

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

Letian Zhang¹ Xiaotong Zhai² Zhongkai Zhao^{3,6} Yongshuo Zong⁴ Xin Wen^{5,6} Bingchen Zhao^{4,6}

¹Tongji University ²University of Warwick ³National University of Singapore

⁴University of Edinburgh ⁵The University of Hong Kong ⁶LunarAI

Abstract

Counterfactual reasoning, a fundamental aspect of human cognition, involves contemplating alternatives to established facts or past events, significantly enhancing our abilities in planning and decision-making. In light of the advancements in current multi-modal large language models, we explore their effectiveness in counterfactual reasoning. To facilitate this investigation, we introduce a novel dataset, C-VQA, specifically designed to examine the counterfactual reasoning capabilities of modern multi-modal large language models. This dataset is constructed by infusing original questions with counterfactual presuppositions, spanning various types such as numerical and boolean queries. It encompasses a mix of real and synthetic data, representing a wide range of difficulty levels. Our thorough evaluations of contemporary vision-language models using this dataset have revealed substantial performance drops, with some models showing up to a 40% decrease, highlighting a significant gap between current models and human-like vision reasoning capabilities. We hope our dataset will serve as a vital benchmark for evaluating the counterfactual reasoning capabilities of models. Code and dataset are publicly available at <https://bzhao.me/C-VQA/>.

1. Introduction

“Counterfactuals are the building blocks of moral behavior as well as scientific thought.”

— Judea Pearl, *The Book of Why*

Counterfactual ability is a pivotal cognitive function in humans, enabling us to envision alternate realities and outcomes based on different choices or events. This capacity underpins our decision-making, moral reasoning, and problem-solving skills. Recent development of multi-modal large language models (MLLMs) [26, 42, 45, 48] have dramatically improved the capabilities of image recognition, image-based dialogue, and language grounding, *etc.* [4, 9,

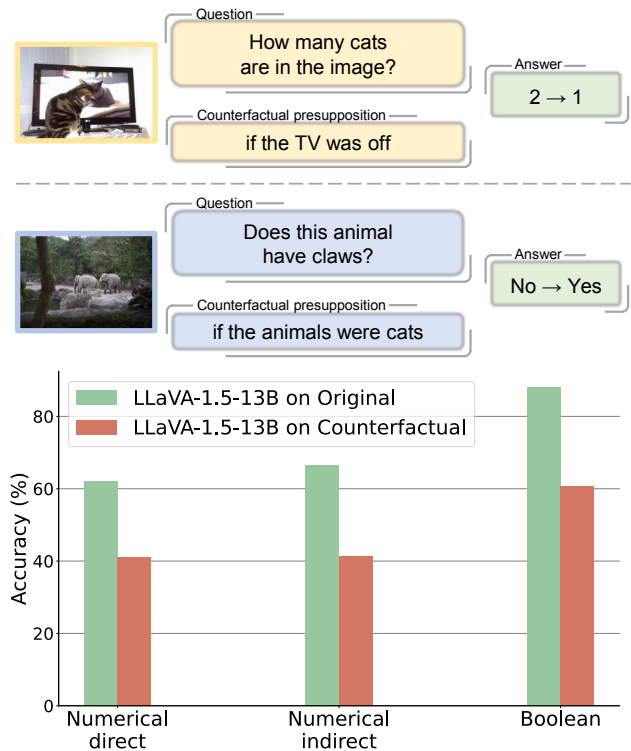


Figure 1. Examples of C-VQA (top), and performance comparison of LLaVA-1.5 [25] w/ and w/o counterfactuality (bottom). C-VQA is constructed by adding counterfactual presuppositions to the questions. We observe that state-of-the-art models all exhibited significant performance drops on the counterfactual questions.

11, 12, 14, 20, 24, 26, 36]. These developments raise the possibility of achieving higher levels of artificial intelligence. Consequently, we pose a critical question: *are contemporary MLLMs equipped for counterfactual reasoning?*

Visual question answering is one of the central tasks to evaluate the ability of vision-language models. Current benchmarks focus on evaluating different aspects of the abilities, such as visual recognition (*e.g.*, VQAv2 [13]), external knowledge (*e.g.*, OKVQA [29]), compositional reasoning (*e.g.*, CLEVR [19]). They only require the model to understand the image contents by grounding the concept in-

volved in the language questions to the image content. With the help of large language models, these tasks are relatively easy to solve, as evidenced by the high scores achieved by modern MLLMs [5, 8, 26]. The most relevant benchmarks are CF-VQA [30] and VQA-CP [1]. However, they focus on constructing the counterfactual questions between the training set and the test set, *e.g.*, showing a green banana in the test set while most of the bananas in the training set are yellow. They do not evaluate the counterfactual ability of the MLLMs - *the skill to imagine “what if” scenarios that differ from what actually happened.*

In light of this, we propose a novel and challenging evaluation scenario where the language model not only needs to query for the correct visual representation of image content but also needs to be able to perform counterfactual reasoning on those representations. This cannot be solved by encoding world knowledge or object spatial relation alone, but also requires the model to further understand and imagine the given scenarios. Specifically, in our proposed benchmark, each visual question is modified by a counterfactual presupposition - making an assumption based on a scenario that did not actually happen, but could have. This requires the MLLMs to understand the visual content and then perform reasoning over the image contents accordingly to answer the questions. For example, in Fig. 1, to answer the question of “How many cats are in the image?”, the model can directly query the image and count the occurrence of the visual concept “cat”. However, with the counterfactual presupposition “if the TV was off”, one of the cats on the screen will not be visible. Thus, it poses a more challenging scenario for the models as the model not only needs to identify the occurrence of the concept “cat”, but also reason about its state after a counterfactual operation is taken.

