Few-Shot Multimodal Explanation for Visual Question Answering

Anonymous Authors

ABSTRACT

A key object in eXplainable Artificial Intelligence (XAI) is to create intelligent systems capable of reasoning and explaining real-world data to facilitate reliable decision-making. Recent studies have acknowledged the importance of providing user-friendly and verifiable explanations to facilitate trustworthy Visual Question Answering (VQA) systems. This paper aims to promote explainable VQA from both data and method perspectives. First, we propose a new Standard Multimodal Explanation (SME) dataset and a new Few-Shot Multimodal Explanation for VQA (FS-MEVQA) task, which aims to generate the multimodal explanation of the underlying reasoning process for solving visual questions with few training samples. Our SME dataset includes 1,028,230 samples composed of questions, images, answers, and multimodal explanations, which can facilitate the research in both traditional MEVQA and FS-MEVQA. To the best of our knowledge, this is the first large-scale dataset with joint language-vision explanations based on standard English and additional visual grounding tokens, which bridge MEVQA to a broad field in Natural Language Processing (NLP). Second, we propose a training-free Multimodal Explaining Agent (MEAgent) method based on an LLM agent with multimodal open-world tools to infer answers and generate multimodal explanations for visual questions. Our MEAgent can learn multimodal explaining from merely $N(= 16)$ training samples and leverage open-world abilities to perform FS-MEVQA on test samples. Comprehensive experimental results evaluated by language quality metrics, visual detection metric, and visual attribution metrics on our SME dataset indicate the superiority of our method for FS-MEVQA, compared to state-of-the-art MEVQA methods and the multimodal LLM GPT-4V. Our code and data are available at https://anonymous.4open.science/r/FS-MEVQA-646D/.

CCS CONCEPTS

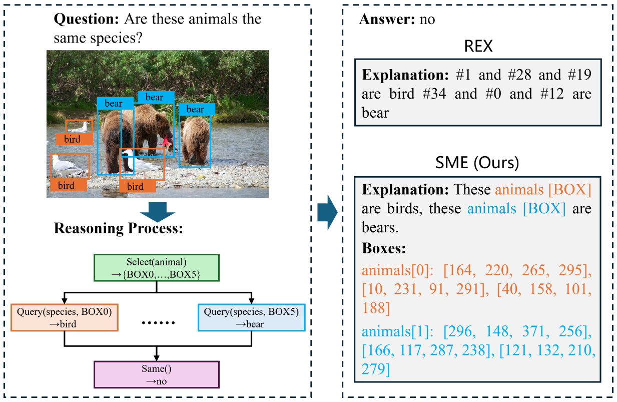
• **Computing methodologies → Knowledge representation and reasoning**; *Computer vision*; *Natural language processing*.

KEYWORDS

Few-shot learning, multimodal generation, explainable artificial intelligence, visual question answering

ACM Reference Format:

Anonymous Authors. 2024. Few-Shot Multimodal Explanation for Visual Question Answering. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn



Question: Are these animals the same species?

bird
bear
bear
bear
bird
bird
bird
Reasoning Process:
Select(animal) →{BOX0,…,BOX5}
Query(species, BOX0) →bird
……
Query(species, BOX5) →bear
Same() →no

Answer: no

REX

Explanation: #1 and #28 and #19 are bird #34 and #0 and #12 are bear

SME (Ours)

Explanation: These animals [BOX] are birds, these animals [BOX] are bears.

Boxes:

animals[0]: [164, 220, 265, 295], [10, 231, 91, 291], [40, 158, 101, 188]

animals[1]: [296, 148, 371, 256], [166, 117, 287, 238], [121, 132, 210, 279]

Figure 1: An example of multimodal explanations constructed in the REX dataset [7] and our SME dataset. In REX, $\#i$ denotes boxes pre-extracted by Faster R-CNN, belonging to only 81 classes. We use $[BOX]$ to represent the detection boxes needed in the explanation and annotate these boxes with their corresponding names mentioned in the explanation, based on the scene graph annotated by humans. Our explanation is in standard English with additional $[BOX]$.

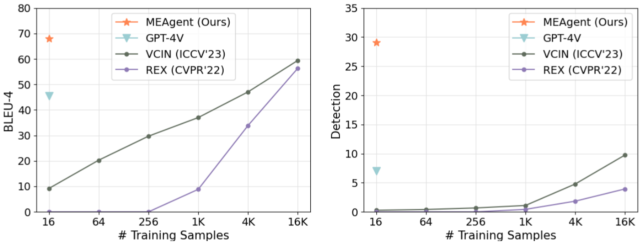
1 INTRODUCTION

Reasoning is a fundamental element of human intelligence and a crucial challenge in artificial intelligence [9, 10, 18]. Traditional reasoning models typically provide answers without offering explanations for their reasoning. This limitation greatly restricts their applicability, particularly in safety-sensitive situations, such as healthcare, transportation, and finance. To address this problem, a key object in eXplainable Artificial Intelligence (XAI) is to create intelligent systems capable of reasoning and explaining real-world data to facilitate reliable decision-making [11, 37, 42]. The recent development of Large Language Models (LLMs) has led to remarkable progress in reasoning on textual data and generating textual explanations for reasoning processes [12, 26]. However, when it comes to the field of multimodal reasoning, tasks necessitate the ability to comprehend multimodal content and generate multimodal explanations to reveal the underlying reasoning processes, which is still a challenging problem.

While Visual Question Answering (VQA) [3, 4, 13, 15, 32] is a typical and important multimodal reasoning task, some pioneering work has been made in Multimodal Explanation for VQA (MEVQA) [7, 50, 51]. Zellers [51] construct the VCR dataset with four explanation choices for every question. The proposed task is to select the correct explanation, which is usually impractical. Moreover, VCR only considers people as visual objects in explanation, limiting its application range. Recently, Chen and Zhao [7] convert the reasoning step graph to multimodal explanations, proposing the REX dataset and a generative task, as shown in Figure 1. However,

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn



(a) Language quality (b) Visual precision

Figure 2: Performance of MEVQA methods with varying number of training samples.

their customized format reduces the readability of the explanations. Moreover, they use a Faster R-CNN [36] trained on MS-COCO [25] to extract object boxes for explanation annotation, which can only detect 81 classes and bring inaccuracy in explanations. Furthermore, existing generative methods for multimodal explaining [7, 50] rely on large-scale training data, which is costly and may be impractical in various vertical domains. As shown in Figure 2, when trained on a small amount of data, state-of-the-art methods experience a notable decline in performance, despite being built upon pretrained Vision-Language Models (VLMs). Especially considering the recent progress in open-world learning (e.g., LLMs), more streamlined and applicable learning frameworks may represent the future direction. Therefore, we propose the Few-Shot MEVQA (FS-MEVQA) task that aims to learn MEVQA from few training samples. Meanwhile, GPT-4V, which may currently be the most powerful multimodal LLM, also performs unsatisfactorily, showing the flaws of LLMs in explaining the implicit multimodal reasoning processes. Inspired by the observation of the above pioneering works in MEVQA, we attempt to promote MEVQA from both data and model perspectives, proposing a new multimodal explanation dataset and a novel LLM agent-based training-free method for FS-MEVQA.

