Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective

Meiqi Chen[1,2*], **Yixin Cao**[3], **Yan Zhang**[1,2†], **Chaochao Lu**[4†]

[1]State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China
[2]School of Intelligence Science and Technology, Peking University
[3]School of Computer Science, Fudan University
[4]Shanghai Artificial Intelligence Laboratory

`meiqichen@stu.pku.edu.cn`, `caoyixin2011@gmail.com`,
`zhyzhy001@pku.edu.cn`, `luchaochao@pjlab.org.cn`

Abstract

Recent advancements in Large Language Models (LLMs) have facilitated the development of Multimodal LLMs (MLLMs). Despite their impressive capabilities, MLLMs often suffer from over-reliance on unimodal biases (e.g., language bias and vision bias), leading to incorrect answers in complex multimodal tasks. To investigate this issue, we propose a causal framework to interpret the biases in Visual Question Answering (VQA) problems. Within this framework, we conduct an in-depth causal analysis to assess the causal effect of these biases on MLLM predictions. Based on the analysis, we introduce 1) a novel `MORE` dataset with 12,000 challenging VQA instances requiring multi-hop reasoning and overcoming unimodal biases. 2) a causality-enhanced agent framework `CAVE` that guides models to comprehensively integrate information from different modalities and mitigate biases. Our experiments show that MLLMs perform poorly on `MORE`, indicating strong unimodal biases and limited semantic understanding. However, when integrated with our `CAVE`, promising improvements in reasoning and bias mitigation can be seen. These findings provide important insights for the development of more robust MLLMs and contribute to the broader goal of advancing multimodal AI systems capable of deeper understanding and reasoning. Our project page is at `https://github.com/OpenCausaLab/MORE`.

1 Introduction

Following the success of Large Language Models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023b), Multimodal LLMs (MLLMs) (OpenAI, 2023; Team et al., 2023) have been proposed for various vision-language tasks (Fu et al., 2023; Liu et al., 2023b). Despite their promising results, it

Figure 1: Examples of over-reliance on unimodal biases. MLLMs (e.g., LLaVA) erroneously generate answers due to language bias (indicated by the underlined text below the left image) and vision bias (the right image).

remains unclear if they truly understand images and text in the context of multimodal reasoning.

As shown in the knowledge-based Visual Question Answering (VQA) problems of Figure 1, when prompted with "*Which country is hosting the next World Cup after this venue?*" MLLMs such as GPT-4V (OpenAI, 2023) and LLaVA (Liu et al., 2023a) may capture the language bias of "*the next World Cup*" and think that the next World Cup will be "*the 2022 FIFA World Cup held in Qatar*" (which is also outdated knowledge), while ignoring the exact venue presented in the image. Similarly, when presented with an image of "*The Shard*" in London, MLLM directly identifies "*The representative building is The Shard*" influenced by vision bias, overlooking the specific constraint "*in Berlin*" mentioned in the question. These inherent issues pose significant challenges to the reasoning capabilities of MLLMs, particularly when faced with more complex questions.

To investigate the issue of **M**LLMs' **O**ver-**RE**liance (`MORE`) on such unimodal biases, we propose a causal framework to interpret and quantify language and vision biases. We begin by defining a

---

Datasets	Knowledge-based	Multi-hop Reasoning	Answer Type	Unimodal Biases Evaluation	Rationale	# Size
Visual7W (Zhu et al., 2016)	✗	✗	Open-ended	✗	✗	327.9K
VQA (v2) (Goyal et al., 2017)	✗	✗	Open-ended	✗	✗	1.1M
FVQA (Wang et al., 2017)	✓	✗	Open-ended	✗	✓	5.8K
OKVQA (Marino et al., 2019)	✓	✗	Open-ended	✗	✗	14K
S3VQA (Jain et al., 2021)	✓	✗	Open-ended	✗	✗	7.5K
A-OKVQA (Schwenk et al., 2022)	✓	✗	Multi-choice	✗	✓	23.7K
INFOSEEK (Chen et al., 2023)	✓	✗	Open-ended	✗	✗	1.4M
MORE (Ours)	✓	✓	Multi-choice	✓	✓	12K

Table 1: Comparison of MORE with other VQA datasets, highlighting its incorporation of external knowledge, multi-hop reasoning, unimodal bias evaluation, and rationale for interpretability.

causal graph of MLLM's prediction on VQA problems, built on key causal factors like images and questions. Then, we identify a set of interventions in the context of VQA problems, thereby ascertaining the causal effect of unimodal biases on MLLM predictions via *do*-calculus (Pearl, 1995). This allows us to evaluate the sensitivity and robustness of MLLMs against unimodal biases.