We build our dataset on top of the commonly used VQAv2 dataset [3], where we collect over 3,000 images suitable for asking counterfactual questions. In addition, we also created over 3,000 synthetic images and corresponding questions to further test the out-of-distribution ability of multi-modal large language models.

We evaluate a wide array of contemporary state-of-the-art multi-modal language models on our dataset, such as LLaVA [26], MiniGPT4 [5, 47], BLIP2 [24], Instruct-BLIP [8], Qwen-VL [4] and CogVLM [37]. Our experiments show several interesting findings: (1) Neuro-symbolic models perform worse than end-to-end models on complex counterfactual reasoning; (2) No model family can consistently address counterfactual questions. All of the models suffer from a large performance drop from our counterfactual questions; (3) Even the strongest GPT-4V model [32] cannot solve our benchmark; (4) The MLLMs also demonstrate a systematic bias in answering gender-related counterfactual questions.

We summarize the main contributions of this paper:

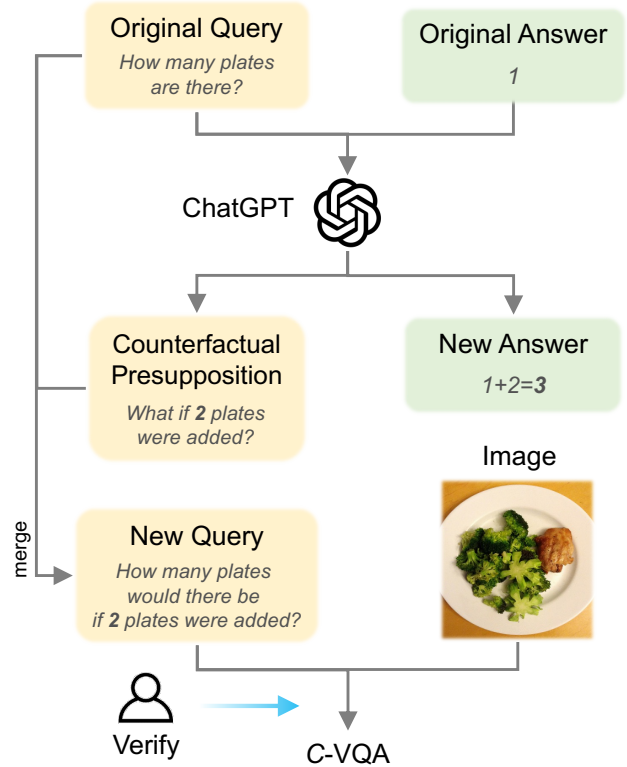


Figure 2. **Our annotation flow for C-VQA-Real.** We select images and questions from the VQAv2 dataset [3], and then utilize ChatGPT to add counterfactual presupposition to the questions and get the corresponding answers. All questions and answers are carefully inspected by human annotators.

1. We propose a novel and challenging dataset *C-VQA* with both real and synthetic image-question pairs. The questions contain counterfactual presuppositions and are of various difficulties.
2. We conduct extensive evaluation of current state-of-the-art vision-language models on our dataset. We show that they struggle with counterfactual reasoning.
3. We summarize our findings from the experiments on proposed *C-VQA*, hoping to provide valuable insights for future research in this field.

2. Related Works

Visual Question Answering. Visual question answering (VQA) aims to evaluate machines’ capabilities of visual understanding, visual reasoning, and the application of commonsense knowledge. Several datasets have been proposed for VQA [3, 6, 13, 18, 19, 23, 29, 34, 39]. COCOQA [34] and VQA [3] are the first to propose the task of VQA, and they contain enormous pictures and questions covering daily life objects. Builds on the VQA dataset. VQAv2 [13] presents a more rigorous evaluation for the VQA task by mitigating language biases and shortcuts in

the VQA dataset. Further works [18, 19, 29] continue to extend VQA evaluation to different aspects of image understanding. CLEVR dataset [19] devises a pipeline for generating synthetic data for evaluating the compositional reasoning ability of VQA models. GQA dataset [18] leverages the scene graph structure to generate reasoning questions on real-world images to test the compositional reasoning ability. In this work, we build our dataset *C-VQA* on the top of VQAv2 [13], where the questions and answers are modified by counterfactual presupposition. Different from previous VQA tasks, our proposed counterfactual dataset provides a new challenging scenario for complementary MLLMs. It further examines the models’ abilities to parse the scene structure and reason about the observed world after the counterfactual presuppositions.