To improve the readability and user-friendliness of multimodal explanations, we propose the Standard Multimodal Explanation (SME) dataset, where the explanations are in standard English with additional $[BOX]$ tokens for visual grounding, as shown in Figure 1. Our constructed explanations are based on reasoning steps for solving the visual questions, with the associated key objects in Visual Genome scene graphs [17] of 1,703 classes. After converting the structured reasoning steps into natural language-like explanations with visual object annotations, we utilize GPT-3.5 [6] to revise the explanations without $[BOX]$ into standard English. Moreover, to facilitate an effective evaluation of visual objects grounded in multimodal explanations, we annotate key visual objects with their corresponding names mentioned in explanations and add special $[BOX]$ tokens to represent grounded visual boxes. We further propose a visual detection metric to evaluate the ability to simultaneously generate names and ground regions of key visual objects to explain the multimodal reasoning processes. As shown in Table 1, compared to the existing explanations for VQA, our multimodal explanations based on standard English with additional $[BOX]$ tokens can be more user-friendly and expressive with far more visual objects involved in explanation, bridging MEVQA to a broad area in Natural Language Processing (NLP).

To overcome the dependency on large-scale training data, we propose a training-free Multimodal Explaining Agent (MEAgent) method based on an LLM agent with multimodal open-world tools to infer answers and generate multimodal explanations for visual questions, given merely $N(=16)$ training samples. Traditionally, $N-shot$ few-shot learning represents given N training samples for every class (i.e., every answer in VQA) [8, 39, 45], which is still costly and cannot address unseen classes in the test. Moreover, considering the recent progress in open-world learning [16, 19, 31, 35, 49], the paradigm of learning reasoning knowledge for every class may be outdated. Therefore, we propose a stronger and more practical few-shot learning setting for MEVQA, where only N training samples are given for all classes. Under such a setting, the value of training samples mostly lies in defining the MEVQA task, while the model needs out-of-training knowledge and open-world tools to solve the test questions. Inspired by the recent work in LLM-based visual programming [14], we construct a GPT-3.5-based LLM agent with multimodal open-world tools, such as an open-world object detector, image croppers, and customized Python functions. With a few-shot program prompt and a few-shot explanation prompt constructed based on $N(=16)$ in-context examples, our MEAgent can generate the multimodal program for solving the input question, execute it via open-world tools to infer the answer, and translate the execution process into a multimodal explanation of the multimodal reasoning process. Additionally, we propose a rethinking mechanism to complete the visual objects needed in the explanation but ignored in multimodal programming. Extensive experiments demonstrate that MEAgent significantly outperforms the state-of-the-art MEVQA methods [7, 50] with even thousands of training samples and the multimodal LLM GPT-4V [1] for FS-MEVQA.

Table 1: Comparison of explanations for VQA. *Visual Object* denotes visual objects grounded in explanation.

Dataset	Modality	Visual Object	Format
VQA-E [22]	Language	None	Standard English
VCR [51]	Language-Vision	Only people	Customized
REX [7]	Language-Vision	81	Customized
SME (Ours)	Language-Vision	1,703	Standard English with $[BOX]$

In brief, the contributions of our work are as follows:

- We propose SME, a new dataset for Multimodal Explanation for Visual Question Answering (MEVQA) comprising 1,028,230 samples, with 1,703 visual objects requiring detection in explanations. To our knowledge, this is the first dataset where the explanations are in standard English with additional visual grounding tokens, bridging MEVQA with a broad area in NLP.
- We propose MEAgent, a novel training-free method based on an LLM agent with multimodal open-world tools for Few-Shot MEVQA (FS-MEVQA). MEAgent can infer answers and generate multimodal explanations of the reasoning processes for visual questions, given only $N(=16)$ training samples. Additionally, we propose a rethinking mechanism to complete the visual objects needed in explanation but ignored in multimodal programming.
- Extensive experiments demonstrate the effectiveness of our dataset and method for FS-MEVQA. We adopt language quality metrics, visual detection metric, and visual attribution metrics to evaluate the generated multimodal explanations. Experimental results show the superiority of our MEAgent compared to the state-of-the-art MEVQA methods and the multimodal LLM GPT-4V.

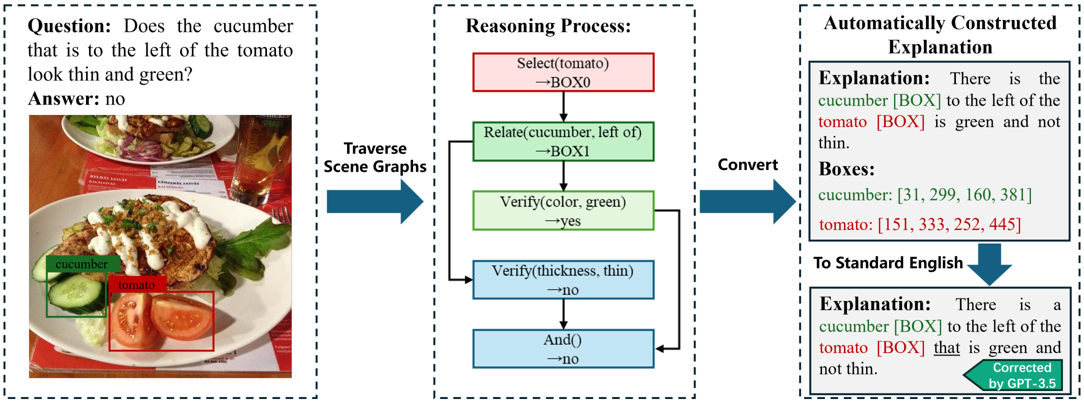

Question: Does the cucumber that is to the left of the tomato look thin and green?
Answer: no
cucumber
tomato
Traverse Scene Graphs
Reasoning Process:
Select(tomato) →BOX0
Relate(cucumber, left of) →BOX1
Verify(color, green) →yes
Verify(thickness, thin) →no
And() →no
Convert
Automatically Constructed Explanation
Explanation: There is the cucumber [BOX] to the left of the tomato [BOX] is green and not thin.
Boxes:
cucumber: [31, 299, 160, 381]
tomato: [151, 333, 252, 445]
To Standard English
Explanation: There is a cucumber [BOX] to the left of the tomato [BOX] that is green and not thin.
Corrected by GPT-3.5

Figure 3: Brief construction procedure of our multimodal explanations: (1) Traverse scene graphs to link reasoning processes with visual objects; (2) Convert reasoning processes into multimodal explanations by automatic programs; (3) Convert the text (excluding $[BOX]$) into standard English by GPT-3.5 and add $[BOX]$ back.

