# Causal-CoG: A Causal-Effect Look at Context Generation for Boosting Multi-modal Language Models

Shitian Zhao[1] Zhuowan Li[2] Yadong Lu[1] Alan Yuille[2] Yan Wang[1*]
[1]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
[2]Johns Hopkins University

## Abstract

*While Multi-modal Language Models (MLMs) demonstrate impressive multimodal ability, they still struggle on providing factual and precise responses for tasks like visual question answering (VQA). In this paper, we address this challenge from the perspective of contextual information. We propose Causal Context Generation,* ***Causal-CoG****, which is a prompting strategy that engages contextual information to enhance precise VQA during inference. Specifically, we prompt MLMs to generate contexts, i.e, text description of an image, and engage the generated contexts for question answering. Moreover, we investigate the advantage of contexts on VQA from a causality perspective, introducing causality filtering to select samples for which contextual information is helpful. To show the effectiveness of Causal-CoG, we run extensive experiments on 10 multimodal benchmarks and show consistent improvements, e.g., +6.30% on POPE, +13.69% on Vizwiz and +6.43% on VQAv2 compared to direct decoding, surpassing existing methods. We hope Casual-CoG inspires explorations of context knowledge in multimodal models, and serves as a plug-and-play strategy for MLM decoding.* [1]

## 1. Introduction

Owing to the widespread adoption of Large Language Models (LLM) [31, 32], there has been a proliferation of research endeavors aimed at integrating visual ability into language models to build Multi-modal Language Models (MLM) [8, 23, 39]. Representative works, *e.g.*, LLaVA [23] and MiniGPT-4 [39], incorporates a pretrained visual encoder and a lightweight alignment module to align visual features into LLMs. Leveraging the intrinsic power of LLM, these models exhibit great capability in following user instructions and show good performance across various multimodal benchmarks, underscoring their potential in multimodal understanding.

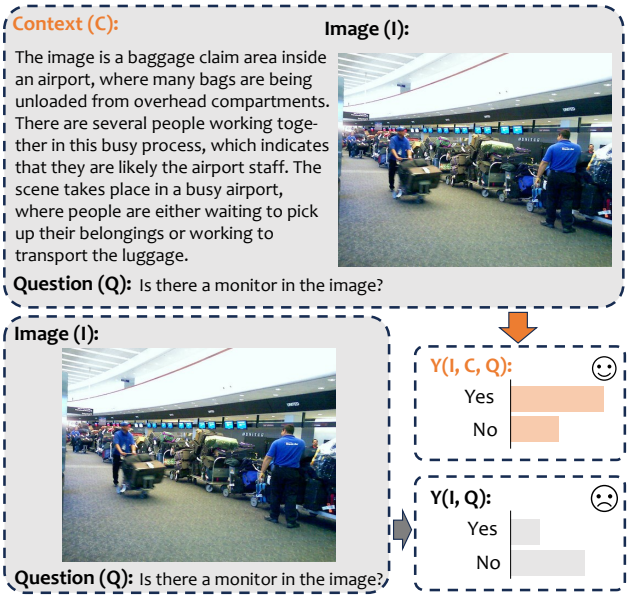


Figure 1. Contextual information helps VQA. In this example, the monitors in the image are too small to see, so the multi-modal language model (MLM) incorrectly predicts "no" for the question "Is there a monitor in this image?". However, if contextual information is provided, describing that this is an airport baggage hall, MLM can predict the right answer, since there are usually monitors in the baggage hall.

However, similar to LLMs, MLMs grapple with issues related to hallucination[2] [19]. The models may predict incorrect responses when inquired with misleading queries like the existence of an object in the image, and may struggle to grasp complex relationships among multiple objects in an image [21]. As the example shown in Fig. 1, when asked the question "is there a monitor in the image" for an image containing small monitors that can hardly be seen, current MLM incorrectly predicts the answer as "no". This inability to provide factual answers based on the visual content is commonly observed in MLMs, due to possible rea-

*Corresponding author: ywang@cee.ecnu.edu.cn
[1]Code is released zhaoshitian/Causal-CoG

[2]Hallucination is referred to *"the generated content is nonsensical or unfaithful to the provided source content"* [13]