Evaluation of Reasoning Abilities. The evaluation of generative models has always been a difficult and challenging problem, as well as for large language models. There have been many efforts for the evaluation of the reasoning ability of LLMs in different aspects [1, 7, 16, 17, 30, 40, 43]. For example, GSM8k [7] and MATH [17] evaluate the mathematical reasoning of LLMs, while MMLU [16] aggregates a diverse range of subjects and tasks for evaluation. For counterfactual reasoning, IfQA [43] proposes the first QA dataset that specifically designed to assess the counterfactual reasoning capabilities of language models. In [40], the counterfactual reasoning abilities of state-of-the-art LLMs (GPT-4 [31], Claude [2]) are evaluated under eleven different tasks with counterfactual presuppositions. It is shown that current LLMs cannot reason with counterfactuals reliably. Most related to our work is CF-VQA [30] and VQA-CP [1] which focus on testing the ability of VQA models to answer “counterfactual” questions which are defined as having different property distribution in training and testing set, such as training with yellow bananas but test with green ones. In our work, we propose to build the first visual question-answering dataset with counterfactual presuppositions directly added to the question text, this enables the testing of current SOTA MLLMs which leverages internet-scale training data. Our evaluation results show current approaches for MLLMs do not facilitate counterfactual reasoning. Thus, further work should be done to create stronger MLLMs.

Multi-Modal LLM Benchmarks With the increasing interest in training novel multi-modal LLMs that can perceive multi-modal inputs, various benchmarks for evaluating the performance of MLLMs are proposed. MME [10] measures the performance of multi-modal LLMs using both perception and cognition abilities on 14 subtasks. MM-Bench [27] covers over 20 different ability dimensions. A robust evaluation method is also proposed that leverages ChatGPT to match the model prediction to given choices. SEED-

Bench [22] proposes a benchmark with over 19k questions, covering both image and video modality to evaluate the performance of multi-modal LLMs. MM-Vet [44] defines 16 emergent tasks of interest from 6 core vision-language capabilities, with an LLM-based evaluator, relative strengths and weaknesses of different system paradigms are identified. In this paper, we propose an orthogonal direction for evaluating the reasoning ability of current multi-modal LLMs: counterfactual reasoning.

3. Dataset

This section presents the construction process of our proposed dataset *C-VQA*, which consists of two parts: *C-VQA-Real* and *C-VQA-Synthetic*. *C-VQA-Real* contains 3,144 image and question-answer pairs, where each question is not only related to the image content but also comes with a counterfactual presupposition. All images and original questions of *C-VQA-Real* come from the VQAv2 dataset. *C-VQA-Synthetic* contains 3,000 image and question-answer pairs, and all images and questions in *C-VQA-Synthetic* are generated automatically. Each image corresponds to an original question and a counterfactual question. The counterfactual presuppositions in *C-VQA* enable a new and challenging scenario for VQA models. Both *C-VQA-Real* and *C-VQA-Synthetic* contain numerical and boolean questions that can be answered with a number or a boolean value respectively. We illustrate the counterfactual presupposition generation and verification steps below.

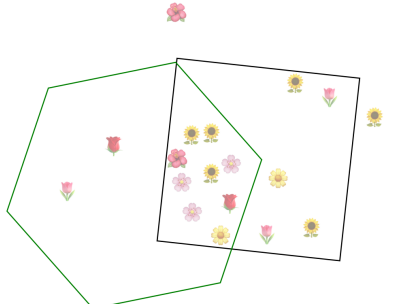
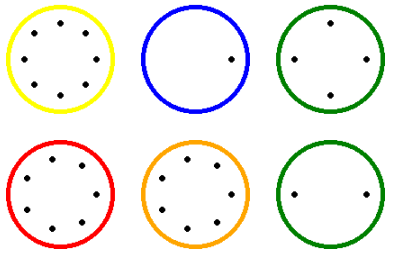
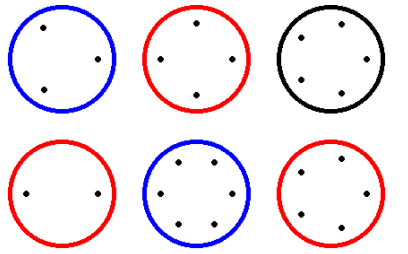
3.1. Annotation

Counterfactual presupposition type of *C-VQA-Real*. When designing counterfactual questions, we employ different types of counterfactual presuppositions as follows.

Numerical questions are split into two groups: direct group and indirect group. In the **numerical direct** group, we add counterfactual presuppositions that change the original answers directly. For example, these questions typically have the form “*How many X would there be if two X were added/removed?*”. In the **numerical 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 “*How many cats would be there if the TV was off?*” requires the model to recognize how many cats are images on the TV and understand that when the TV is off, those cats will no longer be there. 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 Fig. 1.

