Evaluating VQA Models' Consistency in the Scientific Domain

Khanh-An C Quan, Camille Guinaudeau, Shin'Ichi Satoh

Evaluating VQA Models' Consistency in the Scientific Domain

Khanh-An C. Quan[1,2*], Camille Guinaudeau[3], and Shin'ichi Satoh[4]

[1] University of Information Technology, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Université Paris-Saclay / Japanese French Laboratory for Informatics, CNRS, Tokyo, Japan
[4] National Institute of Informatics, Tokyo, Japan
anqck@uit.edu.vn

Abstract. Visual Question Answering (VQA) in the scientific domain is a challenging task that requires a high-level understanding of the given image to answer a given question. Although having impressive results on the ScienceQA dataset, both LLaVA and MM-CoT models exhibit inconsistent answers when a simple modification is applied to the textual input of the question (e.g., choices re-ordering). In this paper, we propose two approaches that slightly modify the image-question pair without changing the question's meaning to gain a deeper comprehension of VQA models' question understanding: choices permutation and question rephrasing. Along with these two proposed approaches, we introduce two metrics, namely Consistency across Choice Variations (CaCV) and Consistency across Question Variations (CaQV), to measure the consistency of the VQA models. The experimental results show that both LLaVA and MM-CoT give inconsistent answers regardless of the accuracy. We further conducted a comparison between the proposed metrics and the Accuracy metric, demonstrating that relying solely on the Accuracy is inadequate. By revealing the limitations of existing VQA models and the Accuracy metric through evaluation results in the scientific domain, we aim to provide insights for motivating future research.

Keywords: Visual Question Answering · LLMs evaluation

1 Introduction

Visual Question Answering (VQA) is a challenging task that requires a high-level understanding of the given image to provide the answer to a given question. In particular, the model must understand various visual elements in this task, including recognizing instances, reading text, comprehending the visual characteristics of objects, or reasoning based on visual data to provide a response. On the other hand, integrating various forms of data, such as images and text, adds complexity to this task as the model needs to comprehend and leverage the relationships between these modalities.

* This work was conducted during Khanh-An C. Quan's internship at the National Institute of Informatics, Tokyo, Japan.

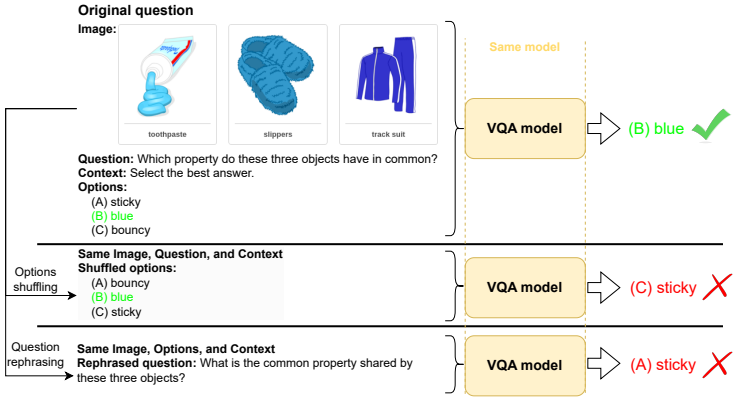

Fig. 1: Illustrate the inconsistent results predicted by the current VQA model using the same image-question pair but altering the order of options or rephrasing the question.

Scientific problem-solving benchmarks [8,17,14] have been employed to evaluate the multi-hop reasoning skills and the interpretability of AI systems. To address the questions in this field, the model must not only comprehend multimodal content but also retrieve external information to determine the correct answer. Among these recently proposed benchmarks for a scientific domain, the ScienceQA dataset [14] is a large-scale multichoice dataset with multimodal science questions along with explanations and has a wide range of domains.

Large Language Models (LLMs) have recently demonstrated impressive performance across a range of Natural Language Processing tasks [19,16]. Additionally, LLMs have shown the capability to address complex reasoning problems through chain-of-thought (CoT) processes by leveraging a small number of demonstration examples [2]. When integrated with input image data, Multimodal Large Language Models (MLLMs) achieve promising visual question answering (VQA) results, both in general contexts and specifically within scientific domains [12,24]. Despite MLLMs achieving notable results in the scientific domain, MLLMs require high computational costs due to their nature. On the other hand, another recent approach to this problem is VLM. In this direction, Multimodal-CoT (MM-CoT) is a starting point and achieves a comparative result to MLLMs. Specifically, MM-CoT combines textual and visual data within a two-stage approach, separating the rationale generation process from the answers inference stage. In comparison to the MLLMs approach, the VLM approach has significantly fewer parameters and a much faster computational speed.