Based on the above causal analysis, we curate a novel dataset termed MORE, comprising 12,000 VQA instances. This dataset advances existing VQA datasets by introducing a dedicated evaluation of unimodal biases. We adopt a Multiple Choice Question (MCQ) format to facilitate the evaluation, where each instance consists of an image, a question, and four candidate options. The image is sourced from an existing VQA dataset (Chen et al., 2023). For question and option curation, we incorporate a knowledge graph (KG) (Wang et al., 2021), allowing us to simulate MLLMs to navigate potential paths within the causal graph. Specifically, the options consist of one correct answer, and three distractors targeting language bias, vision bias, and multi-hop reasoning, respectively. We also provide the reasoning path, designated as *causal rationale*, in the KG for each instance, offering interpretability for evaluation. As summarized in Table 1, compared to existing VQA datasets, MORE features better comprehensiveness.

Furthermore, motivated by the causal analysis, we propose CAVE, a causality-enhanced method to mitigate unimodal biases in MLLMs. CAVE encompasses a diverse set of workflows, including question decomposition, causality-based enhanced self-reflection, external knowledge retrieval, and answer verification. This framework guides models in explicitly and comprehensively integrating information from multiple modalities while helping to

prevent biased decision-making and the selection of incorrect shortcuts.

Through extensive experiments on MORE with several leading MLLMs, we observe that: 1) most MLLMs perform much poorly on MORE, showing strong reliance on unimodal biases and low robustness to disturbances; 2) MLLMs still struggle to achieve precise semantic understanding in multimodal reasoning; 3) With our propose method CAVE, we can mitigate the unimodal bias in MLLMs, yet it still falls short of the ideal. This indicates that addressing the unimodal bias in MLLMs is a highly challenging issue that merits further exploration. Overall, our main contributions are as follows:

- We propose a causal framework to interpret and quantify the biases in VQA problems.
- We construct a new dataset, MORE, which requires multi-hop reasoning and overcoming biases, demonstrating superior comprehensiveness to existing VQA datasets.
- We conduct extensive experiments on MORE and propose a causality-enhanced method CAVE to mitigate the unimodal biases, providing insights for future work.

2 A Causal Framework

Inspired by Stolfo et al. (2023), we introduce a causal graph for MLLM predictions on VQA problems, highlighting language and vision biases and assessing their causal effects via controlled interventions (Pearl, 1995).

2.1 Problem Setup

We consider an entity-centric VQA problem, M, involving a question $Q := (S, T)$ and an image $I := (E, C)$. Here, S is the core semantic content of Q, and T is the textual form unrelated to Q's
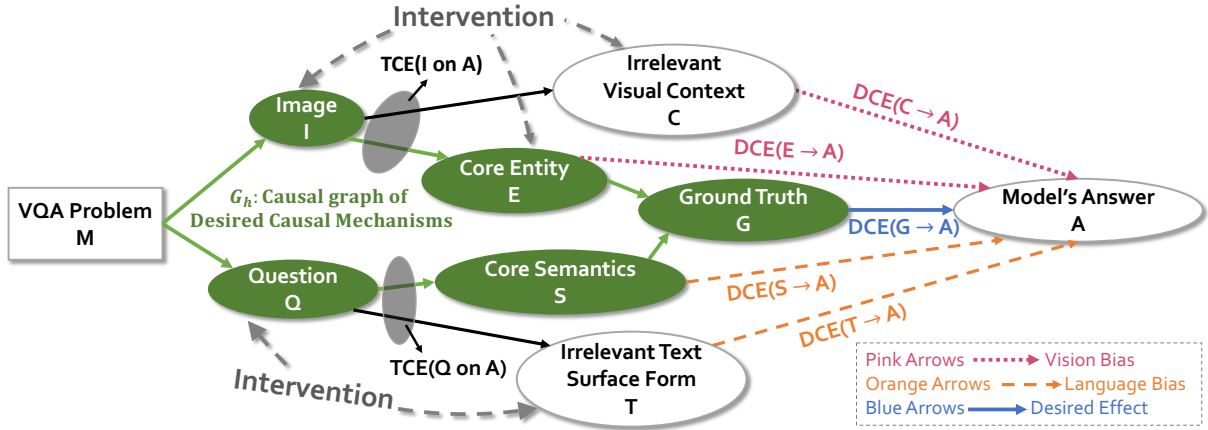
Figure 2: Causal graph of MLLM's Prediction on VQA problems. We use the green subgraph G_h to represent the desired causal mechanisms and compare it with the undesired effects of unimodal biases. We quantify the causal effects of each factor by performing controlled interventions of the images (I, E, C) and of the questions (Q, T).