Question and answer annotation of *C-VQA-Real*. A two-stage annotate-and-prompt process is employed to cre-

Table 1. **Three instances in C-VQA-Synthetic.** They are made from different templates.

| | | |
|--|---|---|
|  |  |  |
| <p>How many flowers would be inside black polygons if all flowers in green polygons were removed? Select the correct answer: A:5 B:1 C:3 D:7</p> | <p>How many dots would there be in all the circles together if 24 dots were removed from the circles? Select the correct answer: A:7 B:11 C:3 D:9</p> | <p>How many dots would a circle contain at most if one of the circles with most dots were removed? Select the correct answer: A:5 B:1 C:3 D:7</p> |

ate *C-VQA*. First, we manually annotate 200 questions and answers for each of groups we have defined above. Then, these manually annotated questions and answers are used as in-context-examples [33] to prompt ChatGPT [31] to generate new counterfactual modified questions for the remaining samples. Additionally, to maximize the correctness of ChatGPT-generated questions, we leverage chain-of-thought [38] and 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 output the new question and answer. We also provide in-context examples to further ensure the sentence structure and to illustrate various counterfactual presuppositions. For numerical question groups, we first prompt ChatGPT to produce a counterfactual presupposition and then generate a new question with it, and the new answer is also calculated. The annotation flow is plotted in Fig. 2. 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 remain unchanged or even cannot be determined. Therefore, we employ a prompting strategy that flips the original answer first and then generates a corresponding counterfactual presupposition.

3.2. Verification

As we leverage ChatGPT in the process of generating the annotation of *C-VQA-Real*, we employ a rigorous verification process to ensure the correctness of the generated annotations in order to remove any hallucination or calculation mistakes¹. Therefore, we further verify the questions and answers manually to address the errors caused by ChatGPT. Therefore, although the counterfactual questions generated by ChatGPT are idiomatic, they may be improper in the

¹At the time of curating the dataset, GPT-4V API was not publicly available yet [32].

context of the scene in the image. Thus we further verify the questions and answers manually to address any potential errors. 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 correctness. We examine the following two key points:

(i) **Whether the new question is image-related?** To make sure that the generated question is indeed asking the model to reason with the scene in the image, we manually examine all generated counterfactual questions and remove 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, with no deterministic answers. We remove these questions from the dataset. Since ChatGPT is text-only and cannot see the images, it has to access to the position or color of the objects. Thus, it 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 again in order to make sure the questions are grammatically correct.

3.3. Implementation of C-VQA-Synthetic

In this section, we introduce the procedure we take to generate the *C-VQA-Synthetic* which contains synthetic images with counterfactual questions and answers. The advantage of *C-VQA-Synthetic* is that this enables the automatic generation of high-quality counterfactual image-related ques-

Table 2. **Six question templates of C-VQA-Synthetic.** Each template has 500 instances with different images and questions. {Color} and N are randomly generated for each instance.

| Type | Question Template |
|-----------------|---|
| Flower-counting | How many flowers would be outside the {Color} polygons if all polygons were {Color}? |
| | How many flowers would be inside {Color} polygons if we removed N flowers in {Color} polygons? |
| | How many flowers would be inside {Color ₁ } polygons if all flowers in {Color ₂ } polygons were removed? |
| Dot-counting | How many dots would there be in all the circles together if N dots were removed from the circles? |
| | How many dots would there be in the top three circles together if the two rightmost circles and dots in them were removed from the circles? |
| | How many dots would a circle contain at most if one of the circles with most dots were removed? |

tions, saving the time and labor cost of manually annotating. We designed six abstract question types in total and generated 500 instances for each type, resulting in a total of 3,000 synthetic images with counterfactual questions. All images are generated via a predefined procedure in [6] randomly, and we use templates to produce original and counterfactual questions. Both original and counterfactual questions are in a multiple-choice question format. Tab. 1 shows three instances of C-VQA-Synthetic. In the following, we detail the design of our C-VQA-Synthetic by explaining the process of generating two sets of synthetic images and questions: Flower-counting and Dot-counting.

Flower-Counting Puzzles. Three types of questions in C-VQA-Synthetic are based on flower-counting tasks. We first randomly sample two polygons S_1 and S_2 in different colors C_1 and C_2 . The two polygons divide the whole image into four parts: $in_1 \& in_2$, $in_1 \& out_2$, $out_1 \& in_2$ and $out_1 \& out_2$, where in_i means inside both S_i , and out_i means outside S_i . Then, we select several flower instances from the Icons-50 dataset [15] and insert them randomly into the four parts. The flowers are guaranteed not to intersect with the edge of polygons to ensure an accurate answer for reasoning questions. The number of flowers in part p is denoted as $num(p)$. With the images procedurally generated, we design questions related to the image contents by filling in question templates. The question templates are shown in Tab. 2. To answer those questions well, a multi-modal model not only needs to reason about the spatial relationship, but also needs to perform counterfactual reasoning. For example, the first question template would require the model to reason about the color and the spatial relationship of the image as well as the counterfactual change of color, and the third question template requires the model to reason about the intersection of regions and the counterfactual removal of items.

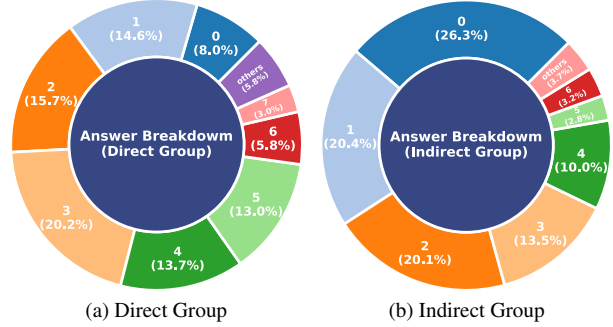


Figure 3. **Breakdown of answers in numerical groups of C-VQA-Real.** We show the percentage of answers in the numerical direct group and numerical indirect group. The share of 0, 1, and 2 in the indirect group are higher while the others are lower.