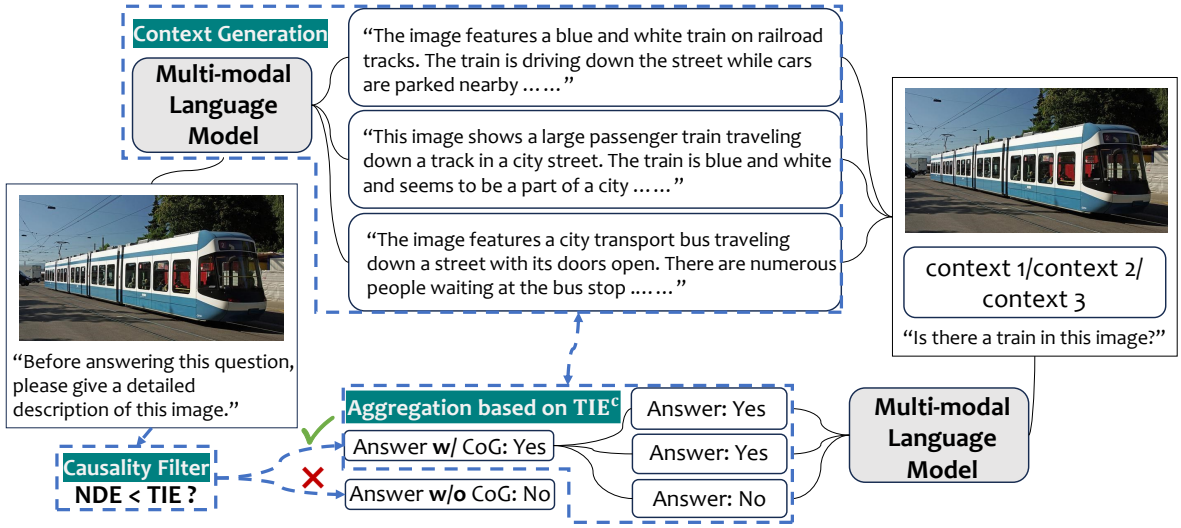


Figure 2. The framework of Causal-CoG: (1) Prompt a multi-modal language model to generate a descriptional context of the image; (2) Repeat this procedure, generating multiple candidates. Calculating and comparing the NDE and TIE of this sample using the generated candidates' answers likelihood distribution, to determine whether to apply CoG on the final answer. (3) For samples determined appropriate for using CoG, aggregate the candidates' answers based on the $TIE^c$ value of each candidate.

sons like shortcuts and noise in the training data, lack of modeling capacity for effective alignment of two modalities, etc.

There have been explorations boosting the off-the-shelf LLMs in the language community, *e.g.*, chain-of-thought [14], tree-of-thought [35] and retrieval-augmented generation [15], *etc*. Meanwhile, existing works in the vision-language community improves multimodal models through shortcut debiasing [27] or better alignment of the vision and language modality by various loss functions [2, 17, 36]. However, these works require training, thus cannot be easily applied to the off-the-shelf MLMs.

In this work, we improve MLM inference from the perspective of contextual knowledge. For example, in Fig. 1, the monitors are too small to see, but if context information like "this is a baggage claim in airport" is provided, then the question can be much easier to answer, since a baggage claims usually have monitors on the top. Given this intuition, we take advantage of the context, *i.e.* the description of the image, for more effective question answering.

To this end, we propose context generation with a causal-effect look, dubbed as Causal-CoG, which is a prompting technique for MLMs. The method is shown in Fig. 2. Concretely, instead of prompting MLMs to answer questions directly, we first instruct the MLM to generate a description, i.e. context, of the image by employing a simple prompt like "describe this image" (rephrased in different flexible ways), then prompt the model to answer the question based on the generated context description. With multiple prompting runs, different context descriptions can be generated, which provides rich information for answering the question. Moreover, to select the most helpful contexts from the multiple generated candidates, we leverage casual inference and take a causal effect look at the contexts. Finally, we propose a candidate aggregation method to attribute greater weight to better candidates considering the impact of the context on the answer. To the best of our knowledge, we are the first to develop "prompting" technique for MLMs.

We conduct extensive experiments to show the effectiveness of Casual-CoG. We run experiments on recent standard benchmarks, including MME [7], SEEDBench [16], MMBench [24], POPE [19], VSR [21], and reformed versions of some traditional datasets: VQAv2 [9], Vizwiz [10], GQA [12], OKVQA [26] and Winoground [30], from ReForm-Eval [20]. On all the benchmarks, Casual-CoG leads to consistent improvements, *e.g.*, +6.30% on POPE, +13.69% on Vizwiz and +6.43% on VQAv2, showing the advantage of context knowledge for VQA tasks. To summarize, our contributions are as follows:

- We propose Causal-CoG, a training-free decoding strategy that can be easily applied to the off-the-shelf MLMs for generating factual response for VQA.
- Causal-CoG explores the usage of context knowledge, with context filtering and aggregation using causality.
- Extensive experiments on 10 datasets show the effectiveness of our method. Causal-CoG consistently boosts the performance of MLMs.

## 2. Related Work

**Multi-modal Language Models.** With the remarkable success of Large Language Models (LLM) in the field of Natural Language Processing (NLP), there has been a significant

surge of interest in extending LLMs to the multi-modal domains. Noteworthy prior research,such as VisualGPT [4] and Frozen [33], has achieved significant progress in cross-modal tasks. VisualGPT [4] employs a series of prompts to facilitate LLMs input, while Frozen [33] trains a visual component to adapt to large language models. Additionally, Flamingo [1] has demonstrated impressive in-context few-shot learning capabilities by utilizing gated cross-attention to align a pre-trained vision encoder and a language model. Similarly, BLIP2 [18] leverages a Q-Former to align the visual and language modalities. Building upon these foundations, MiniGPT-4 [39] utilizes a limited number of high-quality image-text pairs to align MLM, while InstructBLIP [5] and LLaVA [23] enhance comprehensive MLM performance through instruction tuning. These advancements have significantly contributed to the advancement of MLM.

**Prompt Techniques for LLM.** Prompt engineering technique has been utilized on LLM to improve its reasoning ability, *e.g.*, CoT [14], ToT [35] and GoT [3]. The strategy of these methods is to first prompt LLM to generate the reasoning path, then based on the reasoning path, LLM can give a more factual answer. LLM may generate multiple reasoning path, and we need to select better reasoning paths or aggregate these paths, *e.g.*, self-consistent [34] and cumulative reasoning [37]. Prompt engineering technique is under-explored in MLM.

**Visual Question Answering with Context.** Context-aware VQA, which considers contextual information to better understand and answer questions about images, has gained significant attention in recent years. Experiments in ScienceQA [25], MM-CoT [38] and LLaMA-Adapter-V2 [8] have shown that contextual information can help MLM better answer the visual questions. There are also some works focusing on forcing the model to attend the context when answering questions, *e.g.*, CAD [29].

## 3. Causality in VQA with Context

In this section, we provide an introduction to fundamental concepts in causal inference, elucidating the foundational knowledge that underpins Causal-CoG's causality filter in Sec. 4.2.