Although both LLMs and VLMs obtain remarkable results on the ScienceQA dataset, these models can provide inconsistent output. Specifically, by simply altering the order of choices, these models can produce different answers to the same question and image. Figure 1 demonstrates the inconsistency in answer predictions of current VQA models when given the same image-question pair but with different choice orders or rephrased questions.

In this paper, we propose two approaches that modify the image-question pair without changing the question's meaning to have a better comprehension

of the VQA model's question understanding. In the first one, we assess each question in the dataset using all possible permutations of the choices instead of their original order. In the second approach, we rephrase the question into various forms and then evaluate VQA models using these rephrased questions. Essentially, VQA models should produce the same rationales and answers for the same question, regardless of the order of choices or the form of the question.

We demonstrate that relying solely on Accuracy metrics to evaluate VQA models is insufficient. Specifically, despite high-accuracy examples, the models still provide inconsistent answers with two proposed evaluation approaches. To address this limitation, we introduce two metrics: Consistency across Choice Variations (CaCV) and Consistency across Question Variations (CaQV) to measure the consistency of the VQA models. We assess the performance of two recent VQA models, LLaVA [12] and MM-CoT [25], using the ScienceQA dataset [14].

Our contributions can be summarized in four folds as follows:

- we introduce two approaches that make minor adjustments to the image-question pair without altering the question's meaning to gain a better understanding of VQA models: choices permutation evaluation and rephrasing question evaluation;
- we propose two metrics to measure the consistency of VQA models: CaCV and CaQV. we further compare these metrics with Accuracy, highlighting the limitations of using Accuracy as a sole measure;
- we conduct experiments on two current VQA models, LLaVA [12] and MM-CoT [25], and show that they achieve 89.07% and 94.12% respectively of CaCV and 87.48% and 91.77% of CaQV on the ScienceQA dataset [14];
- Finally, we draw insights into these inconsistent sample characteristics.

2 Related Works

Large Language Models (LLMs) Recently, the advancement of LLMs has demonstrated remarkable performance across various natural language tasks [19]. Various methods have been suggested to enhance multimodal understanding by leveraging the robust generality of LLMs, especially when integrated with other modalities like images [16,24,12]. In the vision-language field, Multimodal Large Language Model (MMLMs) yields remarkable results in various downstream tasks, especially in multimodal reasoning and visual question-answering (VQA) [16,12]. However, one of the main difficulties with MLLMS is its high computational cost and the requirement for large-scale, high-quality training data.

MLLMs Evaluation As MLLMs have advanced, many benchmarks have been proposed to evaluate comprehension abilities, such as [3,13,23,22,11]. Recent benchmarks, e.g. MME [3], MMBench [13], and SEED-Bench [11], assess MLLMs' comprehension abilities by creating multiple choice questions that span a range of ability dimensions. Li *et al.* [11] show that most MLLMs still exhibit limited performance on tasks that require fine-grained instance-level comprehension.

Chain-of-Though Reasoning LLMs have recently demonstrated impressive results by utilizing Chain-of-Thought (CoT) prompting techniques [9,21]. Specifically, CoT methods prompt the LLM to produce a step-by-step reasoning chain

to address a problem. There are two primary mechanisms to perform CoT reasoning on LLMs: Zero-Shot-CoT and Few-Shot-CoT. Kojima *et al.* [9] show that LLMs can perform Zero Shot-CoT by simply appending a prompt such as "Let's think step by step" to the question can trigger CoT reasoning. In Few-Shot-CoT, language models learn reasoning through a few examples demonstrating the step-by-step reasoning process. Recent research indicates that fine-tuned language models can evoke CoT reasoning in smaller models [15,4,5].

Visual Question Answering for Scientific Domain Solving science problems is a difficult task requiring an AI system to not only grasp multimodal information within the scientific domain but also require the model to explain how to address the questions. There are many proposed benchmarks for VQA in the scientific domain, such as AI2D [7], DVQA [6], VLQA [18], FOODWEDS [10], and ScienceQA [14]. Among these datasets, ScienceQA [14] incorporated reasoning into the VQA task, establishing a standard for multimodal chain-of-thought analysis. The ScienceQA dataset includes approximately 21,000 multimodal multiple-choice questions covering a wide range of science topics, along with annotated answers, related lectures, and explanations.