Dot-Counting Puzzles. The other three types of questions are based on dot-counting tasks. We arrange six circles in the images and then insert dots into these circles randomly. The questions are about the number of dots in the circles. For example, the first template is “How many dots are there in all the circles together?” for the original question and “How many dots would there be in all the circles together if N dots were removed from the circles?” for the counterfactual question. Suppose the original answer is ans_1 , we can calculate the counterfactual answer ans_2 by the equation $ans_2 = ans_1 - N$. For the third type of counterfactual question, “How many dots would a circle contain at most if one of the circles with most dots were removed?”, in order to fully evaluate the reasoning ability of models, the circle with the most dots is designed to be unique to ensure the counterfactual answer varies from the original one.

We present all six question templates in Tab. 2. All templates are “How many” questions, and we insert random elements in images and templates. The answers are designed to have a wide range so that we can rigorously test the counting and reasoning ability of MLLMs. Additionally, the domain gap between these synthetic images poses a challenge for the current multi-modal models as C-VQA-Synthetic is asking for both domain generalization and counterfactual reasoning, which we argue is an important ability for multi-modal models.

3.4. Dataset Statistics

Question Type and Length. C-VQA-Real contains 3144 questions in total, with 2014 numerical questions and 1130 boolean questions. In numerical questions, 1150 questions are from the direct group and 864 questions are from the indirect group. All the numerical questions are “How many” questions. 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-Real is 13.15 words, much longer than that of original ques-

Table 3. **Evaluation results on three subgroups of C-VQA-Real** by ViperGPT[9], VisProg[14], BLIP2 [24], InstructBLIP [8], MiniGPT4 [5, 47], LLaVA [26], CogVLM [37], and Qwen-VL [4]. All of them suffer from significant performance drop from counterfactual questions.

| Model Type | Model | Numerical Direct | | Numerical Indirect | | Boolean | |
|----------------|---------------------------------------|------------------|---------------------|--------------------|---------------------|-------------|--------------------|
| | | Original | Counterfactual | Original | Counterfactual | Original | Counterfactual |
| Neuro-symbolic | VisProg | 40.3 | 39.9 (-0.4) | 38.4 | 16.8 (-21.6) | 75.4 | 29.5 (-45.9) |
| | ViperGPT | 83.8 | 71.4 (-12.4) | 78.6 | 30.2 (-48.4) | 95.0 | 28.2 (-66.8) |
| End-to-end | BLIP2 (FlanT5 _{XXL}) | 43.4 | 32.3 (-11.1) | 49.1 | 37.5 (-11.6) | 77.8 | 75.1 (-2.7) |
| | InstructBLIP (FlanT5 _{XXL}) | 57.7 | 42.1 (-15.6) | 58.9 | 43.8 (-15.1) | 83.1 | 73.9 (-9.2) |
| | InstructBLIP (Vicuna-7B) | 64.5 | 30.1 (-34.4) | 64.2 | 30.2 (-34.0) | 83.9 | 50.4 (-33.5) |
| | InstructBLIP (Vicuna-13B) | 63.5 | 41.7 (-21.8) | 63.4 | 39.7 (-23.7) | 86.6 | 61.8 (-24.8) |
| | MiniGPT-4 (Vicuna-7B) | 31.2 | 19.6 (-11.6) | 30.6 | 19.6 (-11.0) | 55.8 | 41.2 (-14.6) |
| | MiniGPT-v2 (Llama2-Chat-7B) | 55.0 | 25.9 (-29.1) | 55.2 | 29.1 (-26.1) | 76.8 | 46.7 (-30.1) |
| | LLaVA-7B (Vicuna-7B) | 38.8 | 38.8 (-0.0) | 42.0 | 37.5 (-4.5) | 60.7 | 55.9 (-4.8) |
| | LLaVA-13B (Vicuna-13B) | 31.2 | 31.2 (-0.0) | 38.3 | 31.4 (-6.9) | 67.3 | 64.3 (-3.0) |
| | LLaVA-1.5-7B (Vicuna-7B) | 60.4 | 36.5 (-23.9) | 62.2 | 37.4 (-24.8) | 86.2 | 58.5 (-27.7) |
| | LLaVA-1.5-13B (Vicuna-13B) | 62.0 | 41.0 (-21.0) | 66.4 | 41.2 (-25.2) | 88.0 | 60.7 (-27.3) |
| | CogVLM-Chat | 49.2 | 20.3 (-28.9) | 50.3 | 25.9 (-24.4) | 83.6 | 60.0 (-23.6) |
| | Qwen-VL-Chat | 65.1 | 30.7 (-34.4) | 69.2 | 29.4 (-39.8) | 88.1 | 49.3 (-38.8) |

Table 4. **Evaluation results on C-VQA-Real** by four end-to-end models when chain-of-thought prompt “Let’s think step by step:” is post-pended to the questions.