main meaning. I includes the primary entity E and surrounding visual context C. The model outputs an answer, A. Notation: lowercase letters denote instances of corresponding uppercase variables.

2.2 Causal Graph of MLLM Predictions

Inspired by human cognitive processes (Zellers et al., 2019; Stolfo et al., 2023), we formulate the causal mechanisms in human problem-solving for a VQA problem m: $s = f_{c_1}(q)$, $e = f_{c_2}(i)$, $g = f_{c_3}(s, e)$. This involves decoding the question q to extract its semantic meaning s through cognitive process f_{c_1}, and identifying the core entity e from the image i through f_{c_2}. These are combined by f_{c_3} to produce the result g, as shown in the green subgraph G_h of Figure 2. In contrast, the model's approach to the same VQA problem employs $a = f_b(q, i)$, a black-box function where it is unclear how question and image inputs are utilized and interact to form the prediction a. For deeper analysis, we illustrate possible causal mechanisms in Figure 2. Notable mechanisms include:

Language Bias The model may directly process the question Q in two ways: by focusing on the core semantics S via $Q \rightarrow S \rightarrow A$, or on the irrelevant part T via $Q \rightarrow T \rightarrow A$. Both pathways lead to *language bias*, e.g., the focus on "the next World Cup" in Figure 1 (a).

Vision Bias The model may attend directly to the entity E of the image I via $I \rightarrow E \rightarrow A$, or to the irrelevant part C via $I \rightarrow C \rightarrow A$. Both pathways lead to the emergence of *vision bias*, e.g., the focus on the "*The Shard*" entity in Figure 1 (b).

Desired Causal Mechanisms Correct reasoning in VQA problems hinges on understanding causal mechanisms, as depicted in Figure 2. The subgraph G_h represents comprehension of how image and question jointly affect the ground-truth result G, through $E \rightarrow G$ and $S \rightarrow G$. This understanding should lead to model predictions that are sensitive and robust to changes in G, namely $G \rightarrow A$, with no spurious effects on A unless it passes the mediator G. Model performance could therefore be evaluated by its: 1) *Sensitivity*, which assesses how well the model adapts to changes in the correct answer, i.e., A responds to changes in G. 2) *Robustness*, which measures resistance to unimodal biases, e.g., $C \rightarrow A$ and $T \rightarrow A$, where minimal bias effects imply higher robustness to input changes that do not alter the ground-truth answer.

2.3 Causal Analysis of VQA Biases

We adopt the controlled interventions as outlined by Pearl (1995) to quantify the causal effects of questions and images on model predictions.

Causal Interventions for VQA 1) *Interventions on Q*. The question Q can be modified in two ways: (i) altering both S and T, or (ii) altering T but keeping S unaffected. 2) *Interventions on I*. The image I can be replaced with an alternative image I' in three ways: (i) altering both E and C, or (ii) altering C but keeping E unaffected, or (iii) altering E but keeping C unaffected. Note that we do not solely alter S within Q, because it is not feasible to intervene on the core semantics S of a question without affecting the surface text T of it.

