has 53 words. On average, the explanations contain 21.53 words. In Fig. 3 (right) and Table 1, we can observe that explanations in CLEVR-X tend to be longer than the explanations in the VQA-X dataset. Furthermore, VQA-X has significantly fewer samples overall than the CLEVR-X dataset. The e-SNLI-VE dataset also contains longer explanations (that are up to 125 words long), but the CLEVR-X dataset is significantly larger than the e-SNLI-VE dataset. However, due to the synthetic nature and limited domain of CLEVR, the vocabulary of CLEVR-X is very small with only 96 different words. Unfortunately, VQA-X and e-SNLI-VE contain spelling errors, resulting in multiple versions of the same words. Models trained on CLEVR-X circumvent those aforementioned challenges and can purely focus on visual reasoning and explanations for the same. Therefore, Natural Language Generation (NLG) metrics applied to CLEVR-X indeed capture the factual correctness and completeness of an explanation.

3.4 User study on explanation completeness and relevance

In this section, we describe our user study for evaluating the completeness and relevance of the generated ground-truth explanations in the CLEVR-X dataset. We wanted to verify whether humans are successfully able to parse the synthetically generated textual explanations and to select complete and relevant explanations. While this is obvious for easier explanations like “There is a blue sphere.”, it is less trivial for more complex explanations such as “There are two red cylinders in front of the green cube that is to the right of the tiny ball.” Thus, strong human performance in the user study indicates that the sentences are parsable by humans.

We performed our user study using Amazon Mechanical Turk (MTurk). It consisted of two types of Human Intelligence Tasks (HITs). Each HIT was made up of (1) An explanation of the task; (2) A non-trivial example, where the correct answers are already selected; (3) A CAPTCHA to verify that the user is human; (4) The problem definition consisting of a question and an image; (5) A user qualification step, for which the user

Image:



Question: Are there any large matte things to the left of the big brown matte sphere?

What is the correct answer for the given question and image?

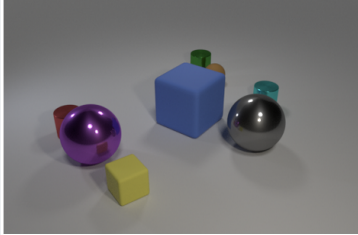
☐ no
 ☐ cylinder
 ☐ sphere
 ☐ yes

Which of these explanations is complete?

☐ Explanation 1
 There is a large matte ball which is on the left side of the big brown matte sphere.

☐ Explanation 2
 There are no large matte things which are to the left of the big brown matte sphere.

Image:



Question: There is a big purple metallic sphere; what number of brown matte spheres are to the left of it?

What is the correct answer for the given question and image?

☐ 0
 ☐ 5
 ☐ 9
 ☐ 6

Which explanation matches the given question and image best?

☐ Explanation 1
 There are no brown matte spheres that are to the left of the big purple metallic sphere.

☐ Explanation 2
 There is a rubber ball which is on the right side of the large green cylinder.

Fig. 4: Two examples from our user study to evaluate the completeness (left) and relevance (right) of natural language explanations in the CLEVR-X dataset.

has to correctly answer a question about an image. This ensures that the user is able to answer the question in the first place, a necessary condition to participate in our user study; (6) Two explanations from which the user needs to choose one. Example screenshots of the user interface for the user study are shown in Fig. 4.

For the two different HIT types, we randomly sampled 100 explanations from each of the 9 question types, resulting in a total of 1800 samples for the completeness and relevance tasks. For each task sample, we requested 3 different MTurk workers based in the US (with high acceptance rate of $> 95\%$ and over 5000 accepted HITs). A total of 78 workers participated in the completeness HITs. They took on average 144.83 seconds per HIT. The relevance task was carried out by 101 workers which took on average 120.46 seconds per HIT. In total, 134 people participated in our user study. In the following, we describe our findings regarding the completeness and relevance of the CLEVR-X explanations in more detail.