| Model | Num. Direct | | Num. Indirect | | Boolean | |
|------------------|-------------|--------------|---------------|--------------|---------|--------------|
| | Ori. | Counterfact | Ori. | Counterfact | Ori. | Counterfact |
| InstructBLIP-13B | 63.2 | 34.2 (-29.0) | 62.3 | 33.7 (-28.6) | 82.8 | 51.0 (-31.8) |
| MiniGPT-v2-7B | 46.8 | 21.4 (-25.4) | 44.0 | 24.5 (-19.5) | 72.9 | 46.3 (-26.6) |
| LLaVA-1.5-13B | 62.1 | 41.7 (-20.4) | 64.9 | 42.2 (-22.7) | 87.9 | 60.9 (-27.0) |
| Qwen-VL-Chat | 65.2 | 29.7 (-35.5) | 68.4 | 24.9 (-43.5) | 87.4 | 47.5 (-39.9) |

tions (5.88 words) [3]. C-VQA-Synthetic contains 3000 questions with a multiple-choice answer selection setting. The average length of counterfactual questions in C-VQA-Synthetic is 25.33 words, also longer than that of original questions (16.00 words).

Answer Statistics. For C-VQA-Real, 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 Fig. 3. Each answer in the boolean type is a single “yes” or a single “no”. The percentage of “no” is 55.13%, while “yes” is 44.87%. For C-VQA-Synthetic, the options are randomly generated, and we control the proportion of each option close to 25%.

4. Experiments

In this section, we provide the evaluation of current state-of-the-art multi-modal LLMs on our proposed C-VQA. Our evaluation covers both neuro-symbolic models like

ViperGPT [9] and VisProg [14] and end-to-end models including LLaVA [26], MiniGPT4 [5, 47], BLIP2 [24], InstructBLIP [8], CogVLM [37], and Qwen-VL [4]. Our evaluation results reveal several interesting findings as detailed below. Implementation details are in the appendix.

4.1. Experiments on C-VQA-Real

Performance of End-to-end Models. As shown in Tab. 3, end-to-end models perform significantly worse with counterfactual questions than the original ones in all three groups. We find that reasoning difficulties have a significant impact on the results. 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. We further evaluate four end-to-end models on C-VQA-Real with a prompt “Let’s think step by step” for chain-of-thought reasoning [21, 38]. As shown in Tab. 4, the performance of the models remains constant or even drops with CoT, which indicates that the CoT strategy contributes little to solving our counterfactual questions. This highlights the need for novel prompting techniques for eliciting counterfactual reasoning ability or new paradigms for fine-tuning end-to-end multi-modal models to help with the counterfactual reasoning tasks.

Performance of Neuro-symbolic Models. Similar to end-to-end models, the performance of neuro-symbolic models on counterfactual questions is worse than that on

Table 5. Evaluation on the GPT-4V model using a subset of C-VQA-Real. Note that the results are not directly comparable with other tables.

| GPT-4V | Original | Counterfactual |
|--------------------|----------|----------------|
| Numerical Direct | 50.4 | 14.6 (-35.8) |
| Numerical Indirect | 56.8 | 40.7 (-16.1) |
| Boolean | 65.9 | 48.0 (-17.9) |

the original ones. And it still holds that reasoning difficulty influences accuracy. In the evaluation of ViperGPT [9], the accuracy drops 12.4% when one-step reasoning is required (*i.e.*, ‘Numerical direct’), and it drops 48.4% when multi-step reasoning is needed (*i.e.*, ‘Numerical indirect’). We note that the gap of neuro-symbolic models between counterfactual performance and original performance is much larger than that of any end-to-end model on ‘Numerical indirect’ and ‘Boolean’ questions. This indicates that despite the fact that neuro-symbolic models like ViperGPT [9] and VisProg [14] directly generate codes, it cannot handle complex counterfactual reasoning tasks.

Finding 1: Performance of neuro-symbolic models are worse than end-to-end models on complex reasoning tasks.

Results of Neuro-symbolic Models With Different Code LLMs. Keeping the API tools, in-context prompts, and questions the same, we can test the coding abilities of different code-generation LLMs by replacing the ChatGPT model inside ViperGPT. The evaluation results are shown in Tab. 6. Notably, the counterfactual presupposition causes significant and consistent performance drops across different model families and scales. This observation emphasizes the inherent limitations of current language models in handling more complex reasoning scenarios. Within this general trend, individual model performance varies. GPT-3.5-turbo generally outperforms other models in the original setting but also suffers from substantial drops in the counterfactual scenarios. CodeLlama [35] shows a relatively moderate performance decrease when counterfactuals are introduced, suggesting some resilience but still a noticeable drop. WizardCoder [28] performs least effectively in both original and counterfactual contexts. Our results here highlight an important and urgent need for improvement in enhancing the models’ abilities to handle counterfactual reasoning.

From the above experiments, we can see that no model families, neither end-to-end nor neuro-symbolic, can handle counterfactual questions well. This indicates that C-VQA is challenging and calls for further explorations.

Finding 2: No model family can consistently address counterfactual questions.