### 3.1. Causal Graph

A causal graph is a directed acyclic graph that serves as a graphical representation of the causal relationships between nodes. It is typically denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ represents the set of variables within the causal graph, and $\mathcal{E}$ denotes the set of causal-effect relationships between pairs of variables. We construct the causal graph specific to VQA with context in Fig. 3. When doing VQA with context, image ($I$), question ($Q$) and context ($C$) are input to MLM, and MLM can output likelihood distribution on different options, termed as $Y$. In Fig. 3, if a causal-effect relationship exists between two variables, *e.g.*, $C$ and $I$, it can be symbolized by $I \rightarrow C$.

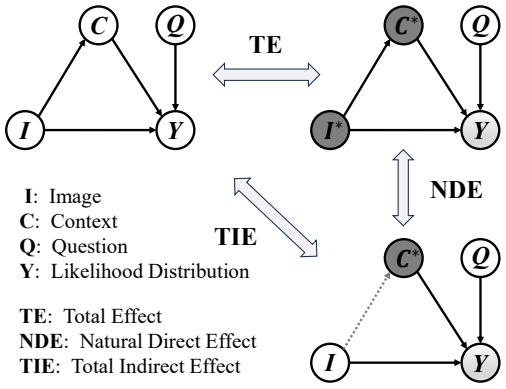


Figure 3. Causal graph of visual question answering with context. Capital letter with "*" as superscript means not input this variable into MLM when doing VQA.

### 3.2. Causal Effect

Causal effect serves as a measure that assesses the contrast between potential outcomes with and without a specific treatment. In Fig. 3, $I$ functions as a treatment for $Y$. There are two distinct types of effects between $I$ and $Y$: the direct effect ($I \rightarrow Y$) and the indirect effect via the generated context ($I \rightarrow C \rightarrow Y$). In the literature of causality, the Total Effect (TE) of $I$ on $Y$ is calculated by comparing $Y(I, C, Q)$ and $Y(Q)$, expressed as:

$$\text{TE} = \mathbb{E}[Y(I, C, Q) - Y(Q)], \tag{1}$$

where $\mathbb{E}[\cdot]$ represents expectation operation and $Y(I, C, Q)$ represents the answer obtained in the VQA task, taking $I$, $C$, and $Q$ as input. The TE encompasses two essential components: the Natural Direct Effect (NDE) and the Total Indirect Effect (TIE). By fixing the variable $C$, the NDE is computed by contrasting the potential outcomes with and without $I$, as formulated by:

$$\text{NDE} = \mathbb{E}[Y(I, Q) - Y(Q)]. \tag{2}$$

TIE represents the difference between TE and NDE:

$$\text{TIE} = \text{TE} - \text{NDE} = \mathbb{E}[Y(I, C, Q) - Y(I, Q)]. \tag{3}$$

## 4. Causal-CoG

Prior investigations have demonstrated the potential of context in augmenting the visual understanding and visual reasoning capabilities of MLM, *e.g.*, ScienceQA [25], MM-CoT [38], and LLaMA-Adapter-V2 [8]. In these prior studies, the contextual information has typically consisted of rationale or image descriptions, often necessitating input from domain experts or the utilization of retrained models.

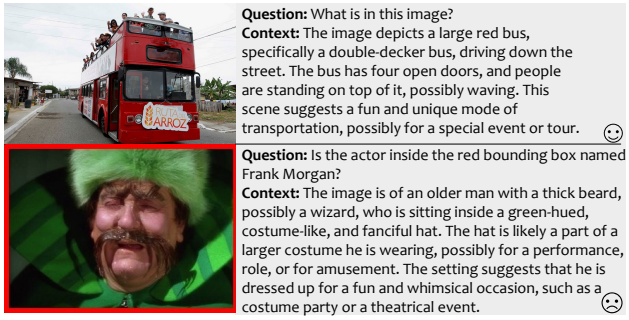


Figure 4. Helpful and unhelpful context. (1) For the upper image, context provides a descriptional information, which is helpful to the question: "What is it in this image?" (2) For the image below, to determine this actor's name, the common sense of the real world is needed, while context can't provide this kind of information.

Different from these approaches, we directly employ this Multi-modal Language Model by using a simple prompt to autonomously generate "context", bypassing the need for retraining the model to generate specific "context".

However, our exploration has revealed that the utility of the generated contexts is not consistently beneficial for addressing the questions posed. These contexts may sometimes be inaccurate or fail to provide valuable information. As a response to this challenge, we propose a causality-based filtering mechanism to determine whether to employ CoG[3] or not. For those instances where CoG is deemed appropriate, we aggregate the outcomes to derive the final answer.

### 4.1. Context Generation

Various forms of contextual information can be employed to facilitate the VQA task, including lecture content, as exemplified in the case of MM-CoT [38], and image descriptions, as utilized in LLaMA-Adapter-V2 [8]. Within the framework of CoG, we instruct the model to produce a detailed description of the provided image through the use of the following prompt: "Before answering this question, please give a detailed description of this image." This generated description is subsequently referred to as the "Context." Note that the MLM which generates the context is the same as the MLM where we apply Causal-CoG.

### 4.2. Causality Filter

The generated context provides relevant descriptions of the image. But it is not always helpful in answering the question, as illustrated in Fig. 4. This illustration underscores that the generated context has the potential to introduce irrelevant or even erroneous information for answering the question, thereby increasing the likelihood of an incorrect model response. In this scenario, the context can be regarded as a source of noise during the question-answering process, which could result in lower performance compared to a model that operates without CoG, as depicted in Fig.4. Therefore, it is important to evaluate whether the generated context is beneficial for a given sample. In essence, we need to devise a filtering mechanism to ascertain the utility of the generated context for individual samples.