2 RELATED WORK

We review three highly related topics in explainable Visual Question Answering (VQA).

Textual explanation for VQA. Researchers have long recognized the importance of explanations in yielding verifiable reasoning results and establishing trustworthy VQA systems. Early attempts focus on providing textual explanations [21, 22, 28, 30, 44]. Zhou et al [53] propose a multi-task model to predict the answer to a visual question and generate the image caption to explain the answer simultaneously. Since their VQA data and image captioning data are from independent sources, the caption may be irrelevant to the question. Following their work, Li et al. [22] propose the VQA-E dataset by adopting the answer-related caption as the explanation. Differently, Wen and Peng [47] propose to retrieve related text from databases to explain the reasoning process.

Though textual explanations are simple and easy to obtain, researchers have also acknowledged their limitations. While VQA involves reasoning on multiple modalities, textual explanations often fail to effectively explain the visual concepts involved in the reasoning processes. Therefore, recent research has begun to explore Multimodal Explanation for VQA (MEVQA).

Multimodal explanation for VQA. Early research on MEVQA attempts to provide individual visual explanations in addition to textual explanations [34, 48]. These works typically relate visual attention maps to the predicted answers or explanations. However, attention maps can be confusing. For example, textual explanations “a man riding on a horse" and “a horse riding on a man" can have similar attention maps. To provide more expressive and accurate explanations of the reasoning processes, joint language-vision explanations are proposed [7, 50, 51]. For example, Chen and Zhao [7] propose to represent visual objects in the textual explanations by their grounding number predicted by a trained Faster R-CNN [36], belonging to only 81 classes. Moreover, their constructed explanations are in unnatural format, reducing the readability and applicability. Following their research, Xue et al. [50] recently propose a variational causal model to improve answer-explanation consistency by establishing the corresponding causal correlation.

However, traditional MEVQA methods rely on large-scale training data, which are costly and may be impractical in various vertical domains. Moreover, considering that the recent open-world models (e.g., LLMs) have already learned open-world abilities [27, 46, 52], more streamlined and applicable learning frameworks may represent the future direction. Therefore, we propose the SME dataset, where the explanations are in standard English with additional visual grounding tokens, supporting the research of Few-Shot MEVQA (FS-MEVQA). To perform FS-MEVQA, our proposed MEAgent can leverage a GPT-3.5-based agent with multimodal open-world tools to infer answers and generate multimodal explanations for visual questions, given only $N(=16)$ training samples.

Few-shot explanation for VQA Recent studies have found that LLMs are powerful few-shot learners [6, 23, 38]. Therefore, several LLM-based few-shot explanation methods have been developed [28, 29], which mainly focus on textual explanation. For example, Lu et al. [28] input the image caption and the question with in-context examples into GPT-3 to generate a chain-of-thought (i.e., textual explanation) and the answer. Different from the existing methods, we propose a few-shot method to generate multimodal explanations for visual questions, thereby enhancing the understanding of multimodal reasoning processes for VQA.

3 DATASET

Explaining reasoning processes for visual questions can aid in understanding and verifying the predicted answers, thereby enhancing the reliability and credibility of VQA systems. Our proposed dataset aims to provide an effective benchmark for MEVQA. Compared to previous multimodal explanations for VQA [7, 34, 51], our dataset has three key advantages: (1) Instead of separate textual and visual explanations [34], we integrate language and vision to construct brief but expressive explanations; (2) Instead of using unnatural formats [7, 51], our explanations are in standard English with additional $[BOX]$ tokens for visual grounding, bridging to broad advancements in Natural Language Processing (NLP). We annotate both object boxes and their corresponding names mentioned in explanations, which support (3) a more effective visual metric for measuring the ability to simultaneously generate names and ground regions of key visual objects to explain the multimodal reasoning processes. Figure 3 demonstrates a brief construction procedure of our multimodal explanations.

Data collection. We leverage VQA samples in the GQA dataset [15]. Given the 127 different operations in GQA, we first represent each reasoning step in the reasoning process by a <operation, relation/attribution, dependency 1, dependency 2> tuple, where "dependency 1" and "dependency 2" can be previous reasoning steps. Then, we follow [7] to categorize all operations into 12 atomic operations that cover the essential semantics, i.e., $Select$, $Exist$, $Filter$, $Query$, $Verify$, $Common$, $Same$, $Different$, $Compare$, $Relate$, And, and Or. For example, to answer "Does the cucumber that is to the left of the tomato look thin and green?", we need to find the tomato first by executing "<$Select$, $tomato$, $None$, $None$>", or "$Select(tomato)$" in abbreviation.

3.1 Multimodal Explanation Construction

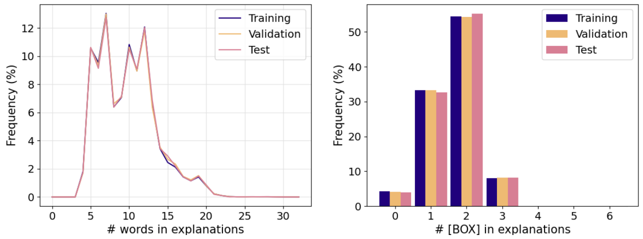
With reasoning step tuples, we need to convert reasoning steps into explanations. Previous work [7] uses a Faster-RCNN [36] trained on MS-COCO [25] to annotate visual objects related to reasoning steps, since their proposed explanation model also uses Faster-RCNN to extract visual objects. However, Faster-RCNN is often inaccurate for annotating objects in the GQA dataset. In contrast, we traverse the human-annotated scene graphs to obtain the results of all steps and link reasoning steps with their related visual objects. Then, we convert the reasoning process graphs into explanations with templates designed for all operations. For example, a $Verify$ operation "<$Verify$, ATT, $DEP1$, $None$>" is converted into "(DEP1) is [RETURN] ATT", where [RETURN] is "not" if this operation returns "no" or empty otherwise. (DEP1) is the phrase converted from the dependent operation. Moreover, we use the $[BOX]$ tokens in explanations to represent the grounding boxes of visual objects related to the reasoning steps. To facilitate the visual metric presented in Section 3.2, we annotate both the values of grounding boxes and their corresponding names mentioned in the explanations. We also design programs to correct some grammar issues in generated explanations, e.g., merging two $Verify$ operations in Figure 3. However, rigid programs cannot address all language errors in the constructed explanations. Therefore, we leverage the powerful language ability of GPT-3.5 [6] to correct our explanations. To avoid disruptions of $[BOX]$, we remove $[BOX]$ in the explanations and utilize GPT-3.5 to convert the explanations into standard English. Then, we add back the $[BOX]$ tokens according to their names. Finally, we take several rounds of manual checks to further clean the constructed explanations. After removing a small number of low-quality samples, we obtain 1,028,230 multimodal explanations in standard English with additional $[BOX]$ tokens for visual grounding. All data are divided into 901,203 training samples, 97,027 validation samples, and 30,000 test samples. In the MEVQA task, the model is required to answer the visual question, generate a textual representation of the explanation (e.g., "There is a cucumber [BOX] to the left of the tomato [BOX] that is green and not thin."), and grounds visual boxes linked to all $[BOX]$ tokens.