Evaluation of GPT-4V. We also provide an additional evaluation of our proposed benchmark with the GPT-4V [32]. Due to the rate limit and the late date when the API becomes available, we only test GPT-4V on a 10% randomly selected subset of C-VQA-Real, results are presented in Tab. 5. We show that despite the strong performance of GPT-4V at various visual problems [41], our counterfactual questions still pose a challenge to GPT-4V as the model shows an over 10% drop in performance for all three types of counterfactual questions in our benchmark.

Finding 3: Even strongest models such as GPT-4V cannot solve our benchmark.

4.2. Experiments on C-VQA-Synthetic

We test the synthetic questions in the same way as the manually annotated questions. However, we find that almost no models show normal performance in the synthetic dataset C-VQA-Synthetic. As shown in Tab. 7, the performance of InstructBLIP (Vicuna-7B) is at a random level 25% and the performance of InstructBLIP (FlanT5_{XXL}) and BLIP2 (FlanT5_{XXL}) is worse than the random level. We further prompt InstructBLIP (Vicuna) to merely count the flowers and it gives random answers, which indicates that it cannot recognize flowers in the synthetic images. Similar to the observation of [49], we notice that the distribution of the options answered by these models is highly biased. For example, BLIP2 (FlanT5_{XXL}) answers (*D*) to most questions. It is important to emphasize that the prompt for multiple-choice questions must be meticulously designed; failure to do so may result in some models being unable to select any of the provided options. For neuro-symbolic models, they perform better in the reasoning part but fail to answer the synthetic questions because no available APIs can handle the queries to the synthetic data. For example, ViperGPT cannot check whether a flower is in a polygon because no tools are available for querying this. However, it is worth noting that it can often produce correct code even when counterfactual presuppositions are added. These results indicate that current models cannot generalize beyond the training domain of real images, let alone handling counterfactual reasoning in the OOD synthetic domain.

4.3. Bias Analysis

As our dataset C-VQA-Real is based on VQAv2 and the images were from the COCO dataset, we further study the bias of the COCO dataset in the counterfactual questions. Prior work [46] studies bias propagation pathways within image captioning on the COCO dataset by annotating the

Table 6. Evaluation results of ViperGPT on C-VQA-Real when combining with different code LLMs.

| LLM of ViperGPT | Numerical Direct | | Numerical Indirect | | Boolean | |
|------------------------|------------------|--------------------|--------------------|---------------------|-------------|---------------------|
| | Original | Counterfactual | Original | Counterfactual | Original | Counterfactual |
| GPT-3.5-turbo | 83.8 | 71.4 (-12.4) | 78.6 | 30.2 (-48.4) | 95.0 | 28.2 (-66.8) |
| CodeLlama-7B-Instruct | 78.0 | 50.8 (-27.2) | 72.2 | 20.4 (-51.8) | 70.5 | 31.9 (-38.6) |
| CodeLlama-13B-Instruct | 78.9 | 72.4 (-6.5) | 73.4 | 26.2 (-47.2) | 74.5 | 41.6 (-32.9) |
| CodeLlama-34B-Instruct | 78.3 | 68.3 (-10.0) | 73.5 | 20.9 (-52.6) | 65.0 | 36.5 (-28.5) |
| WizardCoder-Python-7B | 72.8 | 59.1 (-13.7) | 68.6 | 18.3 (-50.3) | 27.7 | 29.5 (+1.8) |
| WizardCoder-Python-13B | 73.7 | 62.3 (-11.4) | 67.6 | 18.8 (-48.8) | 56.3 | 34.4 (-21.9) |
| WizardCoder-15B | 78.0 | 55.4 (-22.6) | 72.6 | 18.8 (-53.8) | 57.7 | 15.6 (-42.1) |

Table 7. Evaluation results on C-VQA-Synthetic.

| Model | Original | Counterfactual |
|---------------------------------------|----------|----------------|
| InstructBLIP (Vicuna-7B) | 26.9 | 25.7 |
| InstructBLIP (FlanT5 _{XXL}) | 16.7 | 19.9 |
| BLIP2 (FlanT5 _{XXL}) | 17.5 | 20.2 |

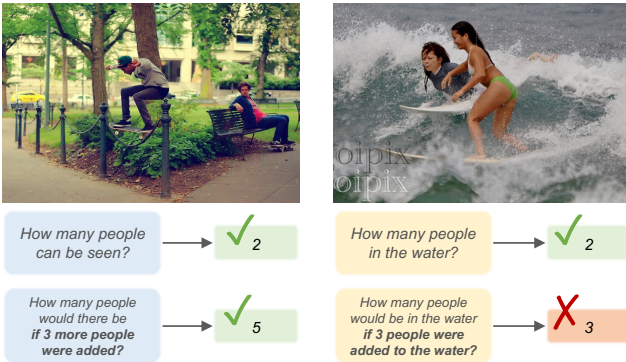


Figure 4. **Qualitative example of biases in MLLMs.** Given similar questions, InstructBLIP (Vicuna-7B) provides correct answers for the male instance but incorrect answers for the female instance.

perceived gender and skin color of 28,315 of the depicted people. In virtue of this dataset, we could get the overlap between our dataset and the selected images and find that 985 images in C-VQA-Real are annotated with gender and skin information. Among these images, we only keep those that contain only males or females, resulting in 717 images eventually (510 male images and 207 female images). We then evaluate the performance of different models by subgroups. The result is shown in Fig. 5.