There are many recent works researching solving this problem, but in general, there are two main directions: utilizing LLM's capabilities and training a Vision-Language Model. Using chain-of-thought prompting, Lu *et al.* [14] demonstrate that a few-shot GPT-3 model can enhance reasoning performance on the ScienceQA dataset and produce reasonable explanations. However, since GPT-3 is an unimodal model that processes only language, captioning models are required to convert visual information into language modality. Employing caption generation models can lead to considerable information loss when dealing with highly complicated images. To overcome this issue, LLaVA [12] proposes a mechanism to embed visual information into LLM and achieve remarkable results on the ScienceQA dataset [14]. On the other hand, Multimodal-CoT (MM-CoT) [25] implements a two-stage framework, which separates the rationale step from the answer step and training with annotated CoT rationales. Compared to LLaVA, although having the same overall result, the computational cost of MM-CoT is significantly lower than LLaVA. Recently, T-SciQ [20]has shown that by combining the MM-CoT architecture and LLM's reasoning, the VQA performance can be further improved.

3 Methodology

3.1 Preliminaries

In this study, we concentrate on the task of Visual Question Answering [1], which requires the model to deliver the answer by utilizing the information given in the question along with the associated image. Specifically, considering a VQA dataset consisting of k $\{X, Y\}$ samples, where X represents multimodal inputs and Y indicates the corresponding ground-truth answers. The multimodal input X can be denoted as $X = \langle T, I \rangle$, where T refers to the text content and I represents the image content associated with the given question. Text content T can be decomposed into $T = \langle Q, C, M \rangle$, where Q represents the question, C

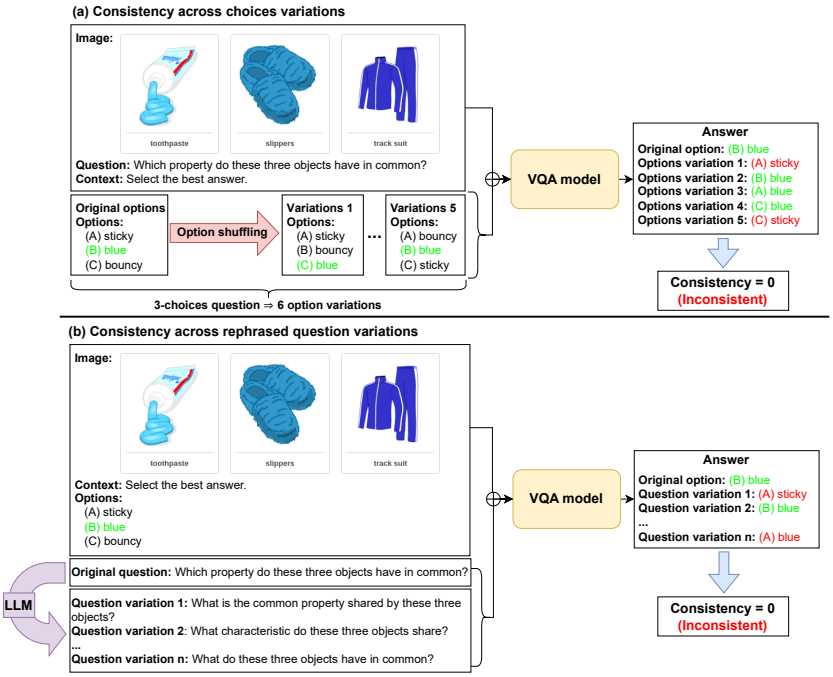

Fig. 2: Illustration of the workflow of our two proposed approaches for assessing the consistency of the current VQA model in the scientific domain.

denotes the context, and $M = (m_1, \ldots, m_k)$ is the list of possible options, and k is the number of choices of the given question. It is important to note that the list of choices M in the input for current VQA models is an ordered list. In simple form, the visual question answering can be described as follows:

$$Y = \arg\max_{Y',} p\left(Y' \mid T, I\right) \tag{1}$$

where $p\left(Y' \mid T, I\right)$ represents the likelihood of the answer Y' given the textual content T and the visual content I. Based on the basic VQA model, the visual question-answering reasoning model (such as GPT-3 [14], LLaVA [12]) involves generating a rationale R, which explains the chain-of-though that supports the answer Y. This can be mathematically described as follows:

$$Y, R = \arg\max_{Y',R'} p\left(Y', R' \mid T, I\right) \tag{2}$$

where $p\left(Y', R' \mid T, I\right)$ is the probability of the answer Y' and the rationale R' given the text content T and the image content I. In MM-CoT [25], rationale generation and answer inference are divided into two distinct stages as follows:

$$\begin{aligned} R &= \arg\max_{R'} p\left(R' \mid T, I\right) \\ Y &= \arg\max_{Y'} p\left(Y' \mid R, T, I\right) \end{aligned} \tag{3}$$

In the following section, we will use Equation 1 to simplify the description of the VQA model, which takes the image-text input and predicts the answer only.