Explanation completeness. In the first part of the user study, we evaluated whether human users are able to determine if the ground-truth explanations in the CLEVR-X dataset are complete (and also correct). We presented the MTurk workers with an image, a question, and two explanations. As can be seen in Fig. 4 (left), a user had to first select the correct answer (yes) before deciding which of the two given explanations was complete. By design, one of the explanations presented to the user was the complete one from the CLEVR-X dataset and the other one was a modified version for which at least one necessary object had been removed. As simply deleting an object from

Table 2: Results for the user study evaluating the accuracy for the completeness and relevance tasks for the nine question types in the CLEVR-X dataset.

	Zero hop	One hop	Two hop	Three hop	Same relate	Compari- son	Compare integer	Single or	Single and	All
Completeness	100.00	98.00	98.67	94.00	100.00	83.67	77.00	84.00	94.33	92.19
Relevance	99.67	99.00	95.67	89.00	95.67	87.33	83.67	90.67	92.00	92.52

a textual explanation could lead to grammar errors, we re-generated the explanations after removing objects from the tracing results. This resulted in incomplete, albeit grammatically correct, explanations.

To evaluate the ability to determine the completeness of explanations, we measured the accuracy of selecting the complete explanation. The human participants obtained an average accuracy of 92.19%, confirming that complete explanations which mention all objects necessary to answer a given question were preferred over incomplete ones. The performance was weaker for complex question types, such as *compare-integer* and *comparison* with accuracies of only 77.00% and 83.67% respectively, compared to the easier *zero-hop* and *one-hop* questions with accuracies of 100% and 98.00% respectively.

Additionally, there were huge variations in performance across different participants of the completeness study (Fig. 5 (top left)), with the majority performing very well (>97% answering accuracy) for most question types. For the *compare-integer*, *comparison* and *single or* question types, some workers exhibited a much weaker performance with answering accuracies as low as 0%. The average turnaround time shown in Fig. 5 (bottom left) confirms that complex question types required less time to be solved than more complex question types, such as *three hop* and *compare integer* questions. Similar to the performance, the work time varied greatly between different users.

Explanation relevance. In the second part of our user study, we analyzed if humans are able to identify explanations which are relevant for a given image. For a given question-image pair, the users had to first select the correct answer. Furthermore, they were provided with a correct explanation and another randomly chosen explanation from the same question family (that did not match the image). The task consisted of selecting the correct explanation that matched the image and question content. Explanation 1 in the example user interface shown in Fig. 4 (right) was the relevant one, since Explanation 2 does not match the question and image.

The participants of our user study were able to determine which explanation matched the given question-image example with an average accuracy of 92.52%. Again, the performance for complex question types was weaker than for easier questions. The difficulty of the question influences the accuracy of detecting the relevant explanation, since this task first requires understanding the question. Furthermore, complex questions tend to be correlated with complex scenes that contain many objects which makes the user’s task more challenging. The accuracy for *three-hop* questions was 89.00% compared to 99.67% for *zero-hop* questions. For *compare-integer* and *comparison* questions, the users obtained accuracies of 83.67% and 87.33% respectively, which is significantly lower than the overall average accuracy.

