What If the TV Was Off? Examining Counterfactual Reasoning Abilities of Multi-modal Language Models

Letian Zhang[1] Xiaotong Zhai[2] Zhongkai Zhao[3] Xin Wen[3] Bingchen Zhao[3]
[1]Tongji University [2]University of Warwick [3]LunarAI

Abstract

Counterfactual reasoning ability is one of the core abilities of human intelligence. This reasoning process involves the processing of alternatives to observed states or past events, and this process can improve our ability for planning and decision-making. In this work, we focus on benchmarking the counterfactual reasoning ability of multimodal large language models. We take the question and answer pairs from the VQAv2 dataset and add one counterfactual presupposition to the questions, with the answer being modified accordingly. After generating counterfactual questions and answers using ChatGPT, we manually examine all generated questions and answers to ensure correctness. This results in over 2k counterfactual question and answer pairs. We evaluate recent vision language models on our newly collected test dataset and found that all models exhibit a large performance drop compared to the results tested on questions without counterfactual presupposition. This result indicates that there still exists space for developing vision language models. We hope our proposed benchmark can help the development of future systems.

1. Introduction

The recent development of multi-modal large language models (MLLMs) [15] greatly improves the performance of complex reasoning on images, image-based dialogue, and language grounding tasks [8, 6, 13, 14, 15]. The current reasoning benchmarks used by these MLLMs usually only require the model to understand the image contents by grounding the concept involved in the language questions to the image content. Because the visual representation learned by CLIP [20] can already represent the image contexts in an aligned space with language, we argue that the current reasoning benchmarks are relatively easy to handle for MLLMs as most of the reasoning questions only need to readout the content of the image. In this work, we aim to propose a harder evaluation scenario where the language model needs to query for the correct vision representation of image content and be able to perform counterfactual operations on those representations.

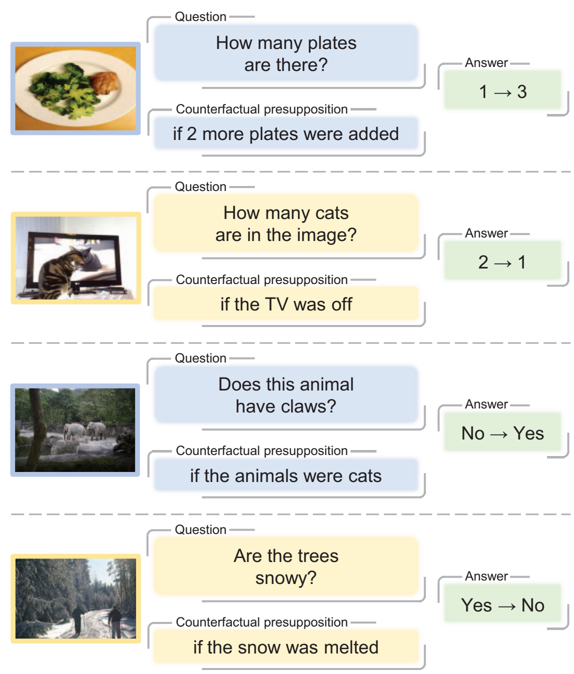


Figure 1. Counterfactual modifications to VQA-v2 questions.

In Figure 1, we present the counterfactual visual question answering. We start with the VQAv2 dataset [7] and add one counterfactual presupposition to the questions. The language model needs to understand the counterfactual presupposition and then adjust its inner representations of the image contents accordingly to be able to correctly answer the questions. For example, in Figure 1, to answer the question of "How many cats are in the image", the model can query and count the occurrence of the visual concept "cat" in the image. However, with the counterfactual presupposition "if the TV was off", one of the cats on the screen will not be visible. Thus this counterfactual presupposition poses a challenging scenario for VQA models. We collect over 2k of these kinds of images and counterfactual questions, each of the images and questions is examined by two people to ensure the counterfactual presupposition is related to the image content and the answers generated are accurate.

Our initial experiment of evaluating current state-of-the-art vision-language models on our dataset shows that all models have a performance drop when evaluated on these counterfactual questions, indicating further room for improvement.