The causality graph associated with the VQA task, showed in Fig. 3, reveals the existence of two causal relationships between the image ($I$) and the answers' likelihood distribution ($Y$): one corresponds to the direct causal pathway ($I \rightarrow Y$), while the other represents the indirect causal pathway ($I \rightarrow C \rightarrow Y$). The direct causal path signifies the immediate influence of the provided image on the answer, while the indirect path encapsulates the influence mediated by the generated context. In causality literature, we have the capacity to compute the Natural Direct Effect (NDE) and the Total Indirect Effect (TIE) between the provided $I$ and $Y$.

In Sec. 3, we have introduced how to calculate the NDE and TIE of $I$ on $Y$ (see Eq. 2 and Eq. 3). It is important to note that the subtraction in these formulations signifies a comparison between two types of outcomes, and this comparison is instantiated using the Jensen-Shannon Divergence (JSD) in our work. Therefore, the practical calculation methods for NDE, and TIE are as follows:

$$\text{NDE} = \mathbb{E}[\text{JSD}(Y(I, Q), Y(Q))], \tag{4}$$

$$\text{TIE} = \mathbb{E}[\text{JSD}(Y(I, C, Q), Y(I, Q))]. \tag{5}$$

Since $\mathbb{E}[\cdot]$ represents the expectation operation, we generate multiple candidates, consisting of generated context and corresponding answers' likelihood distribution, to estimate the expectation value. Overall, $N$ candidates are generated. For the $i$-th candidate, we denote the answers' likelihood distribution as $\hat{Y}_i$ and the candidate's context as $\hat{C}_i$. Here we use a sample with $N$ generated candidates to exemplify the NDE and TIE calculation process:

$$\text{NDE} = \text{JSD}(\hat{Y}(\hat{I}, \hat{Q}), \hat{Y}(\hat{Q})), \tag{6}$$

$$\text{TIE} = \frac{1}{N} \sum_{i=1}^{N} \text{JSD}(\hat{Y}_i(\hat{I}, \hat{C}_i, \hat{Q}), \hat{Y}_i(\hat{I}, \hat{Q})), \tag{7}$$

where different candidates have different $\hat{Y}_i$ and $\hat{C}_i$. $I$ and $Q$ remain the same for all the candidates, denoted as $\hat{I}$ and $\hat{Q}$ for this sample.

For any given sample, if NDE $<$ TIE, it implies that the indirect effect plays a more pivotal role in responding to the question, signifying that the context can be instrumental in addressing the question effectively. Consequently, we select such samples for the application of the CoG technique. In contrast, for samples whose NDE $>$ TIE, we opt to use the answer generated directly by the MLM as the final answer.

[3]CoG in this paper refers to the process of generating multiple candidates and aggregating their answers as the final answer, *i.e.*, removing the causality filter's judgment in Causal-CoG.

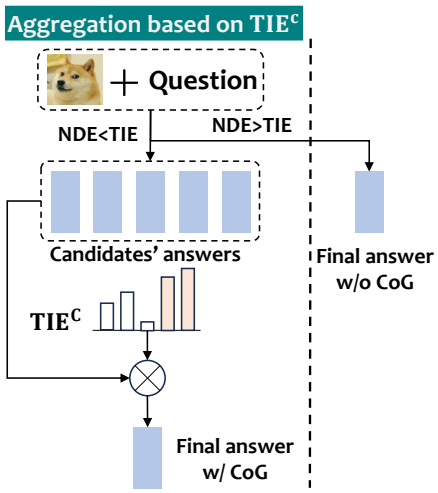


Figure 5. Pipeline of aggregation strategy, exemplified by Top-2 aggregation of 5 candidates' answers. (1) Compare the NDE and TIE values of this sample, determining whether to use CoG to decide the final answer. (2) For samples with TIE>NDE, aggregate answers based on $\text{TIE}^{\text{c}}$.

### 4.3. Candidate Aggregation

Given multiple candidates' contexts, we propose an aggregation method to obtain the ultimate response. It's crucial to acknowledge that, despite applying a causality filter to select these samples, not all candidates produce desirable responses. It's important to assign different weights to different candidates. A trivial approach is to use the answer's likelihood as the weight to aggregate all the candidates' answers or majority voting without weights [34]. But in this work, The determination of higher weight assignment for these candidates hinges on a consideration of the impact of the context on the answer.

**Effect of the Context.** As illustrated in Fig. 4, certain contexts may not contribute significantly to the question's answer. Therefore, we consider candidates with a higher TIE value of the image on the answer, mediated through the generated context, to be "better." Similar to the approach described in Section 3.2, we compute the individual-level TIE value for $i$th candidate and denote it as $\text{TIE}_i^{\text{c}}$, which is expressed as:

$$\text{TIE}_i^{\text{c}} = \text{JSD}(\hat{Y}_i(\hat{I}, \hat{C}_i, \hat{Q}), Y(\hat{I}, \hat{Q})). \tag{8}$$

Here, $\text{TIE}_i^{\text{c}}$ quantifies the degree of indirect effect and aids in the evaluation of candidate effectiveness in leveraging context to answer questions.

**Top-$k$ Aggregation.** For samples applied with CoG, we need to aggregate the candidates' answers to get the final answer, considering each candidate's $\text{TIE}^{\text{c}}$. Assuming that we sampled $N$ candidates and each candidate consists of a context and a corresponding answer. After getting the $\text{TIE}^{\text{c}}$ of each candidate, then

$$\textbf{TIE}^{\text{c}} = [\text{TIE}_1^{\text{c}}, ..., \text{TIE}_N^{\text{c}}], \tag{9}$$

where $\textbf{TIE}^{\text{c}}$ is the set of all candidates' $\text{TIE}^{\text{c}}$.