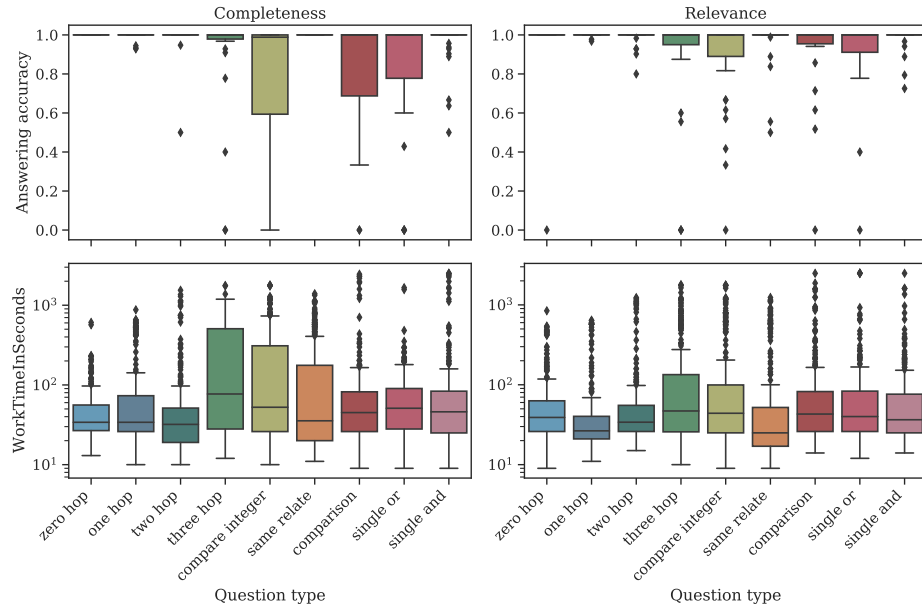


Fig. 5: Average answering accuracies for each worker (top) and average work time (bottom) for the user study (left: completeness, right: relevance). The boxes indicate the mean as well as lower and upper quartiles, the lines extend 1.5 interquartile ranges of the lower and upper quartile. All other values are plotted as diamonds.

We analyzed the answering accuracy per worker in Fig. 5 (top). The performance varies greatly between workers, with the majority performing very well ($>90\%$ answering accuracy) for most question types. Some workers showed much weaker performance with answering accuracies as low as 0% (e.g. for *compare-integer* and *single or* questions). Furthermore, the distribution of work time for the relevance task is shown in Fig. 5 (bottom right). The turnaround times for each worker exhibit greater variation on the completeness task (bottom left) compared to the relevance task (bottom right). This might be due to the nature of the different tasks. For the completeness task, the users need to check if the explanation contains all the elements that are necessary to answer the given question. The relevance task, on the other hand, can be solved by detecting a single non-relevant object to discard the wrong explanation.

Our user study confirmed that humans are able to parse the synthetically generated natural language explanations in the CLEVR-X dataset. Furthermore, the results have shown that users prefer complete and relevant explanations in our dataset over corrupted samples.

4 Experiments

We describe the experimental setup for establishing baselines on our proposed CLEVR-X dataset in Section 4.1. In Section 4.2, we present quantitative results on the CLEVR-X

dataset. Additionally, we analyze the generated explanations for the CLEVR-X dataset in relation to the question and answer types in Section 4.3. Furthermore, we study the behavior of the NLG metrics when using different numbers of ground-truth explanations for testing in Section 4.4. Finally, we present qualitative explanation generation results on the CLEVR-X dataset in Section 4.5.

4.1 Experimental setup

In this section, we provide details about the datasets and models used to establish baselines for our CLEVR-X dataset and about their training details. Furthermore, we explain the metrics for evaluating the explanation generation performance.

Datasets. In the following, we summarize the datasets that were used for our experiments. In addition to providing baseline results on CLEVR-X, we also report experimental results on the VQA-X and e-SNLI-VE datasets. Details about our proposed **CLEVR-X** dataset can be found in Section 3. The **VQA-X** dataset [26] is a subset of the VQA v2 dataset with a single human-generated textual explanation per question-image pair in the training set and 3 explanations for each sample in the validation and test sets. The **e-SNLI-VE** dataset [12,29] is a large-scale dataset with natural language explanations for the visual entailment task.