3.2 Metrics

To evaluate the generated multimodal explanations (aka., reference explanations), we adopt textual and visual metrics. **For textual metrics**, since we only add $[BOX]$ into standard English, we directly adopt widely-used language metrics, i.e., BLEU-4 [33], METEOR [5], ROUGE-L[24], CIDEr [43], and SPICE [2]. **For visual metrics**, we propose a new detection metric to comprehensively evaluate the ability to generate the names of visual objects and ground their corresponding regions in images. For every object name s annotated in a ground truth explanation (e.g., "cucumber" and "tomato" in Figure 3), we match the $[BOX]$ token following s in the reference explanation. Then, we compute the IoU (intersection over union) score between the ground truth boxes B^s_{gt} of s and the reference boxes B^s_{re} related to this $[BOX]$ token, evaluating the detection precision of this object. The final detection score of one explanation is averaged over all object names, as follows:

$$Detection = \frac{1}{N}\sum_{s} IoU(B^s_{gt}, B^s_{re}), \quad (1)$$

where N is the number of object names that occur in the ground truth and reference explanations. Therefore, the redundant boxes in the reference explanation and the missing boxes can punish the final detection score. Moreover, we propose visual attribution metrics to evaluate the ability of models to understand key visual attributions and generate them in explanations for MEVQA, which is introduced in Section 5.5.



(a) Number of words

(b) Number of $[BOX]$

Figure 4: Distributions of the explanation lengths and the numbers of $[BOX]$ in explanations of our SME dataset.

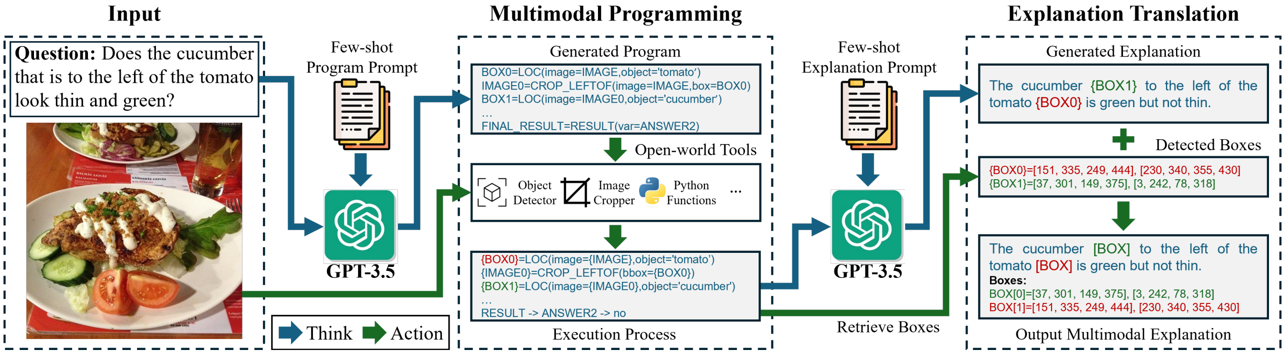
More discussions of our adopted metrics, especially their improvements compared to previous work, are included in Supplementary Materials.

3.3 Data Analysis

Our SME dataset is composed of 901,203 training samples, 97,027 validation samples, and 30,000 test samples.

Distribution of Multimodal Explanations. The main statistics of multimodal explanations in the training, validation, and test sets are shown in Figure 4. Three split sets have similar distributions of the explanation length and the number of $[BOX]$ in explanations. The explanation lengths range from 3 words to 32 words. The majority (97.0%) of explanations consist of between 5 to 19 words, indicating the requirement of generating brief but expressive text. The numbers of $[BOX]$ in explanations range from 0 to 6, which corresponds to the number of visual objects that should be grounded for explaining. The majority (95.8%) of explanations contain at least 1 $[BOX]$, highlighting the crucial role of vision in our multimodal explanations. When comparing three split sets, the test set shows a slightly higher proportion of explanations with 2 or 3 $[BOX]$, which improves the difficulty of the test set.

More information about our dataset is provided in Supplementary Materials.



```
BOX0=LOC(image=IMAGE,object='tomato')
IMAGE0=CROP_LEFTOF(image=IMAGE,box=BOX0)
BOX1=LOC(image=IMAGE0,object='cucumber')
...
FINAL_RESULT=RESULT(var=ANSWER2)
```

```
{BOX0}=LOC(image={IMAGE},object='tomato')
{IMAGE0}=CROP_LEFTOF(bbox={BOX0})
{BOX1}=LOC(image={IMAGE0},object='cucumber')
...
RESULT -> ANSWER2 -> no
```

The cucumber {BOX1} to the left of the tomato {BOX0} is green but not thin.

```
{BOX0}=[151, 335, 249, 444], [230, 340, 355, 430]
{BOX1}=[37, 301, 149, 375], [3, 242, 78, 318]
```

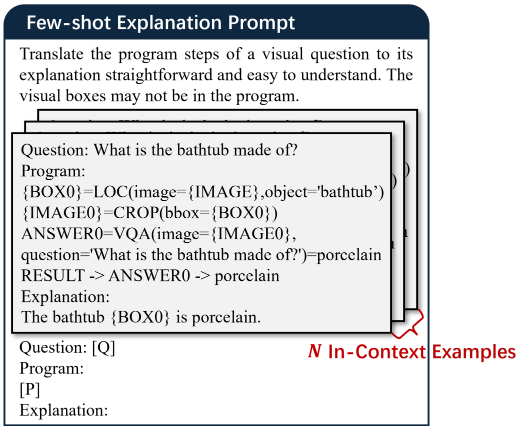
The cucumber [BOX] to the left of the tomato [BOX] is green but not thin.
Boxes:
BOX[0]=[37, 301, 149, 375], [3, 242, 78, 318]
BOX[1]=[151, 335, 249, 444], [230, 340, 355, 430]

Figure 5: Framework of our proposed Multimodal Explaining Agent (MEAgent) method: (1) Convert the input question into the program and execute it by open-world tools, with a few-shot program prompt; (2) Translate the execution process into the explanation with a few-shot explanation prompt and link the detected visual boxes to form the multimodal explanation.