2. Related Works

Visual Question Answering Visual question answering (VQA) aims to enable machines to have the ability of visual understanding, language generation, and commonsense knowledge. Several datasets have been proposed for VQA [2, 7, 11, 16, 21, 12, 3]. COCO-QA [21] and VQA [2] are the first works to propose the task of VQA, the proposed datasets contain enormous pictures and questions covering daily life objects.VQAv2 [7] builds on the VQA dataset. By identifying biases and shortcuts in the VQA dataset, VQAv2 presents a more rigorous evaluation for the VQA task. Further works [16, 12, 11] continue to extend VQA evaluation to different aspects of image understanding, OK-VQA [16] evaluates the ability to draw upon external knowledge for visual question answering. CLEVR dataset [12] devise a pipeline for generating synthetic data for evaluating the compositional reasoning ability of VQA models. GQA dataset [11] leverages the scene graph structure to generate reasoning questions on real-world images to test the compositional reasoning ability. Recent SMART-101 dataset[3] replicates 101 abstract root puzzles into over diverse instances which evaluate the arithmetic and reasoning abilities. In this work, our proposed dataset is built on the real-world images of VQAv2 [7], where the questions and answers are modified by counterfactual presupposition. These counterfactual presuppositions provide a new challenging scenario for VQA models, as they will require the model to parse the scene structure and reason about the observed world after the counterfactual presuppositions.

Reasoning Ability Benchmarks The performance of generative models is hard to evaluate, let alone large language models. There exist several efforts for the evaluation of the reasoning ability of LLMs. GSM8k [4] evaluates this by using a set of grad school math word problems, MMLU [9] extends on this by covering more topics and evaluates the performance via multiple choice questions, later the MATH [10] benchmark evaluates the reasoning of LLMs using text generation instead of multiple choices. IfQA [24] proposes the first question-answering dataset that requires the model to do counterfactual reasoning. In [23], the counterfactual reasoning ability of recent strong LLMs (GPT-4 [18], Claude [1]) is evaluated under eleven different tasks with counterfactual presuppositions. It is shown that current LLMs cannot reason with counterfactuals reliably. [17] combines the idea of counterfactual reasoning with computer vision. It proposes a novel counterfactual inference framework to detect the causal effect of questions on answers. In our work, we propose to build the first visual question-answering dataset with counterfactual presuppositions. Our evaluation results show current approaches for MLLMs do not facilitate counterfactual reasoning. Thus, further work should be done to create stronger MLLMs.

3. Dataset

This section presents the construction process of our proposed dataset C-VQA. C-VQA contains 2,217 image and question-answer pairs, where each questions are not only related to the image content, but also comes with a counterfactual presupposition. These counterfactual presuppositions enable a new and more challenging scenario for VQA models. We will introduce the image selection, counterfactual presupposition generation, and verification steps below.

3.1. Data Selection

In this initial work, we choose to collect numerical and boolean questions as they are easy to evaluate, and adding counterfactual presuppositions can easily change the answer to the counterfactual questions. ViperGPT [6] is used to further filter the dataset. We remove examples in which ViperGPT fails to give a correct answer to the original question in VQAv2. The rationale is that, by removing those examples, we keep only the examples that can be correctly processed by both the visual perception module and the reasoning module of ViperGPT. Thus we can be sure of the error source when adding counterfactual presuppositions to the questions: it only comes from the lack of ability to reason counterfactually.

3.2. Annotation

Counterfactual presupposition type. When designing counterfactual questions, we apply different types of counterfactual presuppositions. For numerical questions, we split them into two groups: direct group and indirect group. In the direct group, we add counterfactual presuppositions that require LLMs to perform one or two arithmetical operations to get correct answers. These questions typically have the form *"How many X would there be if two more X were added?"*. In the indirect group, the counterfactual presuppositions change the original answers indirectly. It requires more reasoning steps to get the new answers. For example, answering the sentence *"Would this animal have claws if the animals were cats?"* requires the model to regard the animals as cats and be aware that cats have claws. For boolean questions, the counterfactual presuppositions are often designed to reverse the fact as well as the answer. For example, *"Would the cat be asleep if it was woken up?"*. Examples of these counterfactual modifications are presented in Figure 1.