Formulation of Causal Effects We assess the causal effects using $\mathrm{do}\,(X : x \to x')$, where $X \in \{Q, T, I, C, E\}$ is a factor in the VQA problem M. The pre-intervention probability distribution $\mathbb{P}(A \mid I, Q)$ is denoted as P, and the post-intervention distribution as P'. Following Pearl (1995), the causal effect is evaluated using a distance metric δ, as $\mathrm{CE} = \delta(P, P')$, where CE denotes the causal effect. It can refer to 1) the total causal effect (TCE), signifying the joint effect across all causal paths from one variable to another; or 2) the direct causal effect (DCE), indicating the effect of the directed causal path devoid of intermediary variables (Pearl, 2022). Following Stolfo et al. (2023), we assess the causal effect of factor X on the model's answer A by comparing the change in predicted results, $\delta_{\mathrm{cp}}(P, P') := \mathbb{I}(a \neq a')$, where $a = \arg\max_x P(x)$ and $a' = \arg\max_x P'(x)$, with $\mathbb{I}$ indicating a change in the answer.

Causal Effects of Questions We assess TCE of a question Q on an answer A by intervening on Q:

$$
\begin{aligned}
&\mathrm{TCE}(Q \text{ on } A) := \mathbb{E}_{q' \sim \mathbb{P}(Q)} \left[\delta\left(P, P'\right) \right], \\
&\text{where } P' = \mathbb{P}\left(A \mid I, \mathrm{do}\left(Q = q'\right)\right).
\end{aligned}
\tag{1}
$$

This TCE contains two different types of paths that show how Q affects A, as illustrated in Figure 2: **1)** The intended decision-making pathway: $Q \to S \to G \to A$, responding to changes in the ground truth. **2)** Potential spurious correlations, e.g., $Q \to T \to A$, where the model may depend on certain linguistic patterns from training data.

Maintaining the core semantics S constant, we can assess DCE of the textual surface T on A:

$$
\begin{aligned}
&\mathrm{DCE}(T \to A) := \mathbb{E}_{q' \sim \mathbb{P}(Q \mid S)} \left[\delta\left(P, P'\right) \right], \\
&\text{where } P' = \mathbb{P}\left(A \mid I, \mathrm{do}\left(Q = q'\right)\right).
\end{aligned}
\tag{2}
$$

As discussed in "Causal Interventions for VQA", solely intervening on S without altering T is impractical. However, Understanding S's causal impact on A is feasible by comparing two known quantities: $\mathrm{TCE}(Q \text{ on } A)$ and $\mathrm{DCE}(T \to A)$.

Causal Effects of Images The causal structure of images mirrors that of questions, as shown in Figure 2. Thus, we can derive TCE of image I on the answer as follows:

$$
\begin{aligned}
&\mathrm{TCE}(I \text{ on } A) := \mathbb{E}_{i' \sim \mathbb{P}(I)} \left[\delta\left(P, P'\right) \right], \\
&\text{where } P' = \mathbb{P}\left(A \mid Q, \mathrm{do}\left(I = i'\right)\right).
\end{aligned}
\tag{3}
$$

Likewise, maintaining E constant during each intervention on I allows us to quantify the DCE of the irrelevant visual context C on A:

$$
\begin{aligned}
&\mathrm{DCE}(C \to A) := \mathbb{E}_{i' \sim \mathbb{P}(I \mid E)} \left[\delta\left(P, P'\right) \right], \\
&\text{where } P' = \mathbb{P}\left(A \mid Q, \mathrm{do}\left(I = i'\right)\right).
\end{aligned}
\tag{4}
$$

Maintaining C and G constant during each intervention on I instead allows us to quantify the DCE of the core entity E on A:

$$
\begin{aligned}
&\mathrm{DCE}(E \to A) := \mathbb{E}_{i' \sim \mathbb{P}(I \mid C, G)} \left[\delta\left(P, P'\right) \right], \\
&\text{where } P' = \mathbb{P}\left(A \mid Q, \mathrm{do}\left(I = i'\right)\right).
\end{aligned}
\tag{5}
$$

Overall, calculating TCE helps us assess models' sensitivity (response to changes in ground truth), while DCE evaluates its robustness (stability of predictions against spurious correlations).