3.2 Method 1: Choices Permutation Evaluation

Choices Permutation In this kind of evaluation, rather than using the predefined order of choices in the dataset, we attempt to assess the VQA models using all possible permutations of the choices. Generally, VQA systems are expected to produce the same answer when presented with the same question and a list of choices, regardless of the order in which the choices are arranged. Specifically, given a list of choices $M = (m_1, \ldots, m_k)$, we construct a set M^*, which includes every possible permutation of the list C. M^* can be denoted as:

$$\begin{aligned} M^* = \{&(m_p, m_q, \ldots, m_r) \mid 1 \leq p, q, \ldots, r \leq k, \\ &\text{and } (p, q, \ldots, r) \text{ is a permutation of } (1, 2, \ldots, k)\} \end{aligned} \tag{4}$$

Typically, a question with k-choices results in $k!$ different permutations of the choices (e.g., with a 3-choices question, we will have $3! = 6$ choices variations.) Subsequently, we create a text input set $T^* = \langle Q, C, M^* \rangle$ derived from all permutations of choices M^*. All variations of choices are input into the VQA models to yield $k!$ answers as follows:

$$Y^* = \{\arg\max_{Y'} p\left(Y' \mid T^*, I\right)\} \tag{5}$$

Finally, based on a list of predicted answers Y^*, we propose a metric - CaCV - to measure the stability of VQA models with the same question content and all permutations of choices.

Propose Metric: Consistency (%) across All Choices Variations (CaCV) The Consistency across all Choice Variations defines whether the model yields the same answer in all different choice variations or not. With an answers' list Y^* predicted by the VQA model based on different choice variations, CaCV for each sample in the dataset is measured as:

$$CaCV = \begin{cases} 1, & \text{if } y_i = y_j \; \forall \, y_i, y_j \in Y^* \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The overall CaCV for the dataset is derived by taking the mean of the CaCV values for each individual sample. We consider samples with a CaCV of zero as inconsistent and those with a CaCV of one as consistent.

3.3 Method 2: Question Rephrasing Evaluation.

Question Rephrasing Along with assessing the models by altering the order of choices in the questions, we also evaluated them based on rephrased questions. The aim of this evaluation is to determine whether the models genuinely understand the questions and provide answers based on their content. In this type of evaluation, we use a large language model (LLM) to rephrase the original question Q into various forms, denoted as $Q_{rephrase} = (Q_1, \ldots, Q_n)$, where n is the number of rephrased questions. In this work, we use ChatGPT-3.5 to rephrase the question with the prompt "Rephrase this question into n different form.". In this paper, we rephrase the original question in 5 different forms. The illustration of the rephrased question is presented in Figure 2. We also manually verified the correctness of the rephrasing question generated by the LLMs. Next,

we concatenate the rephrased questions $Q_{rephrase}$ with the original question Q to create Q^*. We then construct a text input set $T^{**} = \langle Q^*, C, M \rangle$ from all variations of Q^* and input this set into the VQA models.

$$Y^{**} = \{\arg\max_{Y'} p\left(Y' \mid T^{**}, I\right)\} \tag{7}$$

Using a list of predicted answers Y^{**}, we introduce CaQV metric to evaluate the robustness of VQA models when faced with different variations of questions.

Propose Metric: Consistency (%) across Questions Variations (CaQV). Similar to CaCV presented in Section 3.2, this metric aims to measure the consistency of the VQA models with different variations of questions. In particular, given a list of answers predicted by the VQA model from different question variations, CaQV is defined as:

$$CaQV = \begin{cases} 1, & \text{if } y_i = y_j \; \forall \, y_i, y_j \in Y^{**} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The overall CaQV for the dataset is obtained by averaging the CaQV values of each individual sample. Samples with a CaQV of zero are categorized as inconsistent, while those with a CaQV of one are categorized as consistent as the same CaCV.

4 Experiment

4.1 Experiment Setup

Dataset We use the ScienceQA [14] dataset for evaluation and analysis. This multimodal multiple-choice science question dataset includes 21,000 questions across three subjects, covering 26 topics, 127 categories, and 379 unique skills. The dataset is divided into training, validation, and test sets, with 12,726, 4,241, and 4,241 samples respectively. In this paper, we focus on the questions with 2, 3, or 4 choices from the ScienceQA test set, which includes 2,228 questions with 2 choices, 971 questions with 3 choices, and 1,004 questions with 4 choices.