4 METHOD

State-of-the-art MEVQA methods [7, 34, 50] rely on large-scale training data for learning explanation, despite being built upon pretrained Vision-Language Models (VLMs). However, annotating VQA data with explanations is costly and may be impractical in various vertical domains. Meanwhile, recent research [14, 29, 40] has shown that LLM-based visual programming is effective for few-shot VQA. Inspired by the above observations, we propose a training-free Multimodal Explaining Agent (MEAgent) method for Few-Shot MEVQA (FS-MEVQA) in this paper, given merely $N(=16)$ training samples. MEAgent leverages GPT-3.5 and multimodal open-world tools to conduct think and actions for FS-MEVQA, composed of Multimodal Programming and Explanation Translation stages. Figure 5 illustrates the framework of our method.

Multimodal Programming. In Multimodal Programming (MulProg), we decompose the input question Q into program steps, which are then executed with the input image I by multimodal open-world tools. Specifically, we select a set of multimodal open-world tools as our program modules. For example, we use the open-world object detection model OWL-ViT [31] as our location module $LOC(image, object)$. These modules can be flexibly customized and expanded to improve the reasoning capability of our MEAgent. To convert Q into a program instantiated by pre-defined program modules, we use $N(=16)$ in-context examples to form the few-shot program prompt that exemplifies the correspondence between questions and programs. To improve the effectiveness, our designed prompt includes all 16 selected modules. By inputting the prompt and the input question Q, GPT-3.5 can comprehend the correspondence and generate the program for solving Q. Then, multimodal open-world tools are activated to execute the program, inferring the answer with a program-like execution process. We denote the textual representation of the execution process as P, while values of variables are not directly contained in P. More details of Multimodal Programming are included in Supplementary Materials.


Few-shot Explanation Prompt

Translate the program steps of a visual question to its explanation straightforward and easy to understand. The visual boxes may not be in the program.

Question: What is the bathtub made of?
Program:
{BOX0}=LOC(image={IMAGE},object='bathtub')
{IMAGE0}=CROP(bbox={BOX0})
ANSWER0=VQA(image={IMAGE0},
question='What is the bathtub made of?')=porcelain
RESULT -> ANSWER0 -> porcelain
Explanation:
The bathtub {BOX0} is porcelain.

Question: [Q]
Program:
[P]
Explanation:

Figure 6: Our few-shot explanation prompt. $[Q]$ denotes the input question and $[P]$ denotes its execution process.

Explanation Translation. Though the MulProg stage provides a program-like execution process for solving the visual question, such a process is abstract, verbose, and hard to read by general users. Therefore, this stage aims to translate the execution process into a user-friendly multimodal explanation. The textual representation P of the execution process obtained by MulProg can be comprehended by GPT-3.5. Motivated by this, as shown in Figure 6, we construct a few-shot explanation prompt, including $N(=16)$ in-context examples to demonstrate the correspondence between execution processes and explanations in standard English with additional $\{BOXi\}$ tokens that represent the box variables in programs. For the input question Q and its execution process P, the prompted GPT-3.5 can generate the explanation for the reasoning process. However, since MulProg aims to solve the visual question and neglects to explain the reasoning process, some key visual objects for explanation may be ignored in the generated program. To address this problem, we propose a simple yet effective rethinking mechanism to improve the generated explanation. As shown in our prompt, MEAgent allows GPT-3.5 to directly output the needed objects that may not be in the program. Next, MEAgent checks all box variables in the generated explanation and links those existing in the program to their computed values. For visual boxes ignored in the program, MEAgent extracts their object names in the explanation and utilizes the location module LOC to compute their values. Finally, MEAgent can obtain the multimodal explanation with $[BOX]$ tokens to represent boxes of key visual objects.

5 EXPERIMENTS

We conduct extensive experiments on the proposed SME dataset to investigate the effectiveness of our proposed MEAgent method for FS-MEVQA. More details are included in Supplementary Materials.

Table 2: The results of Few-Shot Multimodal Explanation for Visual Question Answering on the SME dataset. N denotes the number of training samples. Gray results are trained on the whole training set.

Method	Backbone	N	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Detection	ACC
REX	VisualBERT	16	0.00	4.37	23.23	0.89	0.00	0.00	17.77
REX	VisualBERT	64	0.00	4.84	24.08	0.91	0.00	0.00	17.77
REX	VisualBERT	256	0.00	4.87	24.76	0.98	0.00	0.00	18.05
REX	VisualBERT	1K	8.84	10.83	38.60	34.86	4.10	0.43	18.19
REX	VisualBERT	4K	33.88	24.98	59.24	99.31	26.63	1.84	27.50
REX	VisualBERT	16K	56.35	39.22	75.49	275.16	48.42	3.94	33.04
REX	VisualBERT	All	84.26	56.22	90.50	763.63	78.88	7.66	61.38
REX	LXMERT	All	87.45	58.90	92.15	798.55	82.69	7.82	71.91
VCIN	VisualBERT	16	9.17	19.82	33.34	4.28	13.39	0.28	17.77
VCIN	VisualBERT	64	20.27	25.61	47.53	9.72	25.97	0.41	17.77
VCIN	VisualBERT	256	29.69	28.94	53.96	24.35	27.96	0.67	18.05
VCIN	VisualBERT	1K	37.00	31.99	59.07	54.01	29.84	1.10	24.10
VCIN	VisualBERT	4K	47.07	35.49	67.91	142.59	39.65	4.79	28.62
VCIN	VisualBERT	16K	59.42	41.54	75.54	309.85	50.93	9.77	35.71
VCIN	VisualBERT	All	90.64	61.80	93.73	833.37	86.63	21.90	64.00
VCIN	LXMERT	All	91.52	62.96	94.44	847.43	88.41	23.04	73.31
GPT-4V	GPT-4V	16	45.51	35.17	52.67	269.68	37.67	7.00	42.30
MEAgent (Ours)	**GPT-3.5**	**16**	**67.91**	**50.55**	**79.41**	**510.44**	**64.09**	**29.09**	**51.45**

5.1 Experimental Setup

Baseline methods. We adopt three baseline methods, including two state-of-the-art methods for MEVQA, as follows:

REX [7] grounds visual objects by Faster-RCNN [36] and utilizes an LSTM-based generator to generate multimodal explanations, based on features extracted by pretrained Vision-Language Models (VLMs).

VCIN [50] is the state-of-the-art method for MEVQA, which utilizes a gating Transformer to generate explanations and establish causal correlations to improve explanation-answer consistency, based on Faster-RCNN and pretrained VLMs.

Experiments reported in [50] adopt VisualBERT [20] and LXMERT [41] as backbone VLMs. Given that LXMERT is pretrained on the GQA dataset which overlaps with the questions and images in our dataset, we only adopt VisualBERT in few-shot experiments.