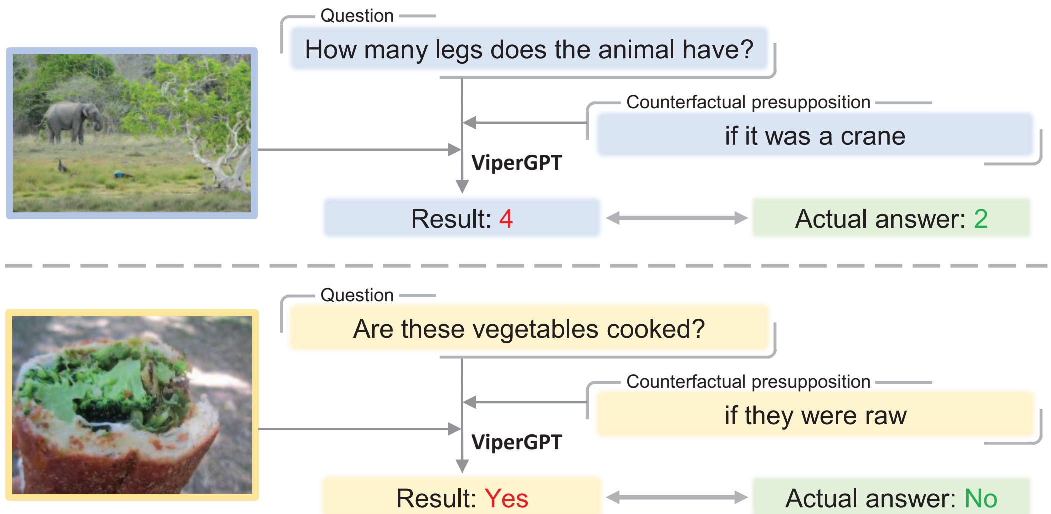


Figure 2. Two examples of wrongly answered questions. When counterfactual presuppositions are added, ViperGPT gives wrong answers to the questions that it can answer correctly without the counterfactual parts.

Question and answer annotation. A two-stage annotate-and-prompt process is employed to create C-VQA. First, we manually annotate 200 questions and answers for each group. Then, we prompt ChatGPT with the manually annotated questions and answers to generate new counterfactual modified questions and answers for the rest of the examples. We use chain-of-thought prompting [22] and in-context learning [19] to promote the quality of the questions produced by ChatGPT and keep the accuracy of new answers. Specifically, we divide the whole task into several smaller tasks: read the original question, figure out a proper counterfactual presupposition, figure out how the answer will change, and write down the new question and answer. The in-context examples will further help ensure the sentence voice is correct and show different counterfactual presuppositions. For numerical groups, we first make ChatGPT produce a counterfactual presupposition and then generate a new question with it, and the new answer is also calculated. However, this strategy cannot be directly applied to boolean questions. We notice the randomly generated counterfactual presuppositions often fail to flip the original answer, and thus a large proportion of answers keeps unchanged or indeterminate. Therefore, we design a new prompting strategy flipping the original answer first and then generating a corresponding counterfactual presupposition.

3.3. Verification

Though ChatGPT does not tend to make mistakes such as spelling mistakes and grammatical mistakes that are common in human writing, it introduces errors such as calculation mistakes. Moreover, each question is paired with a corresponding image in the original dataset, while ChatGPT cannot read the image. Therefore, although the counterfactual questions generated by ChatGPT are idiomatic, they may be improper in the context of the scene in the image. We further verify the questions and answers manually to address the errors caused by ChatGPT. Our verification mainly consists of two stages: (i) image-related verification and (ii) answer-reasonability verification. And each image is at least verified by two people to ensure the correctness.

(i) Whether the new question is image-related? To make sure that the generated question is indeed asking about the model to reason with the scene in the image, we manually examined all generated counterfactual questions and removed all questions that were modified to be not related to the image content.

(ii) Whether the new answer is reasonable? Automatically generated answers may be wrong since ChatGPT may make calculation mistakes, reasoning mistakes, etc. We correct these answers manually. Furthermore, some questions may be ambiguous, so there are no deterministic answers. We remove these questions from the dataset.

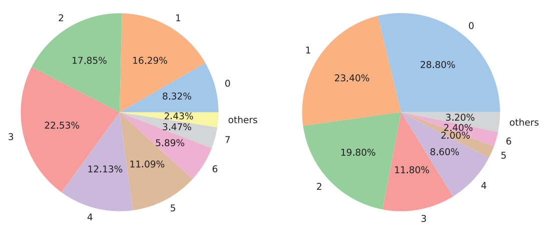
Since ChatGPT cannot access the images, it does not know anything about the position or color of the objects. Thus, ChatGPT tends to add counterfactual presuppositions that remove all objects rather than just manipulate a few of the objects. To ensure a rigorous evaluation, we manually annotate all numerical indirect questions. After completing all the annotations, we use ChatGPT again to inspect and polish all the questions in order to make sure the questions are grammatically correct.

3.4. Dataset Statistics

Question Type and Length. C-VQA contains 2217 questions in total, with 1077 numerical questions and 1140 boolean questions. In numerical questions, there are 577 questions from the direct group and 500 questions from

Model type	Model	Numerical direct		Numerical indirect		Boolean	
		original	counterfactual	original	counterfactual	original	counterfactual
Neuro-symbolic	ViperGPT	-	**80.6** (↓19.4)	-	31.6 (↓68.4)	-	27.6 (↓72.4)
	VisProg	40.0	43.0 (↑3.0)	40.8	18.6 (↓22.2)	73.6	31.1 (↓42.5)
Neuro	InstructBLIP (FlanT5$_{\text{XXL}}$)	47.3	32.9 (↓14.4)	51.6	**32.0** (↓19.6)	66.1	48.4 (↓17.7)
	InstructBLIP (Vicuna-7B)	**66.0**	40.0 (↓26.0)	**65.2**	31.4 (↓33.8)	**84.9**	55.4 (↓29.5)
	InstructBLIP (Vicuna-13B)	59.6	46.4 (↓13.2)	60.2	31.6 (↓28.6)	84.6	**68.8** (↓15.8)
	LLaVA-7B	36.9	27.0 (↓9.9)	40.2	25.0 (↓15.2)	63.3	58.5 (↓4.8)
	LLaVA-13B	36.7	24.8 (↓11.9)	42.0	20.8 (↓21.2)	70.0	56.3 (↓13.7)
	BLIP2 (FlanT5$_{\text{XXL}}$)	46.6	32.8 (↓13.8)	51.8	31.4 (↓20.4)	65.4	51.8 (↓13.6)