3 the MORE Dataset

From Section 3.1 to 3.3, we construct the MORE dataset that exploits the unimodal biases defined in Section 2.2 and requires multi-hop reasoning. In Section 3.4, we discuss how to quantify the causal effects of images and questions (defined in Section 2.3) using our constructed dataset.

3.1 Preparatory Steps

Images and Knowledge Graph Collection We begin with an existing VQA dataset, INFOS-EEK (Chen et al., 2023), which links image entities to Wikipedia information, requiring a VQA model to answer related questions. For instance in Figure 3, the entity "*Chrysler PT Cruiser*" prompts the question, "*What equipment or engine is used by this vehicle to provide power?*" Here, terms like "*Chrysler PT Cruiser*" and "*this vehicle*" all refer to the same entity in the image. We then identify all n-order neighbors ($n \in \{1, 2\}$) of the associated entity within a knowledge graph (KG), Wikidata5M (Wang et al., 2021), which is built upon Wikipedia data.

Subgraph Sampling Then, we identify a subgraph of an entity and its n-order neighbors in the KG. We filter paths in this subgraph that meet two criteria: 1) *Uniqueness of Paths*: the path from the associated entity to the selected neighbor is unique, and 2) *Shared-Type Relations*: they share a unique relation pointing to different entities. These criteria guarantee the uniqueness of the correct answer and introduce interference to challenge MLLMs' reasoning ability. For example, the filtered path in Figure 3 shows "*Fiat 500X*" linked to "*Chrysler PT Cruiser*" by a unique "*followed by*" relation, and
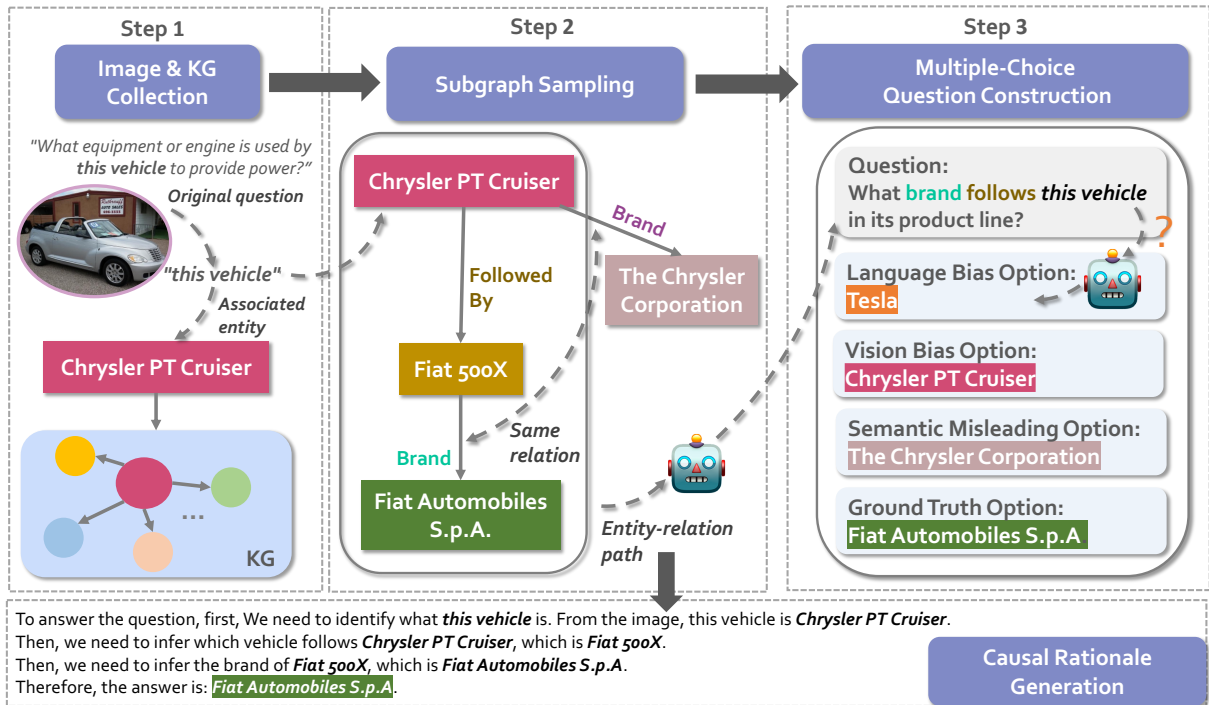
Figure 3: Our framework for generating data of MORE. We first prepare the image source and link the visual entity in a knowledge graph. Then, motivated by the visual and language bias analysis through the causal lens, we construct multiple-choice questions that require MLLMs to overcome unimodal biases and conduct multi-hop reasoning in a sampled subgraph. We also generate the causal (reasoning) rationale for each instance to provide interpretability.