GPT-4V(ision) [1] may be currently the most powerful multimodal LLM. Since GPT-4V can generate text and detect visual objects, we also construct a prompt with the same $N(=16)$ examples in our method to facilitate question answering and multimodal explanation generation via GPT-4V. The details are included in Supplementary Materials.

Evaluation. As introduced in Section 3, we adopt BLEU-4 [33], METEOR [5], ROUGE-L[24], CIDEr [43], and SPICE [2] to evaluate the language quality. To evaluate the ability to generate the names of visual objects and ground their corresponding regions in images, we leverage our proposed visual detection metric. Moreover, we report the answering accuracy (abbr., ACC).

Implementation details. While our SME dataset of 1,028,230 samples can also be adopted in the research of traditional MEVQA, we focus on FS-MEVQA in our experiments. In the few-shot setting, different from traditional N-shot learning [8, 39, 45], we randomly sample N training samples for training and evaluate models on the whole test set. All results are averaged on 5 runs with different random seeds. More details are in Supplementary Materials.

5.2 Results and Discussions

The experimental results on our SME dataset are shown in Table 2. From the results, we have the following observations:

- The performance of traditional MEVQA methods (i.e., REX and VCIN) significantly drops when N is small. This shows these methods cannot effectively conduct FS-MEVQA, though being built upon pretrained VLMs. These methods rely on large-scale training data, which may not be an ideal training paradigm in practice, especially considering the recent progress in open-world learning and LLM.
- The detection scores of REX and VCIN are much lower than that of our MEAgent, even trained on all 901,203 training samples. This can be attributed to their adopted Faster R-CNN for pre-extracting 36 objects for every image, which may not contain the needed objects and lacks open-world objects. Differently, we use an open-world detector OWL-ViT to detect objects with their names mentioned in explanations.
- Our MEAgent significantly outperforms GPT-4V with the same N and state-of-the-art MEVQA methods with even thousands of training samples, showing its effectiveness for FS-MEVQA. Though 16 samples are insufficient for learning question answering and explaining, they are effective in defining our standard English-like multimodal explanations. Based on this, our LLM agent-based method can leverage out-of-training knowledge and multimodal open-world tools to perform FS-MEVQA.
- The language quality (i.e., BLEU-4, METEOR, ROUGE-L, CIDEr, and SPICE) of REX and VCIN trained on all training samples significantly outperforms GPT-4V and our MEAgent with few training samples. This shows the quality of our constructed training set and the current performance gap between traditional MEVQA and FS-MEVQA, suggesting more future work on FS-MEVQA.

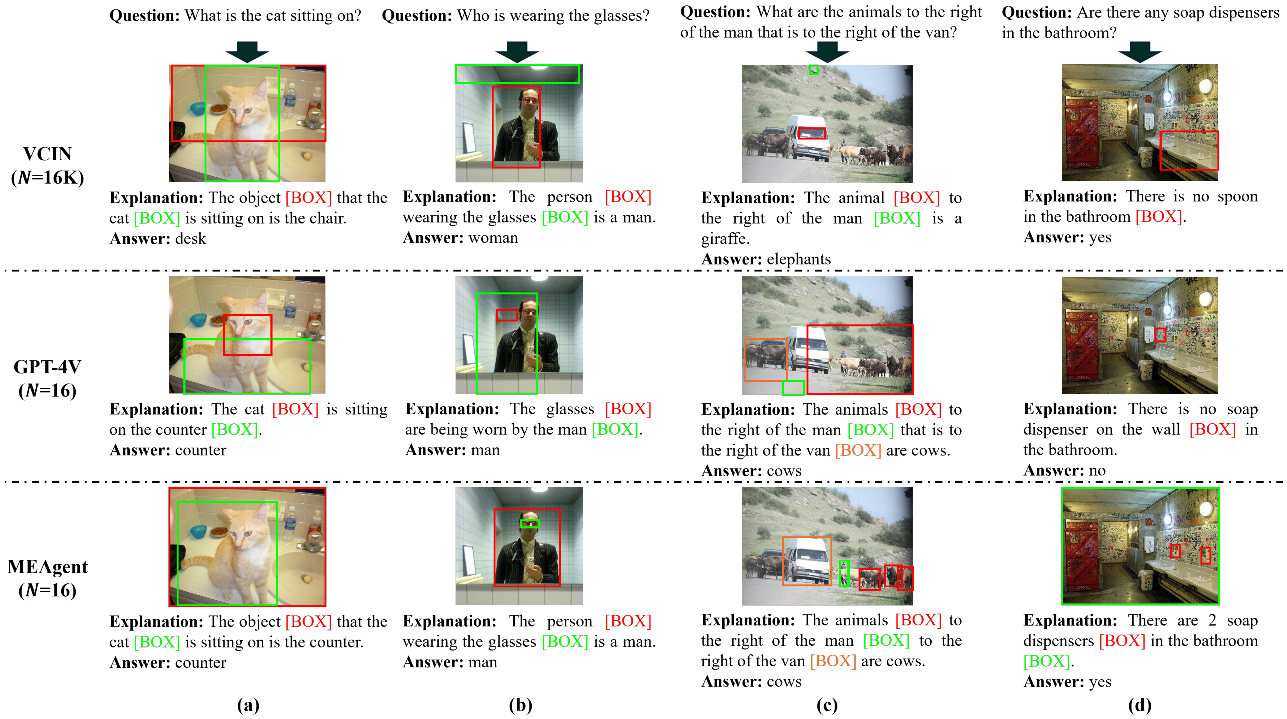

Question: What is the cat sitting on?
Question: Who is wearing the glasses?
Question: What are the animals to the right of the man that is to the right of the van?
Question: Are there any soap dispensers in the bathroom?
VCIN (N=16K)
Explanation: The object [BOX] that the cat [BOX] is sitting on is the chair.
Answer: desk
Explanation: The person [BOX] wearing the glasses [BOX] is a man.
Answer: woman
Explanation: The animal [BOX] to the right of the man [BOX] is a giraffe.
Answer: elephants
Explanation: There is no spoon in the bathroom [BOX].
Answer: yes
GPT-4V (N=16)
Explanation: The cat [BOX] is sitting on the counter [BOX].
Answer: counter
Explanation: The glasses [BOX] are being worn by the man [BOX].
Answer: man
Explanation: The animals [BOX] to the right of the man [BOX] that is to the right of the van [BOX] are cows.
Answer: cows
Explanation: There is no soap dispenser on the wall [BOX] in the bathroom.
Answer: no
MEAgent (N=16)
Explanation: The object [BOX] that the cat [BOX] is sitting on is the counter.
Answer: counter
Explanation: The person [BOX] wearing the glasses [BOX] is a man.
Answer: man
Explanation: The animals [BOX] to the right of the man [BOX] to the right of the van [BOX] are cows.
Answer: cows
Explanation: There are 2 soap dispensers [BOX] in the bathroom [BOX].
Answer: yes
(a)
(b)
(c)
(d)

Figure 7: Qualitative results of the generated multimodal explanations and the predicted answers. The $[BOX]$ tokens link to visual boxes of the same colors in the images.