Methods. We used multiple frameworks to provide baselines on our proposed CLEVR-X dataset. For the **random words** baseline, we sample random word sequences of length w for the answer and explanation words for each test sample. The full vocabulary corresponding to a given dataset is used as the sampling pool, and w denotes the average number of words forming an answer and explanation in a given dataset. For the **random explanations** baseline, we randomly sample an answer-explanation pair from the training set and use this as the prediction. The explanations from this baseline are well-formed sentences. However, the answers and explanations most likely do not match the question or the image. For the random-words and random-explanations baselines, we report the NLG metrics for all samples in the test set (instead of only considering the correctly answered samples, since the random sampling of the answer does not influence the explanation). The Pointing and Justification model **PJ-X** [26] provides text-based post-hoc justifications for the VQA task. It combines a modified MCB [16] framework, pre-trained on the VQA v2 dataset, with a visual pointing and textual justification module. The Faithful Multimodal (**FM**) model [53] aims at grounding parts of generated explanations in the input image to provide explanations that are *faithful* to the input image. It is based on the Up-Down VQA model [3]. In addition, FM contains an explanation module which enforces consistency between the predicted answer, explanation and the attention of the VQA model. The implementations for the PJ-X and FM models are based on those provided by the authors of [29].

Implementation and training details. We extracted $14 \times 14 \times 1024$ grid features for the images in the CLEVR-X dataset using a ResNet-101 [19], pre-trained on ImageNet [11]. These grid features served as inputs to the FM [53] and PJ-X [26] frameworks. The CLEVR-X explanations are lower case and punctuation is removed from the sentences. We selected the best model on the CLEVR-X validation set based on the highest mean of the four NLG metrics, where explanations for incorrect answers were set to an empty

string. This metric accounts for the answering performance as well as for the explanation quality. The final models were evaluated on the CLEVR-X test set. For PJ-X, our best model was trained for 52 epochs, using the Adam optimizer [33] with a learning rate of 0.0002 and a batch size of 256. We did not use gradient clipping for PJ-X. Our strongest FM model was trained for 30 epochs, using the Adam optimizer with a learning rate of 0.0002, a batch size of 128, and gradient clipping of 0.1. All other hyperparameters were taken from [26,53].

Evaluation metrics. To evaluate the quality of the generated explanations, we use the standard natural language generation metrics BLEU [39], METEOR [8], ROUGE-L [36] and CIDEr [50]. By design, there is no correct explanation that can justify a wrong answer. We follow [29] and report the quality of the generated explanations for the subset of correctly answered questions.

4.2 Evaluating explanations generated by state-of-the-art methods

In this section, we present quantitative results for generating explanations for the CLEVR-X dataset (Table 3). The random words baseline exhibits weak explanation performance for all NLG metrics on CLEVR-X. Additionally, the random answering accuracy is very low at 3.6%. The results are similar on VQA-X and e-SNLI-VE. The random explanations baseline achieves stronger explanation results on all three datasets, but is still significantly worse than the trained models. This confirms that, even with a medium-sized answer space (28 options) and a small vocabulary (96 words), it is not possible to achieve good scores on our dataset using a trivial approach.

We observed that the PJ-X model yields a significantly stronger performance on CLEVR-X in terms of the NLG metrics for the generated explanations compared to the FM model, with METEOR scores of 58.9 and 52.5 for PJ-X and FM respectively. Across all explanation metrics, the scores on the VQA-X and e-SNLI-VE datasets are in a lower range than those on CLEVR-X. For PJ-X, we obtain a CIDEr score of 639.8 on CLEVR-X and 82.7 and 72.5 on VQA-X and e-SNLI-VE. This can be attributed to the smaller vocabulary and longer sentences, which allow n -gram based metrics (e.g. BLEU) to match parts of sentences more easily.

In contrast to the explanation generation performance, the FM model is better at answering questions than PJ-X on CLEVR-X with an answering accuracy of 80.3% for FM compared to 63.0% for PJ-X. Compared to recent models tuned to the CLEVR task, the answering performances of PJ-X and FM do not seem very strong. However, the PJ-X backbone MCB [15] (which is crucial for the answering performance) preceded the publication of the CLEVR dataset. A version of the MCB backbone (CNN+LSTM+MCB in the CLEVR publication [27]) achieved an answering accuracy of 51.4% on CLEVR [27], whereas PJ-X is able to correctly answer 63% of the questions. The strongest model discussed in the initial CLEVR publication (CNN+LSTM+SA in [27]) achieved an answering accuracy of 68.5%.