both connected to their respective companies via a "*brand*" relation. This forms the multi-hop query: "*Chrysler PT Cruiser*" $\xrightarrow{\text{followed by}}$ "*Fiat 500X*" $\xrightarrow{\text{brand}}$ "*Fiat Automobiles S.p.A*". Entities in the subgraph may act as distractors to challenge MLLMs' reasoning, which will be discussed in the next section.

3.2 Multiple-Choice Question Construction

We detail the multiple-choice questions construction with four options, guided by the unimodal bias definition in Section 2.2.

Question Generation We generate questions by analyzing entity-relation paths within a specified subgraph, converting these paths into fluent, coherent queries with an LLM. We utilize the in-context learning (ICL) technique (Brown et al., 2020) and a standardized prompt (shown in Appendix A.1) to ensure quality. Among various LLMs tested, ChatGPT was selected for its superior multi-hop question generation. To avoid information leakage, entity names in the question are anonymized as "this <ENTITY_NAME>". For instance, a final generated question asks "*What brand follows this vehicle in its product line?*" in Figure 3.

Language Bias Option Language bias occurs when models overly attend to question-related information via pathways like $Q \to T \to A$. To assess this, we simulate conditions where only the question is provided and evaluate MLLMs' responses. We use the answers of GPT-4 (i.e., the *text-only* version of GPT-4V), to ensure uniform final options across MLLMs. For example, in Figure 3, GPT-4 reponds "*Tesla*" to the question. Additional results using different models for generating language bias options are detailed in Section 5.3 and the prompt template is available in Appendix A.2.

Vision Bias Option Vision bias occurs when visual information dominates (e.g., via $I \to E \to A$). We use the visually associated entity name (e.g., "*Chrysler PT Cruiser*") as an option, to see if the model directly selects it upon encountering an option that aligns with the visual information.

Semantic Misleading Option We introduce a semantic misleading option, such as "*The Chrysler Corporation*", to challenge MLLMs' multi-hop reasoning. This option refers to the entity that is pointed by the relation commonly owned by both the associated entity and its sampled neighbor. For example in Figure 3, upon encountering a question

16453

containing "*brand*" and "*Chrysler PT Cruiser*", MLLMs might simply output a direct answer (e.g., "*The Chrysler Corporation*"), ignoring other constraints in the question (e.g., "*followed by*"), hence struggling to choose the correct answer (e.g., "*Fiat Automobiles S.p.A.*").

Ground Truth Option Corresponding to the causal path via $E \rightarrow G$ and $S \rightarrow G$, this option represents the final entity in the entity-relation path (e.g., "*Fiat Automobiles S.p.A.*"). Finally, we check and ensure that each option is unique to prevent overlap samples.

Causal Rationale Generation Using entity-relation paths, we generate a causal rationale that aids in answering questions through a heuristic rule-based approach. As shown in Figure 3, this process begins with the associated entity and proceeds step-by-step to the ground truth. These rationales help confirm the accuracy of MLLMs' reasoning and improve their interpretability. They also can contribute to fine-tuning MLLMs for better multi-hop reasoning.

3.3 Dataset Statistics and Quality Analysis

Statistics of Different Hops We automatically generate training data from INFOSEEK's train set, and development/test data from its validation set, as shown in Table 3 in Appendix. Focusing on questions with 2-hop and 3-hop depths, we set $n = 1, 2$ respectively. We avoid longer-hop questions to prevent potential ambiguity and complexity.