Next, top-$k$ candidates in $\textbf{TIE}^{\text{c}}$ are picked and whose indices are grouped into a set $\Omega_{\text{TIE}^{\text{c}}}$. Then, we only keep the top-$k$ candidates' values while setting other candidates' values to zero via:

$$\overline{\text{TIE}}_i^{\text{c}} = \begin{cases} \text{TIE}_i^{\text{c}} & : i \in \Omega_{\text{SIM}} \\ 0 & : i \notin \Omega_{\text{SIM}}. \end{cases} \tag{10}$$

Finally, we can get the final answer for the sample using the weighted majority vote, according to each candidate's $\overline{\text{TIE}}^{\text{c}}$.

## 5. Experiment

### 5.1. Setup

We conduct experiments on 10 VQA benchmarks, comparing the performance of Causal-CoG with competitors. The results show that Causal-CoG can boost the MLM's ability on both perception and reasoning tasks, as well as reduce object hallucination.

**Evaluation Benchmarks.** We evaluate on 4 comprehensive benchmarks containing multiple subtasks ranging from object identification to visual reasoning: MME [7], SEEDBench [16], MMBench [24] and VQAv2 [9]. Besides, we use POPE [19] for evaluating object hallucination; VSR [21] and Winoground [30] for evaluating visual spacial understanding ability; OKVQA [26] for testing MLM's ability to leverage outside knowledge; Vizwiz [10] for assisting blind people; GQA [12] for real-world visual reasoning. Noted that Winoground [30], OKVQA [26], VQAv2 [9], Vizwiz [10] and GQA [12] used in this paper are not the original versions, but the split and reformed versions in ReForm-Eval [20] and they are reformed into the single-choice forms, along with $\cdot^*$ in the rest of the paper.

**Multi-modal Language Models.** We evaluate Causal-CoG on two representative MLMs: LLaVA and LLaVA-v1.5.

- **LLaVA** [23]: LLaVA is trained on image-text pair data and visual instruction data. The 7B checkpoint is used in our experiment.
- **LLaVA-v1.5** [22]: LLaVA-v1.5 is an improved version of LLaVA, trained on more multi-modal data. We use the 7B version in our experiment.

**Implementation Details of Candidate Generation.** We denote the context and the likelihood distribution of the $i$th candidate on all the answer options as $\hat{C}_i$ and $\hat{Y}_i$.

- **Context**: $\hat{C}_i$ is generated by the language model in MLM, thus we can choose the hyperparameters in the sampling strategy to control the generation process of $\hat{C}_i$. For all the MLMs mentioned above, we follow the similar setting as [34], setting the temperatur to 0.9 [11], and truncating

| | MME | SEEDBench | MMBench | POPE | VSR | Winoground* | OKVQA* | VQAv2* | Vizwiz* | GQA* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Ensemble* | 57.03 | 39.26 | 41.43 | 66.46 | 54.42 | 61.25 | 30.75 | 42.86 | 29.47 | 37.23 |
| *One-shot* | 54.72 | 39.82 | 41.57 | 68.66 | 51.55 | 62.50 | 30.56 | 42.91 | 32.25 | 37.55 |
| Naive-CoG | 60.43 | 39.20 | 42.53 | 68.63 | 58.10 | 63.75 | 30.75 | 47.81 | 38.05 | 39.78 |
| LLaVA | 57.49 | 38.65 | 41.16 | 64.55 | 55.56 | 61.25 | 30.36 | 43.10 | 28.54 | 37.71 |
| +Causal-CoG | **61.23** | **40.66** | **43.52** | **70.85** | **58.92** | **66.25** | **32.74** | **49.44** | **42.23** | **41.45** |
| Δ | +3.74 | +2.01 | +2.36 | +6.30 | +3.36 | +5.00 | +2.38 | +6.43 | +13.69 | +3.74 |
| LLaVA-v1.5 | **72.27** | 45.45 | 43.96 | 85.76 | 58.02 | 63.75 | 33.92 | 52.89 | 39.21 | **49.64** |
| +Causal-CoG | 71.84 | **45.93** | **45.43** | **86.34** | **61.62** | **66.25** | **34.92** | **54.34** | **46.64** | 48.77 |
| Δ | -0.43 | +0.48 | +1.47 | +0.64 | +3.60 | +2.50 | +1.00 | +1.45 | +7.43 | -0.87 |

Table 1. Accuracy results on 10 benchmarks. (1) Causal-CoG improves LLaVA [23] and LLaVA-v1.5 [22]'s performance on 10 benchmarks, *e.g.*, +13.69% on Vizwiz* [10] and +6.43% on VQAv2* [9]. (2) *Ensemble* and *One-shot* is conducted on LLaVA [23]. On some benchmarks, these two methods show a slight improvement. (3) In Naive-CoG, the number of generated candidates is set to 1. Though Naive-CoG's improvement is not as significant as Causal-CoG, it still boosts LLaVA [23] on most benchmarks, *e.g.*, +9.51% on Vizwiz*. (4) *Dataset** is the split and reformed version of *Dataset* from ReForm-Eval [20].

the top-$k$ ($k = 40$) tokens with highest probability [6]. In our experiments we apply Causal-CoG on LLaVA with 40 candidates and on LLaVA-v1.5 with 20 candidates.