4.3 Analyzing results on CLEVR-X by question and answer types

In Fig. 6 (left and middle), we present the performance for PJ-X on CLEVR-X for the nine question and three answer types. The explanation results for samples which

Table 3: Explanation generation results on the CLEVR-X, VQA-X, and e-SNLI-VE test sets using BLEU-4 (B4), METEOR (M), ROUGE-L (RL), CIDEr (C), and answer accuracy (Acc). Higher is better for all reported metrics. For the random baselines, Acc corresponds to $100/\# \text{ answers}$ for CLEVR-X and e-SNLI-VE, and to the VQA answer score for VQA-X. (Rnd. words: random words, Rnd. expl: Random explanations)

Model	CLEVR-X					VQA-X					e-SNLI-VE				
	B4	M	RL	C	Acc	B4	M	RL	C	Acc	B4	M	RL	C	Acc
Rnd. words	0.0	8.4	11.4	5.9	3.6	0.0	1.2	0.7	0.1	0.1	0.0	0.3	0.0	0.0	33.3
Rnd. expl	10.9	16.6	35.3	30.4	3.6	0.9	6.5	18.4	21.6	0.2	0.4	5.4	9.9	2.6	33.3
FM [53]	78.8	52.5	85.8	566.8	80.3	23.1	20.4	47.1	87.0	75.5	8.2	15.6	29.9	83.6	58.5
PJ-X [26]	87.4	58.9	93.4	639.8	63.0	22.7	19.7	46.0	82.7	76.4	7.3	14.7	28.6	72.5	69.2

require counting abilities (counting answers) are lower than those for attribute answers (57.3 vs. 63.3). This is in line with prior findings that VQA models struggle with counting problems [48]. The explanation quality for binary questions is even lower with a METEOR score of only 55.6. The generated explanations are of higher quality for easier question types; *zero-hop* questions yield a METEOR score of 64.9 compared to 62.1 for *three-hop* questions. It can also be seen that *single-or* questions are harder to explain than *single-and* questions. These trends can be observed across all NLG explanation metrics.

4.4 Influence of using different numbers of ground-truth explanations

In this section, we study the influence of using multiple ground-truth explanations for evaluation on the behavior of the NLG metrics. This gives insights about whether the metrics can correctly rate a model’s performance with a limited number of ground-truth explanations. We set an upper bound k on the number of explanations used and randomly sample k explanations if a test sample has more than k explanations for $k \in \{1, 2, \dots, 10\}$. Fig. 6 (right) shows the NLG metrics (normalized with the maximum value for each metric on the test set for all ground-truth explanations) for the PJ-X model depending on the average number of ground-truth references used on the test set.

Out of the four metrics, BLEU-4 converges the slowest, requiring close to 3 ground-truth explanations to obtain a relative metric value of 95%. Hence, BLEU-4 might not be able to reliably predict the explanation quality on the e-SNLI-VE dataset which has only one explanation for each test sample. CIDEr converges faster than ROUGE and METEOR, and achieves 95.7% of its final value with only one ground-truth explanation. This could be caused by the fact, that CIDEr utilizes a tf-idf weighting scheme for different words, which is built from all reference sentences in the subset that the metric is computed on. This allows CIDEr to be more sensitive to important words (e.g. attributes and shapes) and to give less weight, for instance, to stopwords, such as “the”. The VQA-X and e-SNLI-VE datasets contain much lower average numbers of explanations for each dataset sample (1.4 and 1.0). Since there could be many more possible explanations