Metrics For evaluation, we use the *Accuracy* metric, which compares the answer predicted by the model with the ground-truth from the dataset. In the evaluation of choice permutation, the accuracy of each sample is calculated by averaging the accuracy of all choice variations generated by the evaluated model. Similarly, in question rephrasing evaluation, the accuracy of each sample is measured by averaging the accuracy across all rephrased questions. We also use the proposed metrics *CaCV* (See Section 3.2) and *CaQV* (See Section 3.3) for choice permutation evaluation and question rephrasing evaluation, respectively.

Competing VQA Methods In this paper, we utilize two VQA models for benchmarking: LLaVA [12] and MM-CoT [25]. For the LLaVA model, we use the ScienceQA version pre-trained with 13 billion parameters and set its temperature to 0 for reproducibility. For the MM-CoT model, we employ the largest ScienceQA pre-trained model with 768 million parameters that achieved the highest performance.

Table 1: Overall percentage of Consistency (CaCV and CaQV) and Accuracy (Acc.) for different types of questions for LLaVA [12] and MM-CoT [25] models with choice shuffling (on the left, in white) and question rephrasing (on the right in blue) approaches on ScienceQA dataset [14]. The best results between the two models are highlighted in **bold**.

	Choice shuffling				**Question rephrasing**			
	LLaVA		**MM-CoT**		**LLaVA**		**MM-CoT**	
Question type	**CaCV**	**Acc.**	**CaCV**	**Acc.**	**CaQV**	**Acc.**	**CaQV**	**Acc.**
2-choices	95.37	**92.93**	**99.64**	92.14	85.18	**91.15**	**89.99**	90.78
3-choices	74.15	85.35	**88.36**	**86.06**	87.12	84.74	**92.79**	**86.03**
4-choices	89.54	**94.33**	**91.53**	92.80	92.92	93.71	**94.72**	**97.94**
Overall	89.07	**91.53**	**94.12**	90.96	87.48	90.28	**91.77**	**91.39**

Table 2: Results obtained in two types of situations. On the left, accuracy for consistent (Con.) vs. inconsistent (Inc.) examples. On the right, consistency (CaCV or CaQV) for questions with (Img.) and without images (W/o).

	Consistent / Inconsistent				**With / Without Image**			
	LLaVA		**MM-CoT**		**LLaVA**		**MM-CoT**	
	Inc.	**Con.**	**Inc.**	**Con.**	**Img.**	**W/o**	**Img.**	**W/o**
Choice shuffling	51.42	96.45	45.72	93.78	86.48	91.93	92.87	95.33
Question rephrasing	56.56	95.10	60.16	94.19	87.95	88.97	92.59	90.04

4.2 Results and Analysis

Overall The overall result of both choice shuffling and question rephrasing evaluations is shown in Table 1. Despite having the same accuracy, MM-CoT shows higher Consistency compared to LLaVA, which are 94.12% and 89.07% for choice shuffling and 87.48% and 91.77%, respectively. In comparison to the choice shuffling evaluation, question rephrasing exhibited lower Consistency for both LLaVA and MM-CoT. This might be due to the challenges posed by comprehending rephrased questions in various forms. It can be seen that Consistency is not related to the number of choices. In particular, the lowest Consistency is the 3-choices question, while 2-choices have the highest Consistency for both the LLaVA and MM-CoT models.

Table 2 (right) illustrates the Consistency comparison between questions with images and text-only questions within the ScienceQA dataset [14]. When evaluated with choice shuffling, questions that included images showed lower Consistency in comparison to those that were text-only. Although the Consistency for questions with images remained unchanged during the question rephrasing evaluation, there was a slight decrease in Consistency for text-only questions.

Comparison with Accuracy Table 2 (left) highlights the accuracy of both inconsistent and consistent samples using two evaluation methods. While consistent samples have a high accuracy of around 95% for both models, inconsistent samples still have an accuracy of around 50%. It can be seen that the model somehow can predict the correct answer by chance; however, the accuracy of inconsistent samples still contributes significantly to the overall accuracy across all choice variations and the overall accuracy across rephrased questions.