- **Answer's likelihood distribution**: For the $i$th candidate, the answer options list of this sample is denoted as $L = [A_1, A_2, ..., A_M]$. $\hat{Y}_i$ is the likelihood distribution on $L$. $A_j$'s likelihood, termed as $\hat{Y}_{i,j}$, can be calculated as:

$$\hat{Y}_{i,j}(\hat{X}_i) = \exp^{\frac{1}{K_j}\sum_{k=1}^{K_j} \log P(t_k|\hat{X}_i, t_1, t_2, ..., t_{k-1})} \tag{11}$$

where $\hat{X}_i$ is the input MLM take, *e.g.*, $\hat{X}_i$ is $[\hat{I}, \hat{C}_i, \hat{Q}]$ in $\hat{Y}_i(\hat{I}, \hat{C}_i, \hat{Q})$. $P(t_k|\hat{X}_i, t_1, t_2, ..., t_{k-1})$ is the probability of generating the $k$th token $t_k$ conditioned on $\hat{X}_i$ and previous generated tokens $t_1 \sim t_{k-1}$. $K_j$ is the length of the $j$th answer. As all the benchmarks we use are single-choice formulations, the output answer, regardless of with the application of Causal-CoG or not, is decided based on the options' likelihood distribution, *i.e.*, choose the option with highest likelihood as the output answer.

**Competitors.** Since the study of MLM's decoding strategy[4] is still under-explored, we compare our Causal-CoG with two methods that can be used directly during the decoding stage of MLM: **(1) ensemble MLM through various prompts** [28] and **(2) one-shot in-context learning** [1]. The implementation details are as follows.

- **Ensemble**: We ensemble the outputs by averaging the 5 answers' likelihood distribution using 5 different system prompts[5] on LLaVA. Full list of all system prompts we use is provided in the supplementary material.
- **One-shot**: Similar to Flamingo [1], we use the one-shot in-context learning on LLaVA.

[4]MLM's decoding strategy refers to methods can be used directly without retraining or fine-tuning the pretrained MLM during the inference time.

[5]System prompt refers to the prompt that is contained in every conversation's beginning, telling the MLM to act as an multi-modal chatbot. An example of system prompt is: *"You are a helpful language and vision assistant. And you are able to understand the visual content that the user provides, and assist the user with a variety of tasks using natural language."* [23]

## 5.2. Main Results

We apply Causal-CoG on LLaVA and LLaVA-v1.5 respectively, and compare the results with the ones without Causal-CoG. Table 1 shows the overall results on 10 benchmarks. Causal-CoG improves the performance of both LLaVA and LLaVA-v1.5 significantly on most of the benchmarks. For LLaVA, Causal-CoG boosts the accuracy by more than 5.00% over POPE, Winoground*, VQAv2* and Vizwiz*, 2.00%-3.00% absolute improvement over other benchmarks. When it comes to LLaVA-v1.5, we can still observe nearly 2.00%-3.00% improvement. It is worth mentioning that LLaVA-v1.5 is an advanced version of LLaVA. Compared with LLaVA, it performs better on all benchmarks significantly, while Causal-CoG can still work on this advanced and more developed MLM, which shows the potential of Causal-CoG for working on well-developed MLMs. In Table 1, *Ensemble* and *One-shot* works on some datasets, improving the accuracy incrementally, while these two methods can cause performance drop on other datesets, *e.g.*, VSR, VQAv2* and GQA*.

**Performance on Perception, Recognition and Object Hallucination.** In MME [7], SEEDBench [16] and MMBench [24], all the subtasks are split into two main parts: perception and cognition. Table 2 shows results on perception tasks and cognition tasks. POPE [19] is a benchmark for testing MLM's object hallucination issue. It contains three versions: POPE-Popular, POPE-Adversarial and POPE-Random. Detailed results of Causal-CoG is included in Table 2. We can see that Causal-CoG works better on perception tasks and object hallucination issues than cognition tasks. This may be due to that context MLM generated mainly focuses on the image's description, which is obviously helpful for perception tasks, and may be not helpful for cognition tasks in some cases, *e.g.*, code_reasoning.

## 5.3. Ablation and Analysis

### 5.3.1 Ablation of the Proposed Components

There are three main components in Causal-CoG: **Context Generation**, **Causality Filter** and **Top-$k$ Aggregation**. We

| | Cognition | | | Perception | | | Object hallucination | | |
|---|---|---|---|---|---|---|---|---|---|
| | $MME_c$ | $SEEDBench_c$ | $MMBench_c$ | $MME_p$ | $SEEDBench_p$ | $MMBench_p$ | POPE-P | POPE-A | POPE-R |
| LLaVA | 53.85 | 37.50 | 44.30 | 59.11 | 40.94 | 38.02 | 61.27 | 57.67 | 74.71 |
| +Causal-CoG | 56.43 | 39.43 | 45.26 | 63.37 | 43.12 | 41.78 | 65.97 | 63.60 | 82.99 |
| Δ | +2.59 | +1.93 | +0.96 | +4.26 | +2.18 | +3.76 | +4.70 | +5.93 | +8.28 |
| LLaVA-v1.5 | 53.57 | 46.88 | 44.37 | 80.58 | 42.59 | 43.55 | 87.13 | 82.03 | 88.11 |
| +Causal-CoG | 53.93 | 47.17 | 45.80 | 79.81 | 43.45 | 45.07 | 86.47 | 83.57 | 88.97 |
| Δ | +0.36 | +0.29 | +1.43 | -0.77 | +0.86 | +1.52 | -0.66 | +1.54 | +0.86 |

Table 2. Results on cognition tasks, perception tasks and object hallucination issues. (1) $Dataset_c$ represent the average of results over cognition tasks. While $Dataset_p$ represents the average of results over perception tasks. The list of cognition and perception tasks in MME [7], SEEDBench [16] and MMBench [24] is provided in supplementary material. (2) POPE-P, POPE-A and POPE-R are the abbreviation of POPE-Popular, POPE-Adversarial and POPE-Random respectively, which are three versions of POPE [19].