Specifically, for each model we obtain the performance of both original and counterfactual questions of different subgroups. We then compute the difference between the accuracy of original and counterfactual questions in each subgroup. Formally, we have the formula $\text{diff}(\text{male}) = \text{male}(\text{ori}) - \text{male}(\text{cf})$ and $\text{diff}(\text{female}) = \text{female}(\text{ori}) - \text{female}(\text{cf})$. As shown in Fig. 5, most end-to-end models have a larger $\text{diff}(\text{female})$ than $\text{diff}(\text{male})$, which indicates that when presenting

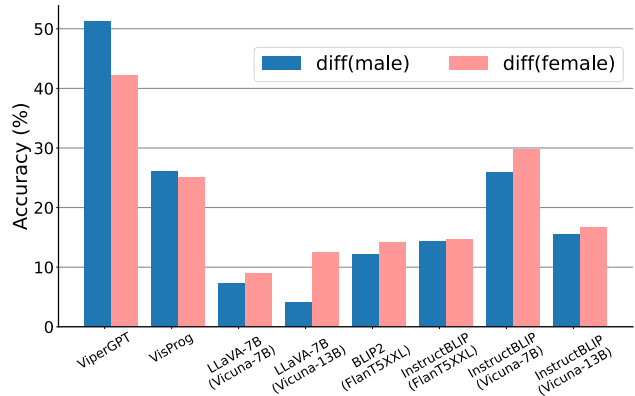


Figure 5. **Performance difference of original and counterfactual questions on the male and female subgroup on C-VQA-Real.** We can see that end-to-end models are often biased toward the male subgroup, and neuro-symbolic models are biased toward the female subgroup. The larger the gap between the performance differences, the larger the bias.

counterfactual presuppositions, they struggle more with females images. Further studies are needed to understand the reasons and improve the fairness in the model reasoning process.

Finding 4: MLLMs demonstrate systematic bias in answering gender-related counterfactual questions.

5. Conclusion

In this paper, we study the ability of current multi-modal language models to handle counterfactuals – a core cognition ability of human intelligence. We build a novel dataset C-VQA to test this counterfactual reasoning ability, the dataset is designed to have three types of counterfactual questions, with mixed real and synthetic images and questions. Our evaluations on C-VQA reveal several findings, the most significant finding is that no model in the current multi-modal model literature can consistently handle our counterfactual questions. We have released our code and the dataset to help the community move forward on achieving human-level multi-modal intelligence.

References

- [1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Anirudha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 2, 3
- [2] Anthropic. Claude. <https://claude.ai>, 2023. 3
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 2, 6
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023. 1, 2, 6
- [5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*, 2023. 2, 6
- [6] 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, 5
- [7] 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. 3
- [8] 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. *NeurIPS*, 2023. 2, 6
- [9] Suris D'Ádac, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *ICCV*, 2023. 1, 6, 7, 14
- [10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023. 3
- [11] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 1
- [12] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *ICLR 2024*, 2023. 1
- [13] 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, 3
- [14] Tanmay Gupta and Anirudha Kembhavi. Visual programming: Compositional visual reasoning without training. In *CVPR*, 2017. 1, 6, 7
- [15] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv:1807.01697*, 2019. 5
- [16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021. 3
- [17] 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. 3
- [18] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2, 3
- [19] 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. 1, 2, 3
- [20] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *ICML*, 2023. 1
- [21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 6
- [22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*, 2023. 3
- [23] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*, 2022. 2
- [24] 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, 2, 6
- [25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1
- [26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 6, 14
- [27] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 3
- [28] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv:2306.08568*, 2023. 7
- [29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 1, 2, 3
- [30] 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, 3
- [31] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 3, 4
- [32] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 2, 4, 7

- [33] Dong Qingxiu, Li Lei, Dai Damai, Zheng Ce, Wu Zhiyong, Chang Baobao, Sun Xu, Xu Jingjing, Li Lei, and Sui Zhi-fang. A survey on in-context learning. *arXiv:2304.08485*, 2023. 4, 11
- [34] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NeurIPS*, 2015. 2
- [35] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv:2308.12950*, 2023. 7
- [36] Haoqin Tu, Bingchen Zhao, Chen Wei, and Cihang Xie. Sight beyond text: Multi-modal training enhances llms in truthfulness and ethics. *arXiv:2309.07120*, 2023. 1
- [37] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079*, 2023. 2, 6
- [38] 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. 4, 6, 11
- [39] Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 2
- [40] 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. 3
- [41] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv:2309.17421*, 9, 2023. 7
- [42] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1
- [43] Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. Ifqa: A dataset for open-domain question answering under counterfactual presuppositions. *arXiv:2305.14010*, 2023. 3
- [44] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490*, 2023. 3
- [45] Bingchen Zhao, Quan Cui, Hao Wu, Osamu Yoshie, and Cheng Yang. Vision learners meet web image-text pairs. *arXiv:2301.07088*, 2023. 1
- [46] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *ICCV*, 2021. 7
- [47] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023. 2, 6
- [48] Yongshuo Zong, Oisín Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*, 2023. 1
- [49] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv:2310.01651*, 2023. 7