Figure 3 shows the relationship between the Consistency and Accuracy of each question in the dataset by using the two proposed approaches. The plot re-

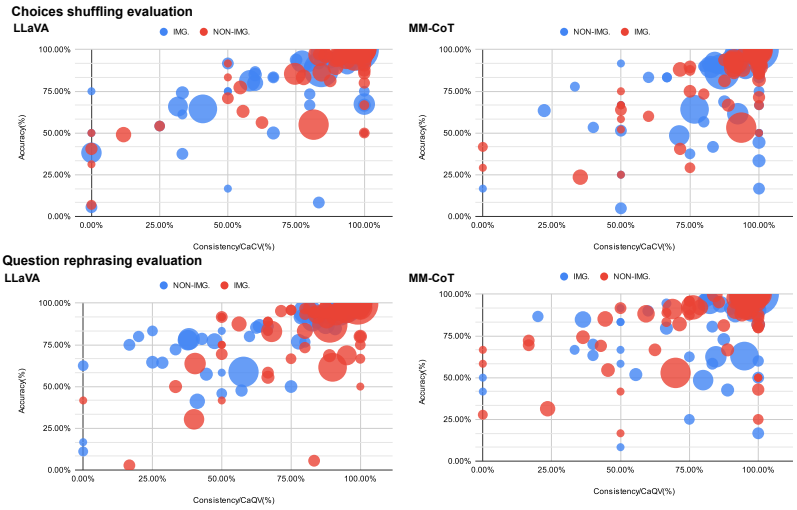

Fig. 3: Scatter plot of Consistency and Accuracy for each question in the ScienceQA dataset (circle sizes represents the number of question examples).

veals that there are many cases where Consistency and Accuracy disagree. With the high-accuracy and low-consistency cases, the VQA model does not perform well and might give the correct answer by chance. On the other hand, with low-accuracy and high-consistency cases, the VQA model does not fully comprehend the question and predicts incorrect answers. Thus, the proposed Consistency metrics effectively complements Accuracy, providing a deeper understanding of the VQA model's behaviors.

We also measure the impact of Consistency on Accuracy in Table 4 in 4 cases: original choices, all choices variations, best cases, and worst cases. In the case of original choices, we assess the accuracy of VQA models exclusively on the original choice provided by the ScienceQA dataset, whereas, in the all-choice variations scenario, the accuracy is evaluated across all possible choice variations. For the best cases, a final answer is deemed correct if any of the choice variations are correct. In contrast, in the worst cases, a final answer is considered incorrect if any of the choice variations is incorrect. In all choice variations, accuracy remains consistent with the original choice; however, in the best-case/worst-case scenarios, there is a notable increase/decrease in this figure.

Choice Shuffling Evaluation We found that the majority of inconsistent samples in the choice shuffling evaluation were questions containing images, accounting for 67.32% and 64.37% for LLaVA and MM-CoT, respectively. Table 2 (right) also highlights that the Consistency of questions with images is also lower than text-only questions with choice shuffling evaluation. We also illustrate the top-5 questions having the most inconsistent samples through choice shuffling evaluation as presented in Table 3. We also group all questions that require understanding map-image in 'geography' topic into 'Question related to map'. By analyzing the inconsistent samples, we notice that there are some characteristics of the question that can affect the Consistency as follows:

Table 3: Top-5 question with most inconsistent samples of both LLaVA [12] and MM-CoT [25] models with choices shuffling and question rephrasing approaches. (Has img: whether the question has an image, #Total: the number of samples with the question listed in the dataset, #Inc: number of inconsistent samples predicted by VQA models, Acc.: average accuracy across choice variations/rephrased question with choices shuffling/rephrasing approach.)

Question	Num. of choices	Has image	#Total	LLaVA		MM-CoT	
				#Inc.	Acc.	#Inc.	Acc.
Top-5 questions with the most inconsistent samples with *choices shuffling approach*							
Think about the magnetic force between the magnets in each pair. Which of the following statements is true?	3	✓	120	71	64.44	28	64.16
Question related to map.	4	✓	266	46	89.09	38	87.29
Which solution has a higher concentration of {} particles?	3	✓	45	45	38.15	13	48.53
Which property do these three/four objects have in common?	3	✓	70	29	80.95	11	91.66
Use guide words skill question.	2		140	26	55.00	9	50.71
Top-5 questions with most inconsistent sample with *question rephrasing approach*							
Use guide words skill question.	2		140	59	58.81	42	52.86
Suppose {} decides to {}. Which result would be a cost?	2		45	28	77.78	14	91.11
Will these magnets attract or repel each other?	2	✓	52	31	63.78	8	62.50
Question related to map.	4	✓	266	28	88.29	19	97.43
Which solution has a higher concentration of {} particles?	3	✓	45	27	30.37	9	48.52