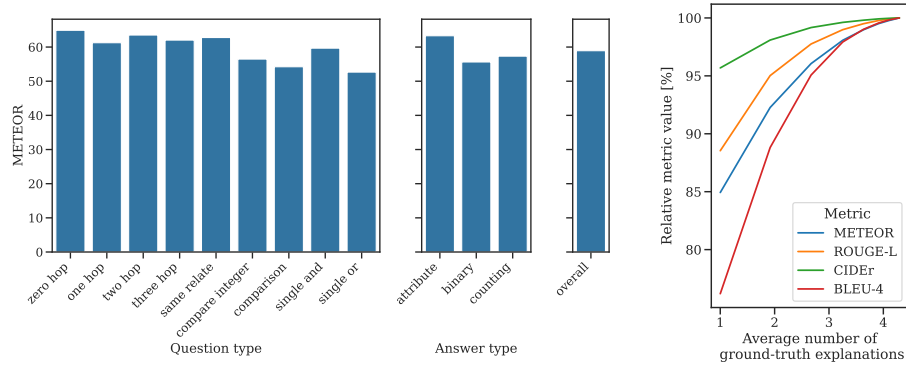


Fig. 6: Explanation generation results for PJ-X on the CLEVR-X test set according to question (left) and answer (middle) types compared to the overall explanation quality. Easier types yield higher METEOR scores. NLG metrics using different numbers of ground-truth explanations on the CLEVR-X test set (right). CIDEr converges faster than the other NLG metrics.

for samples in those datasets that describe different aspects than those mentioned in the ground truth, automated metric may not be able to correctly judge a prediction even if it is correct and faithful w.r.t. to the image and question.

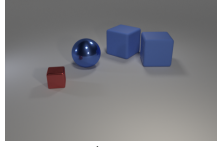
4.5 Qualitative explanation generation results

We show examples for explanations generated with the PJ-X framework on CLEVR-X in Fig. 7. As can be seen across the three examples presented, PJ-X generates high-quality explanations which closely match the ground-truth explanations.

In the left-most example in Fig. 7, we can observe slight variations in grammar when comparing the generated explanation to the ground-truth explanation. However, the content of the generated explanation corresponds to the ground truth. Furthermore, some predicted explanations differ from the ground-truth explanation in the use of another synonym for a predicted attribute. For instance, in the middle example in Fig. 7, the ground-truth explanation describes the size of the cylinder as “small”, whereas the predicted explanation uses the equivalent attribute “tiny”. In contrast to other datasets, the set of ground-truth explanations for each sample in CLEVR-X contains these variations. Therefore, the automated NLG metrics do not decrease when such variations are found in the predictions. For the first and second example, PJ-X obtains the highest possible explanation score (100.0) in terms of the BLEU-4, METEOR, and ROUGE-L metrics.

We show a failure case where PJ-X predicted the wrong answer in Fig. 7 (right). The generated answer-explanation pair shows that the predicted explanation is consistent with the wrong answer prediction and does not match the input question-image pair. The NLG metrics for this case are significantly weaker with a BLEU-4 score of 0.0, as there are no matching 4-grams between the prediction and the ground truth.

Question: How many tiny red things are the same material as the big sphere?

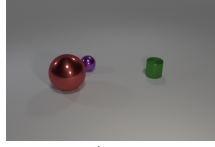


GT Answer | Explanation:
1 | The tiny red metal block has the same material as a big sphere.

Pred. Answer | Expl.
1 | There is the tiny red metal block which has the identical material as a big sphere.

B4 / M / RL / C:
100.0 / 100.0 / 100.0 / 744.0

Question: The cylinder has what size?

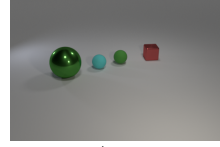


GT Answer | Explanation:
Small | The cylinder is small.

Pred. Answer | Expl.
Small | The cylinder is tiny.

B4 / M / RL / C:
100.0 / 100.0 / 100.0 / 462.4

Question: Are there any small matte cubes?



GT Answer | Explanation:
No | There are no small matte cubes.

Pred. Answer | Expl.
Yes | There is a small matte cube.