Table 3: Ablation results on the SME test set. N denotes the number of training samples. MEAgent-RTK denotes MEAgent without the rethinking mechanism.

Method	N	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE	Detection
MEAgent-RTK	16	65.15	46.31	75.66	456.03	55.39	23.58
MEAgent	4	52.50	41.72	69.61	372.97	45.32	15.77
MEAgent	8	62.93	47.39	75.85	477.38	59.06	22.52
MEAgent	**16**	**67.91**	**50.55**	**79.41**	**510.44**	**64.09**	**29.09**

5.3 Ablation Study

In this section, we ablate our rethinking mechanism and investigate the effect of N in our method. The ablation results are shown in Table 3, from which we have the following observations: (1) MEAgent-RTK removes the rethinking mechanism and only uses visual objects detected in multimodal programming, which performs worse than MEAgent, especially on the detection metric. This indicates the effectiveness of the rethinking mechanism which can generate and ground objects ignored in the previous multimodal programming. (2) MEAgent performs better with more training samples (i.e., N). However, compared to the results in Table 2, our MEAgent with $N = 4$ can still outperform GPT-4V with $N = 16$ or REX and VCIN with $N = 4$K. Our MEAgent with $N = 8$ can still outperform REX and VCIN with $N = 16$K. These results further verify the effectiveness of our method.

5.4 Qualitative Results

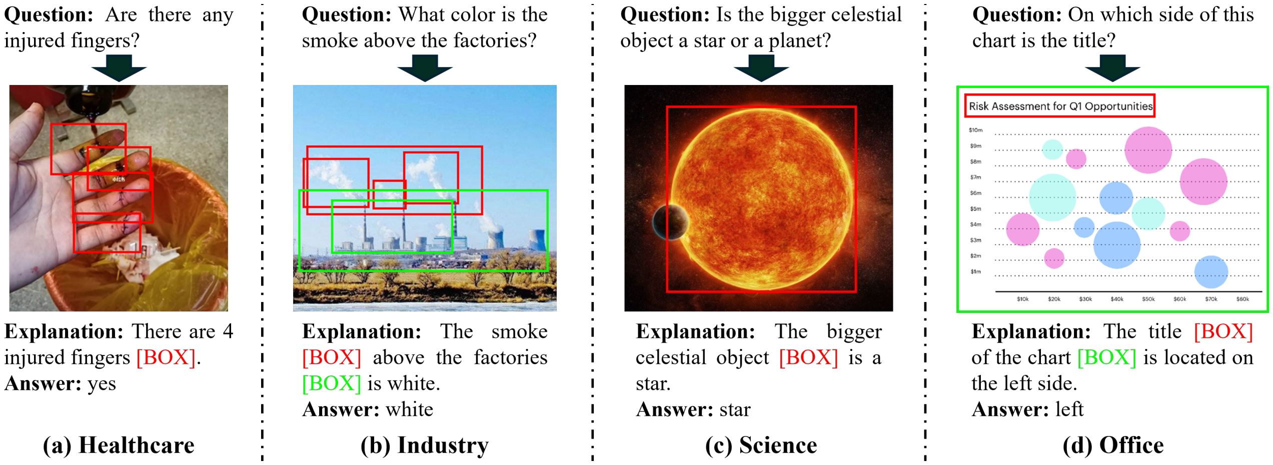
In addition to quantitative results, we demonstrate qualitative results of the predicted answers and explanations for FS-MEVQA in Figure 7. Compared to VCIN (N=16K) and GPT-4V (N=16), our MEAgent (N=16) can generate more rational, accurate, and coherent explanations of the reasoning process: (1) In (b)-(d), VCIN cannot ground the key visual objects in the images for explanation, while our MEAgent can utilize the names of objects and an open-world detector to locate accurate boxes. (2) VCIN usually predicts inconsistent answers and explanations. This is because VCIN needs large-scale data to learn the causal correlation between the answer and explanation. Differently, we utilize LLM to translate the execution process of inferring the answer to its explanation, ensuring inherent consistency between answers and explanations. (3) In (b)-(d), the detection accuracy of GPT-4V is also unsatisfactory. In (d), GPT-4V grounds an unimportant object (i.e., "wall") in the explanation but ignores key objects (i.e., "soap dispensers" and "bathroom"), which shows GPT-4V fails to capture the correct reasoning process for solving this question. Differently, our MEAgent explicitly generates the multimodal program for solving the input question and generates the explanation accordingly. Therefore, MEAgent can capture rational reasoning processes more effectively.

5.5 Results of Visual Attribution Metrics

We further propose metrics to evaluate the ability of models to understand key visual attributions and generate them in explanations for FS-MEVQA. We have collected 38, 41, 16, 96, 6, 11, 14, 9, 79, 38, and 43 keywords about colors, materials, shapes, activities, sizes, poses, sports, directions, animals, people, and plants that occur in the test explanations, separately. Then, we compute the percentage of correctly generated keywords in the explanations for every

Table 4: The accuracy of explaining key visual attributions on the SME test set. N denotes the number of training samples.

Method	N	Color	Material	Shape	Activity	Size	Pose	Sport	Direction	Animal	Person	Plant
VCIN	16K	31.65	24.45	10.50	59.76	39.21	33.40	16.29	55.23	29.79	70.39	31.07
GPT-4V	16	63.11	59.07	61.43	65.27	70.48	71.56	68.42	69.13	72.73	48.68	60.00
MEAgent	**16**	**82.67**	**76.90**	**83.60**	**95.25**	**79.77**	**92.46**	**89.71**	**87.92**	**82.97**	**83.45**	**67.85**



Question: Are there any injured fingers?

Explanation: There are 4 injured fingers [BOX].
Answer: yes

(a) Healthcare

Question: What color is the smoke above the factories?

Explanation: The smoke [BOX] above the factories [BOX] is white.
Answer: white

(b) Industry

Question: Is the bigger celestial object a star or a planet?

Explanation: The bigger celestial object [BOX] is a star.
Answer: star

(c) Science

Question: On which side of this chart is the title?