– *Question requires fine-grained instance-level comprehension of the image* We observed that many inconsistent samples are related to questions needing a detailed understanding of the provided image. For example, with the question 'Think about the magnetic force between the magnets in each pair. ...' (First row of Table 3, Figure 4.(1)) requires VQA models to identify two pairs of magnets in the image, as well as the relative size, orientation of each magnet, or distance between magnets in each pair. In general, this demands that the model have particular capabilities such as text recognition, instance identification, and instance interaction. With this type of question, both LLaVA and MM-CoT show remarkably poor Consistency and Accuracy. In addition, these models also output inconsistent rationales about understanding the image. With questions related to map (Second row of Table 3), we noticed that inconsistent samples require fine-grained understanding, as shown in Figure 4.(4).
– *Require specific knowledge/logical reasoning* With questions require logical reasoning such as using guide words skill question (Figure 4.(3)), which requires comparing words in alphabetical order. The words given in these questions in the test set are completely different from the training set. Thus, these questions show low Consistency and Accuracy, and the VQA's reasoning for these questions is entirely incorrect. With compare properties of objects question (Forth row of Table 3), although having higher accuracy than other inconsistent questions, the attributes of each object in the VQA's rationales remain incorrect, even inconsistent samples. (Figure 4.(2)).

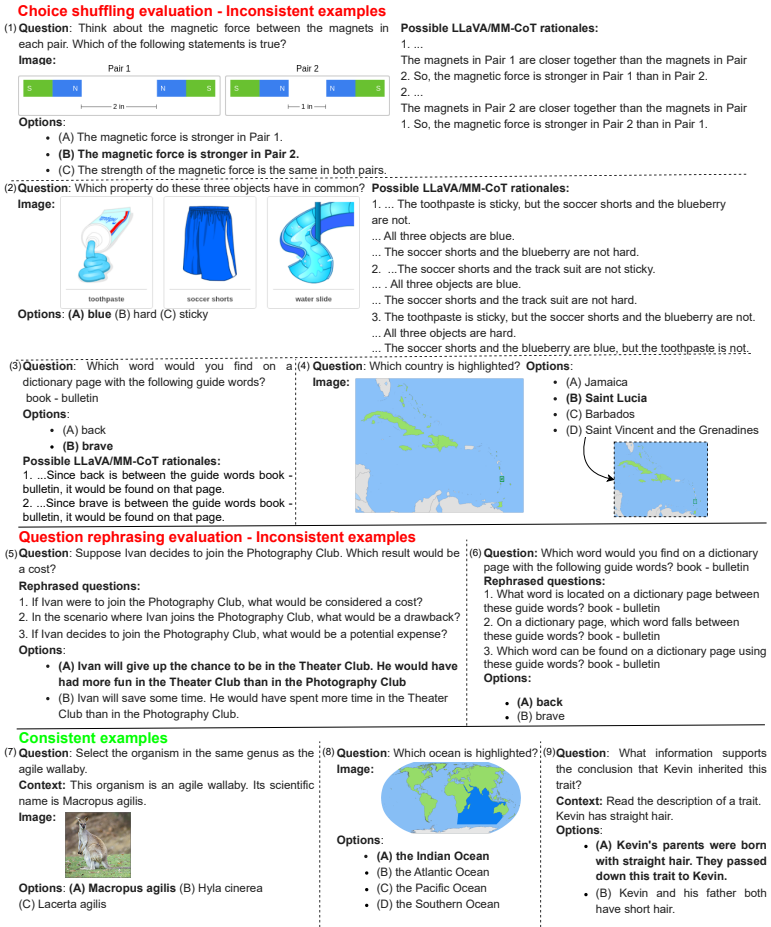

Fig. 4: Illustrate the rationales and answers predicted by the VQA models with our two proposed evaluations.

– *Missing image (data issue)* We noticed that some questions require understanding the provided image; However, no image was supplied. Consequently, this question exhibited poor Consistency and Accuracy.

Question Rephrasing Evaluation Compared to the choice-shuffling evaluation, the question rephrasing evaluation shows the same analysis with inconsistent samples as described above. Along with the reduction in the Consistency of text-only questions, as shown in Table 2 (right), there are also new text-only inconsistent samples compared to the choice shuffling evaluation. Most of the new text-only inconsistent samples require the logical reasoning of the model.

Consistent Samples' Characteristic By analyzing the remaining consistent samples of both the choices shuffling and question rephrasing evaluations, we found that the majority of these are questions that require an understanding of the provided text to deduce the answer from this understanding Figure 4.(9). Some of these samples include images, which serve merely as illustrations and

Table 4: Comparison of Overall Accuracy(%) of LLaVA [12] and MM-CoT [25] with different scenerios.