Table 1. The evaluation results of three groups in our dataset C-VQA by ViperGPT, VisProg, InstructBLIP, LLaVA and BLIP2. As mentioned in section 3.1, all original questions can be correctly answered by ViperGPT.



(a) direct group (b) indirect group

Figure 3. **Breakdown of answers in numerical groups.** We show the percentage of answers in the numerical direct group and numerical indirect group. The share of 0, 1, and 2 in the indirect group are higher while the others are lower.

the indirect group. All the numerical questions are *"How many"* questions. And most of the questions in the boolean group start with "Is" or "Are" before they are changed into counterfactual ones. The average length of questions in C-VQA is 13.12 words, much longer than that of original questions (5.75 words).

Answer Statistics. Each answer in the numerical type is an exact number, with no ambiguous answers such as "a lot" or "many". The distribution of the answers is shown in Figure 3. Each answer in the boolean type is a single "yes" or a single "no". The percentage of "no" is 65.70%, while "yes" is 34.30%.

4. Experiments

We evaluate both neuro-symbolic models and neuro models on C-VQA. Specifically, we test ViperGPT [6], VisProg [8], InstructBLIP [5], LLaVA [15] and BLIP2 [14].

4.1. Qualitative Results

Figure 2 provides two examples of wrongly answered questions. In both examples, ViperGPT gives wrong answers to the counterfactual questions. We further inspect the codes generated by ViperGPT when solving these two questions. The first example causes ViperGPT to generate long codes with many redundancy steps leading to a wrong result, and in the second example, the generated code ignores the counterfactual presupposition. The code generated by ViperGPT will be put into the supplementary.

4.2. Quantitative Results

As shown in Tab. 1, both neuro-symbolic and neuro models perform significantly worse with counterfactual questions than the original ones in all three groups. It should be noted that reasoning difficulty impacts the results significantly. For the numerical direct group, models require only one simple reasoning step to get the answer. As a consequence, the difference between the evaluation accuracy of the original and counterfactual questions is smaller compared to other groups. In the evaluation of ViperGPT, the accuracy drops 19.4% when one-step reasoning is added, and it drops about 70% when multi-step reasoning is added. Obviously, complex reasoning challenge models greatly. In most cases, ViperGPT fails to consider complex counterfactual presuppositions or reasons improperly. We notice that InstructBLIP-Vicuna shows overall improvements in counterfactual questions when the model scale turns bigger. However, there is an opposite case that LLaVA performs worse in counterfactual questions when its scale is bigger. For neural models, InstructBLIP-Vicuna outperforms other models in most question types. These results indicate that C-VQA is challenging and further explorations on C-VQA can lead to the future development of MLLMs.

5. Conclusion

In this paper, we introduce a novel benchmark C-VQA with over 2k images and questions. Each question in C-VQA is based on a counterfactual presupposition and contains an "if" clause. We test both neuro-symbolic and neuro models and show that the performance of state-of-the-art VQA models drops significantly on C-VQA. The evaluations reveal that C-VQA is highly challenging and that the reasoning ability of existing multi-modal large language models has a large room for improvement.

References

[1] Anthropic. Claude. *https://www.anthropic.com/index/introducing-claude*, 2023. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2

[3] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A Smith, and Joshua B Tenenbaum. Are deep neural networks smarter than second graders? In *CVPR*, 2023. 2

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021. 2

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023. 4

[6] Surís Dídac, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 1, 2, 4

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 1, 2

[8] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2017. 1, 4

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 2

[10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021. 2

[11] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 2

[13] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. 2023. 1

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 4

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023. 1, 4

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

[17] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 2021. 2

[18] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 2

[19] Dong Qingxiu, Li Lei, Dai Damai, Zheng Ce, Wu Zhiyong, Chang Baobao, Sun Xu, Xu Jingjing, Li Lei, and Sui Zhifang. A survey on in-context learning. *arXiv:2304.08485*, 2023. 3

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[21] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NeurIPS*, 2015. 2

[22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 3

[23] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv:2307.02477*, 2023. 2

[24] Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv:2305.14010*, 2023. 2