Explanation: The title [BOX] of the chart [BOX] is located on the left side.
Answer: left

(d) Office

Figure 8: Out-of-distribution results of our MEAgent for FS-MEVQA. The $[BOX]$ tokens link to visual boxes of the same colors in the images.

attribution, which are demonstrated in Table 4. From the results, we have the following observations: (1) Explaining plants appears to be the most challenging one, which may be due to the limited training data on subdivided plant species in current research. Therefore, the experimented LLMs and VLMs may have less knowledge about plant species. (2) Interestingly, GPT-4V demonstrates superior performance compared to VCIN across all attributions except for *Person*, despite having lower language scores in Table 2. This can be attributed to the extensive open-world knowledge of GPT-4V, while VLM-based VCIN suffers from insufficient knowledge in few-shot learning. This comparison also shows the complementarity of visual attribution metrics in evaluating visual knowledge for explanation. (3) GPT-4V achieves an especially low score for *Person*, which indicates GPT-4V may be too cautious in identifying the biological sexes, ages, and occupations of people. (4) Our MEAgent significantly outperforms both baselines, showing its superior ability to understand visual attributions and explain them. Though GPT-3.5 can only understand and generate text, the GPT-3.5-based MEAgent can leverage open-world visual tools (e.g., object detector and image croppers) to perform multimodal explanation, even outperforming the stronger multimodal LLM GPT-4V.

5.6 Out-Of-Distribution Results

Given the limited knowledge within training samples in FS-MEVQA, a key to performing FS-MEVQA is leveraging open-world knowledge. Therefore, we further gather Out-Of-Distribution (OOD) visual questions that involve concepts not present in our SME dataset to test our MEAgent method. In Figure 8, we show the results for four OOD samples belonging to healthcare, industry, science, and office domains. We are surprised to find that MEAgent can predict rational answers and generate good explanations for these questions. In (a), since we implement a *COUNT* module to count the number of detected boxes, MEAgent can even explain the number of injured fingers. In (d), while MEAgent can accurately locate the title and the chart, its answer is not very accurate (“upper left" may be a better answer). These results may suggest a boarder application range of our MEAgent in various domains. By employing an LLM agent with open-world tools for visual reasoning and multimodal explanation, MEAgent is capable of overcoming the limited knowledge in training data and generalizing to encompass open-world questions. Moreover, we argue that utilizing domain-specific programming modules and examples can further improve the capability of MEAgent in a particular domain.

6 DISCUSSION AND CONCLUSION

In this paper, we propose a Standard Multimodal Explanation (SME) dataset with 1,028,230 samples for Visual Question Answering (VQA) with elaborately constructed multimodal explanations of the underlying multimodal reasoning processes. Based on our dataset, we propose a new Few-Shot Multimodal Explanation for VQA (FS-MEVQA) task, which aims to answer the visual question and explain the reasoning process with a limited number (denoted as N) of training samples. To the best of our knowledge, SME is the first large-scale dataset for MEVQA with joint language-vision explanations based on standard English and additional visual grounding tokens, which bridge MEVQA to a broad field in Natural Language Processing (NLP). Moreover, we propose a training-free Multimodal Explaining Agent (MEAgent) method based on an LLM agent with multimodal open-world tools for FS-MEVQA. Extensive experiments demonstrate that our MEAgent significantly outperforms traditional MEVQA methods and GPT-4V for FS-MEVQA.

REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 382–398.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[5] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Shi Chen and Qi Zhao. 2022. Rex: Reasoning-aware and grounded explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15586–15595.

[8] Yinbo Chen, Zhuang Liu, Huijuan Xu, Trevor Darrell, and Xiaolong Wang. 2021. Meta-baseline: Exploring simple meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9062–9071.

[9] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding* 163 (2017), 90–100.

[10] Ian J Deary, Lars Penke, and Wendy Johnson. 2010. The neuroscience of human intelligence differences. *Nature reviews neuroscience* 11, 3 (2010), 201–211.

[11] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.

[12] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2024. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems* 36 (2024).

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[14] Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14953–14962.

[15] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.

[16] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. 2021. Towards open world object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5830–5840.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.

[18] Wendy Lehnert. 1977. Human and computational question answering. *Cognitive Science* 1, 1 (1977), 47–73.

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. 19730–19742.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[21] Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018. Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1338–1346.

[22] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. 2018. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 552–567.

[23] Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. 2024. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18608–18616.

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[26] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2024. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems* 36 (2024).

[27] Keming Lu, Xiaoman Pan, Kaiqiang Song, Hongming Zhang, Dong Yu, and Jianshu Chen. 2023. PIVOINE: Instruction Tuning for Open-world Entity Profiling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 15108–15127.

[28] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[29] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2024. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[30] Ana Marasović, Chandra Bhagavatula, Jae Sung Park, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. *arXiv preprint arXiv:2010.07526* (2020).

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple open-vocabulary object detection. In *European Conference on Computer Vision*. Springer, 728–755.

[32] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2022. Coarse-to-fine reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4558–4566.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[34] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8779–8788.

[35] Shengsheng Qian, Hong Chen, Dizhan Xue, Quan Fang, and Changsheng Xu. 2023. Open-world social event classification. In *Proceedings of the ACM Web Conference 2023*. 1562–1571.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[37] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots-a comprehensive review. *Comput. Surveys* 55, 10 (2023), 1–41.

[38] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2023. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2998–3009.

[39] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *Comput. Surveys* 55, 13s (2023), 1–40.

[40] Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11888–11898.

[41] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5100–5111.

[42] Erico Tjoa and Cuntai Guan. 2020. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems* 32, 11 (2020), 4793–4813.

[43] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[44] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2413–2427.

[45] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)* 53, 3 (2020), 1–34.

[46] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Shawn Ma, and Yitao Liang. 2024. Describe, explain, plan and select: interactive planning with LLMs enables open-world multi-task agents. *Advances in Neural Information Processing Systems* 36 (2024).

[47] Zhang Wen and Yuxin Peng. 2020. Multi-level knowledge injecting for visual commonsense reasoning. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 3 (2020), 1042–1054.

[48] Jialin Wu and Raymond Mooney. 2019. Faithful Multimodal Explanation for Visual Question Answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 103–112.

[49] Manjie Xu, Guangyuan Jiang, Wei Liang, Chi Zhang, and Yixin Zhu. 2024. Active Reasoning in an Open-World Environment. *Advances in Neural Information Processing Systems* 36 (2024).

[50] Dizhan Xue, Shengsheng Qian, and Changsheng Xu. 2023. Variational Causal Inference Network for Explanatory Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2515–2525.

[51] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6720–6731.

[52] Sipeng Zheng, Yicheng Feng, Zongqing Lu, et al. 2023. Steve-Eye: Equipping LLM-based Embodied Agents with Visual Perception in Open Worlds. In *The Twelfth International Conference on Learning Representations*.

[53] Yiyi Zhou, Rongrong Ji, Jinsong Su, Yongjian Wu, and Yunsheng Wu. 2017. More than an answer: Neural pivot network for visual qestion answering. In *Proceedings of the 25th ACM international conference on Multimedia*. 681–689.