B4 / M / RL / C:
0.0 / 76.9 / 57.1 / 157.1

Fig. 7: Examples for answers and explanations generated with the PJ-X framework on the CLEVR-X dataset, showing correct answer predictions (left, middle) and a failure case (right). The NLG metrics obtained with the explanations for the correctly predicted answers are high compared to those for the explanation corresponding to the wrong answer prediction.

5 Conclusion

We introduced the novel CLEVR-X dataset which contains natural language explanations for the VQA task on the CLEVR dataset. Our user study confirms that the explanations in the CLEVR-X dataset are complete and match the questions and images. Furthermore, we have provided baseline performances using the PJ-X and FM frameworks on the CLEVR-X dataset. The structured nature of our proposed dataset allowed the detailed evaluation of the explanation generation quality according to answer and question types. We observed that the generated explanations were of higher quality for easier answer and question categories. One of our findings is, that explanations for counting problems are worse than for other answer types, suggesting that further research into this direction is needed. Additionally, we find that the four NLG metrics used to evaluate the quality of the generated explanations exhibit different convergence patterns depending on the number of available ground-truth references.

Since this work only considered two natural language generation methods for VQA as baselines, the natural next step will be the benchmarking and closer investigation of additional recent frameworks for textual explanations in the context of VQA on the CLEVR-X dataset. We hope that our proposed CLEVR-X benchmark will facilitate further research to improve the generation of natural language explanations in the context of vision-language tasks.

6 Acknowledgements

The authors thank the Amazon Mechanical Turk workers that participated in the user study. This work was supported by the DFG – EXC number 2064/1 – project number 390727645, by the DFG: SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms - project number: 276693517, by the ERC (853489 - DEXIM), and by the BMBF (FKZ: 01IS18039A). L. Salewski thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support.

References

1. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: EMNLP. pp. 1955–1960. Association for Computational Linguistics (2016) [3](#)
2. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don’t just assume; look and answer: Overcoming priors for visual question answering. In: CVPR. pp. 4971–4980 (2018) [3](#)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR. pp. 6077–6086 (2018) [13](#)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015) [3](#)
5. Arras, L., Osman, A., Samek, W.: Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion* **81**, 14–40 (2022) [4](#)
6. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion* **58**, 82–115 (2020) [4](#)
7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **10**(7), e0130140 (2015) [3](#)
8. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop. pp. 65–72 (2005) [14](#)
9. Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al.: Toward trustworthy ai development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213 (2020) [4](#)
10. Camburu, O.M., Rocktäschel, T., Lukasiewicz, T., Blunsom, P.: e-snli: Natural language inference with natural language explanations. In: NeurIPS (2018) [4](#)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) [13](#)
12. Do, V., Camburu, O.M., Akata, Z., Lukasiewicz, T.: e-snli-ve-2.0: Corrected visual-textual entailment with natural language explanations. arXiv preprint arXiv:2004.03744 (2020) [2](#), [4](#), [13](#)
13. Fong, R.C., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV. pp. 2950–2958 (2019) [3](#)
14. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV. pp. 3429–3437 (2017) [3](#)
15. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP. pp. 457–468 (2016) [3](#), [14](#)

16. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. In: EMNLP. pp. 457–468 (2016) 13
17. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: IEEE DSAA. pp. 80–89 (2018) 4
18. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In: CVPR. pp. 6904–6913 (2017) 4
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 13
20. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Generating counterfactual explanations with natural language. arXiv preprint arXiv:1806.09809 (2018) 3
21. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Grounding visual explanations. In: ECCV. pp. 264–279 (2018) 3
22. Holzinger, A., Saranti, A., Mueller, H.: Kandinskypatterns - an experimental exploration environment for pattern analysis and machine intelligence. arXiv preprint arXiv:2103.00519 (2021) 4
23. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. In: NeurIPS (2019) 4
24. Hudson, D.A., Manning, C.D.: Compositional attention networks for machine reasoning. In: ICLR (2018) 4
25. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR. pp. 6693–6702 (2019) 3
26. Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: CVPR. pp. 8779–8788 (2018) 2, 4, 13, 14, 15
27. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR. pp. 2901–2910 (2017) 2, 3, 4, 14
28. Johnson, J., Hariharan, B., Van Der Maaten, L., Hoffman, J., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Inferring and executing programs for visual reasoning. In: ICCV. pp. 2989–2998 (2017) 4
29. Kayser, M., Camburu, O.M., Salewski, L., Emde, C., Do, V., Akata, Z., Lukasiewicz, T.: e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In: ICCV. pp. 1244–1254 (2021) 2, 4, 13, 14
30. Kim, J.M., Choe, J., Akata, Z., Oh, S.J.: Keep calm and improve visual feature attribution. In: ICCV. pp. 8350–8360 (2021) 3
31. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: ECCV. pp. 563–578 (2018) 3
32. Kim, S.S., Meister, N., Ramaswamy, V.V., Fong, R., Russakovsky, O.: Hive: Evaluating the human interpretability of visual explanations. arXiv preprint arXiv:2112.03184 (2021) 3
33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015) 14
34. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In: NAACL. pp. 582–595 (2019) 4
35. Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In: ECCV. pp. 552–567 (2018) 4
36. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: ACL. pp. 74–81 (2004) 14
37. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: Clevr-ref+: Diagnosing visual reasoning with referring expressions. In: CVPR. pp. 4185–4194 (2019) 4

38. Marasović, A., Bhagavatula, C., Park, J.s., Le Bras, R., Smith, N.A., Choi, Y.: Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In: EMNLP. pp. 2810–2829 (2020) 4
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002) 14
40. Patro, B., Patel, S., Namboodiri, V.: Robust explanations for visual question answering. In: WACV. pp. 1577–1586 (2020) 4
41. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer. In: AAAI. vol. 32 (2018) 4
42. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMVC. p. 151 (2018) 3
43. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV. pp. 618–626 (2017) 3
44. Shi, J., Zhang, H., Li, J.: Explainable and Explicit Visual Reasoning Over Scene Graphs. In: CVPR. pp. 8376–8384 (2019) 4
45. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: CVPR. pp. 4613–4621 (2016) 3
46. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR Workshop (2014) 3
47. Suarez, J., Johnson, J., Li, F.F.: Ddrprog: A clevr differentiable dynamic reasoning programmer. arXiv preprint arXiv:1803.11361 (2018) 4
48. Trott, A., Xiong, C., Socher, R.: Interpretable counting for visual question answering. In: ICLR (2018) 3, 15
49. Vedantam, R., Szlam, A., Nickel, M., Morcos, A., Lake, B.M.: CURI: A benchmark for productive concept learning under uncertainty. In: ICML. pp. 10519–10529 (2021) 4
50. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR. pp. 4566–4575 (2015) 14
51. de Vries, H., Bahdanau, D., Murty, S., Courville, A.C., Beaudoin, P.: CLOSURE: assessing systematic generalization of CLEVR models. In: NeurIPS Workshop (2019) 4
52. Wu, J., Chen, L., Mooney, R.: Improving vqa and its explanations by comparing competing explanations. In: AAAI Workshop (2021) 4
53. Wu, J., Mooney, R.: Faithful Multimodal Explanation for Visual Question Answering. In: ACL Workshop. pp. 103–112 (2019) 4, 13, 14, 15
54. Xie, N., Lai, F., Doran, D., Kadav, A.: Visual entailment: A novel task for fine-grained image understanding. arXiv preprint arXiv:1901.06706 (2019) 4
55. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV. pp. 451–466 (2016) 3
56. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR. pp. 21–29 (2016) 3
57. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833 (2014) 3
58. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. *International Journal of Computer Vision* **126**(10), 1084–1102 (2018) 3
59. Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D.: Yin and yang: Balancing and answering binary visual questions. In: CVPR. pp. 5014–5022 (2016) 3
60. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: Grounded question answering in images. In: CVPR. pp. 4995–5004 (2016) 3