| LLaVA | MME | SEEDBench | MMBench | POPE | VSR | Wino ground * | OKVQA* | VQAv2* | Vizwiz* | GQA* |
|---|---|---|---|---|---|---|---|---|---|---|
| Casusal-CoG | 61.23 | 40.66 | 43.52 | 70.85 | 58.92 | 66.25 | 32.74 | 49.44 | 42.23 | 41.45 |
| w/o Top-$k_{TIE^c}$ | 61.58 | 39.93 | 43.44 | 70.66 | 59.17 | 65.00 | 32.74 | 49.49 | 41.30 | 40.89 |
| w/o Causality filter | 58.43 | 37.60 | 41.62 | 69.07 | 58.02 | 60.00 | 29.17 | 44.26 | 38.98 | 38.58 |
| Weighted sum (likelihood) | 59.94 | 39.89 | 42.69 | 70.03 | 58.10 | 60.00 | 30.56 | 46.18 | 37.59 | 39.46 |
| w/o Causality filter | 59.13 | 38.56 | 41.37 | 70.12 | 58.10 | 62.50 | 28.97 | 43.75 | 37.82 | 38.90 |
| Weighted sum (similarity) | 60.23 | 40.23 | 42.74 | 70.09 | 58.59 | 60.00 | 30.75 | 47.99 | 37.82 | 38.82 |
| w/o Causality filter | 59.34 | 38.93 | 41.62 | 70.20 | 58.02 | 62.50 | 29.37 | 47.20 | 38.52 | 38.11 |
| Unweighted sum | 60.24 | 40.17 | 42.71 | 70.03 | 58.10 | 60.00 | 30.75 | 47.90 | 38.05 | 38.75 |
| w/o Causality filter | 59.43 | 38.86 | 41.21 | 70.12 | 57.61 | 62.50 | 29.17 | 47.11 | 38.75 | 38.03 |

Table 3. Results of ablation and analysis experiments. (1) Either we remove the Top-$k$ aggregating strategy or the causality filter, performance drops on most datasets, signifying the necessity of these two modules. (2) During the aggregating stage, if the candidates' answers are aggregated based on other values, *e.g.*, likelihood of each answer, similarity between context and image, or simple majority vote, we find that Causal-CoG does not perform as well as before, showing the importance of $TIE^c$ values.

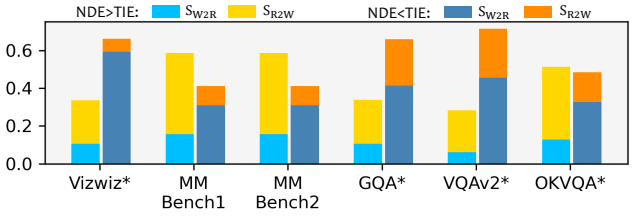


Figure 6. (1) MMBench1 and MMBench2 are two subtasks in MMBench: atribute_reasoning and finegrained_perception. (2) $S_{W2R}$ stands for samples on which CoG is helpful; $S_{R2W}$ means samples on which CoG is harmful. For samples with TIE>NDE, CoG is helpful for most of them.

conduct the ablation experiments of Causality Filter and Top-$k$ Aggregation respectively on 10 benchmarks mentioned above, illustrating the necessity and effectiveness of these two components.

**Candidates with Higher $TIE^c$ are More Important.** Table 3 shows in Causal-CoG, if we consider all candidates' $TIE^c$ when aggregating candidates' answers instead of using the top-$k$ strategy, accuracy on most benchmarks drops.

In Causal-CoG, when aggregating top-$k$ candidates, we select candidates with the 1st to 5th high $TIE^c$. We also explore the situation that uses candidates with 6th to 10th, 11th to 15th, 16th to 20th high $TIE^c$ to do weighted aggregation. As shown in Fig. 7, using candidates with lower $TIE^c$ cause a significant accuracy drop, showing the effectiveness of $TIE^c$ as the aggregation weight.

**The Causality Filter Works by Identifying Samples on which CoG is Helpful.** In Table 3, results without causality filter drop a lot compared with ones with causality filter, signifying the importance of causality filter in Causal-CoG.

The motivation of designing causality filter is to select the samples, where context has more consequence on the answer than image, *i.e.*, TIE>NDE. Thus, we count the samples that can benefit from CoG, *i.e.*, CoG corrects these samples' answers from wrong to right, termed as $S_{W2R}$, and samples that are harmed by CoG, *i.e.*, the original answers output by MLM directly are right but CoG changes them to wrong, termed as $S_{R2W}$. The distribution of $S_{W2R}$ and $S_{R2W}$ in the samples set with NDE>TIE and NDE<TIE are illustrated as in Fig. 6. As we can see, in samples with NDE<TIE, $S_{W2R}$ accounts for the majority, *i.e.*, CoG works well on most samples with NDE<TIE, thus we should apply CoG on these samples. While in samples with NDE>TIE, $S_{R2W}$ is the most, *i.e.*, CoG makes errors

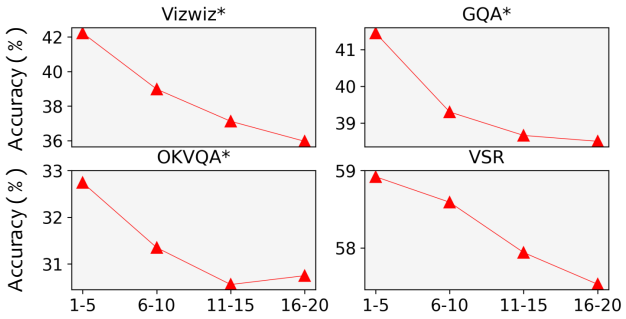


Figure 7. During aggregation, instead of aggregate candidates' answers based on top-k $TIE^c$ value, if we aggregate the answers from candidates with 6th to 10th, 11th to 15th, 16th to 20th high $TIE^c$ values, the performance drops a lot, showing that candidates with higher $TIE^c$ value are better when aggregating.

on most samples with NDE>TIE, so we should avoid applying CoG on these samples. These insightful results inform us how the causality filter works: **selecting the samples that can benefit from CoG**.