Methods	Original choices (Baseline)	All choices variations	Best cases	Worst cases
LLaVA [12]	91.60%	91.53% (-0.07%)	96.17% (+4.57%)	85.91% (-5.63%)
MM-CoT [25]	90.91%	90.96% (+0.05%)	93.58% (+2.67%)	88.27% (-2.64%)

are not essential for answering the question (Figure 4.(7)). Questions related to maps, which necessitate a global understanding of the image, also exhibit high consistency Figure 4.(8). Although there are some low-consistency samples that require detailed discrimination of features within the image, the other still shows good Consistency and Accuracy.

5 Conclusion

In this paper, we introduce two new kinds of evaluation, namely choice shuffling and question rephrasing, in order to understand the behavior of VQA models in the scientific domain. Along with two new approaches, we also proposed two metrics, which are Consistency across all Choice Variations (CaCV) and Consistency across Question Variations (CaQV), to evaluate the consistency of the VQA models. We show that depending solely on Accuracy metrics to assess VQA models is inadequate, and combining both Accuracy and Consistency metrics provides a more comprehensive understanding of VQA models. Our experimental results on the ScienceQA dataset [14] show that current VQA models might have inconsistent results with the same question-image pairs regardless of the accuracy. By understanding and objectively evaluating LLaVA [12] and MM-CoT [25], we have observed the following findings.

1. Number of choices is not correlated with Consistency. Table 1 illustrates that increasing the number of choices does not decrease the consistency. In particular, the 3-choice questions exhibit the lowest Consistency among the cases with three different choices. When compared to 3-choice questions, the 4-choice questions demonstrate higher Consistency. Instead, Consistency is more related to the question's content.
2. In both kinds of evaluations, LLaVA and MM-CoT show **poor consistency** for the question that requires fine-grained instance-level comprehension of the image, specific knowledge/logical reasoning, or data issues.
3. In contrast, LLaVA and MM-CoT show **good consistency** for questions that require an understanding of the provided text to provide the answer or a global understanding of the given image.
4. Despite having the same accuracy, MM-CoT demonstrates higher consistency than LLaVA.

These findings aim to offer valuable insights that can inspire future research on VQA models within the scientific domain. Future endeavors may explore the integration of stronger image representations into MLLMs to further enhance performance.

Acknowledgments. The first author of this paper is supported by the NII International Internship Program and stayed in NII from March to September 2024.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
3. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024)
4. Ho, N., Schmid, L., Yun, S.Y.: Large language models are reasoning teachers. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) ACL. pp. 14852–14882 (Jul 2023)
5. Hsieh, C.Y., Li, C.L., Yeh, C.k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) ACL. pp. 8003–8017 (Jul 2023)
6. Kafle, K., Price, B., Cohen, S., Kanan, C.: Dvqa: Understanding data visualizations via question answering. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5648–5656 (2018)
7. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV. pp. 235–251. Springer International Publishing (2016)
8. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5376–5384 (2017)
9. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. NeurIPS **35**, 22199–22213 (2022)
10. Krishnamurthy, J., Tafjord, O., Kembhavi, A.: Semantic parsing to probabilistic programs for situated question answering. In: Su, J., Duh, K., Carreras, X. (eds.) EMNLP. pp. 160–170 (Nov 2016)
11. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 (2023)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
13. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024)
14. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
15. Magister, L.C., Mallinson, J., Adamek, J., Malmi, E., Severyn, A.: Teaching small language models to reason. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 1773–1781. ACL (Jul 2023)

16. OpenAI: GPT-4 technical report. CoRR **abs/2303.08774** (2023)
17. Sampat, S.K., Yang, Y., Baral, C.: Visuo-linguistic question answering (VLQA) challenge. In: Cohn, T., He, Y., Liu, Y. (eds.) EMNLP. pp. 4606–4616 (Nov 2020)
18. Sampat, S.K., Yang, Y., Baral, C.: Visuo-linguistic question answering (VLQA) challenge. In: Cohn, T., He, Y., Liu, Y. (eds.) EMNLP. pp. 4606–4616 (Nov 2020)
19. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
20. Wang, L., Hu, Y., He, J., Xu, X., Liu, N., Liu, H., Shen, H.: T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. AAAI **38**, 19162–19170 (03 2024)
21. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS **35**, 24824–24837 (2022)
22. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265 (2023)
23. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. Advances in Neural Information Processing Systems **36** (2024)
24. Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., Qiao, Y.: Llama-adapter: Efficient fine-tuning of language models with zero-init attention. arXiv preprint arXiv:2303.16199 (2023)
25. Zhang, Z., Zhang, A., Li, M., hai zhao, Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research (2024)