#### 5.3.2 Additional Analysis

**Other Aggregation Strategies.** In the aggregation strategy we used in this paper, $TIE^c$ is used as the weights. However, there are other measures we can use as the weights when aggregating. For example, we can use each answer's likelihood or the similarity[6] between the image and the generated context as the weights. Table 3 shows the results using the answer's likelihood and the similarity between the given image and the generated context as weights to aggregate in Causal-CoG's framework. We also take place the weighted aggregation with unweighted aggregation (simple majority vote), and results are shown in Table 3.

We find that aggregating with $TIE^c$ as weights outperforms the other three aggregation strategies, informing us that $TIE^c$ is a better metric for aggregation.

Besides, when removing the causal filter, performance drops even if candidates' answers are aggregated by the other three strategies, further showing the effectiveness of causality filter.

**Contextual Information Assists VQA Tasks in MLM.** To verify the direct effect of contextual information, we directly use the first generated candidate's answer as the final answer (getting rid of aggregation in Causal-CoG), termed as **Naive-CoG**, results are shown in Table 1. Naive-CoG improves MLM's performance on 10 benchmarks, showing the helpfulness of contextual information itself for VQA.

We also compare some samples' attention maps on the image queried by the answer of this visual question with and without Causal-CoG in Fig. 8. As we can see, Causal-CoG helps the MLM to attend the context region of the image.

[6]The similarity is calculated using pretrained *openai/clip-vit-patch14*

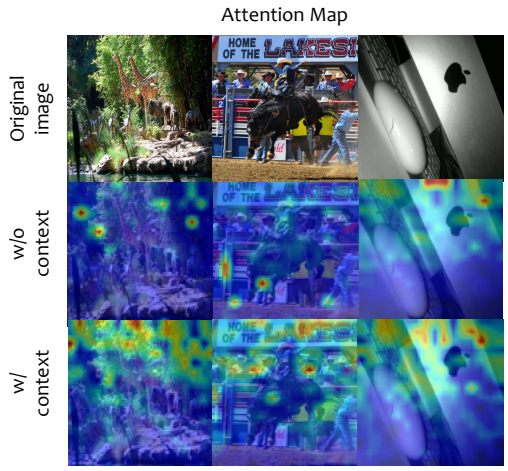


Figure 8. This figure reflects the changes in attention distribution after being provided context. (1) In these three samples, the question is: " Is there a cow/horse/screen in this image?" And the attention map is queried by the right answer of this question above. (2) After providing context when doing VQA, MLM pays more attention to the context in the images, which is helpful for answering the question. For example, in the left-most image, provided with context, the trees and various animals are attended, so MLM can infer to the existence of a cow based on the contextual information and the image.

**Causal-CoG on Other MLMs.** Besides LLaVA and LLaVA-v1.5, we also apply Causal-CoG on MiniGPT-4 [39], and compare the results with ones without Causal-CoG on MME [7]. Causal-CoG boosts MiniGPT-4's accuracy on MME's coarse-grained perception tasks by +2.09%, showing the Causal-CoG can generalize to other MLMs. More results are provided in supplementary material.

## 6. Conclusion

Motivated by contextual information can help MLM answer the visual questions better, we propose Causal-CoG to boost MLM's performance on VQA. In Causal-CoG, we also design a causality filter to determine whether the contextual information is helpful, thus deciding if we should use the answer from CoG or not. We prove the effectiveness of Causal-CoG by conducting the experiments over 10 VQA benchmarks. Causal-CoG improves the accuracy over all these 10 benchmarks significantly compared to the original MLM.

## 7. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62101191), ONR (Grant No. N00014-23-1-2641), Shanghai Natural Science Foundation (Grant No. 21ZR1420800), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3, 6

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *NeurIPS*, 2022. 2

[3] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *arXiv 2308.09687*, 2023. 3

[4] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 3

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv 2305.06500*, 2023. 3

[6] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 2018. 6

[7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2, 5, 6, 7, 8

[8] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter V2: parameter-efficient visual instruction model. *arXiv 2304.15010*, 2023. 1, 3, 4

[9] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017. 2, 5, 6

[10] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 2, 5, 6

[11] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020. 5

[12] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 2, 5

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 2023. 1

[14] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 2, 3

[15] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[16] 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. 2, 5, 6, 7

[17] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. 2

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 2023. 3

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv 2305.10355*, 2023. 1, 2, 5, 6, 7

[20] Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, Xuanjing Huang, and Zhongyu Wei. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv 2310.02569*, 2023. 2, 5, 6

[21] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *arXiv 2205.00363*, 2022. 1, 2, 5

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv 2310.03744*, 2023. 5, 6

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv 2304.08485*, 2023. 1, 3, 5, 6

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv 2307.06281*, 2023. 2, 5, 6, 7

[25] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. 3

[26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 2, 5

[27] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 2

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021. 6

[29] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv 2305.14739*, 2023. 3

[30] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 2, 5

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv 2302.13971*, 2023. 1

[32] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv 2307.09288*, 2023. 1

[33] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021. 3

[34] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 3, 5

[35] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv 2305.10601*, 2023. 2, 3

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022. 2

[37] Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. *arXiv 2308.04371*, 2023. 3

[38] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv 2302.00923*, 2023. 3, 4

[39] 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. 1, 3, 8