Question Distribution We categorize the generated questions into distinct types based on their starting n-grams in Figure 15 in the Appendix. The MORE dataset showcases an extensive lexical diversity in the questions generated.

Question Quality We analyze the lexical diversity and fluency of the generated questions, with baselines and metrics detailed in Appendix C.1. From Figure 4, MORE shows superiority in lexical diversity and fluency, even compared to human-generated datasets.

Human Evaluation Our human evaluation confirms the high quality of generated questions and rationales, with 91% of questions and 98% of rationales deemed valid by annotators (details are in Appendix C.2).
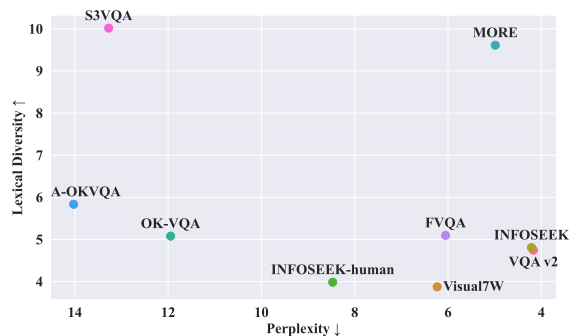

Figure 4: Question quality of MORE compared to other VQA datasets in terms of lexical diversity and fluency.

3.4 Causal Effect Calculation on MORE

Based on the analysis in Section 2.3, we discuss how to quantify the causal effects of unimodal biases on model predictions using MORE.

Causal Effect of Questions When intervening on questions, we keep the images constant: (1) For $\text{TCE}(Q \text{ on } A)$, we alter the question to another from the dataset pertaining to the same entity, changing both its text and core semantics, which shifts the ground truth. (2) For $\text{DCE}(T \rightarrow A)$, we let ChatGPT rephrase the question, altering its textual form but preserving its semantic meaning, resulting in no ground truth change post-intervention.

Causal Effect of Images When intervening on images, we keep the questions constant: (1) For $\text{TCE}(I \text{ on } A)$, we replace the image with another from the dataset corresponding to the same question but featuring different entities, altering both core entity and visual context to change the ground truth post-intervention. (2) For $\text{DCE}(C \rightarrow A)$, we replace the image with another that depicts the same entity, to keep the ground truth consistent after the intervention. Besides, it is practically challenging to calculate $\text{DCE}(E \rightarrow A)$. Because intervening on core entity E while maintaining the constancy of the question Q, visual context C, and ground truth answer G is hard to achieve, hence not considered in further analysis.

Given the impracticality of testing all perturbations of T and C, we randomly select 100 samples for each type of intervention and compute the average effects to determine TCE and DCE in Section 5.3. Overall, a higher TCE indicates better sensitivity, while a lower DCE indicates better robustness.

4 `CAVE` for Bias Mitigation

To mitigate unimodal biases in MLLMs and improve their reasoning capabilities, we propose `CAVE`, a causality-enhanced agent framework in this section.

For a given instance, `CAVE` starts by using a question decomposer to break down complex questions into simpler, step-by-step subquestions, explicitly avoiding spurious paths that may lead to incorrect answers. For each subquestion, `CAVE` uses a causality-enhanced reasoner to evaluate the correctness of the decomposed subquestion. Specifically, based on the aforementioned causal analysis in Section 2, it assesses whether the decomposed subquestion changes under two conditions: 1) rephrasing the question text, or 2) replacing the image with another captured by the same entity from a different angle or at a different time. If it identifies that the decomposed subquestion needs to be altered, this indicates that the previous understanding is influenced by irrelevant factors or biases, failing to capture the true semantics. In such cases, it will enforce a new round of question decomposition until it determines that the current subquestions are appropriate and accurate. Next, `CAVE` uses a verifier to strategically employ external tools and evaluate their outputs, acquiring the necessary context or information, such as image and text retrieval, to provide a precise answer. This iterative process of answering and verifying continues until all subquestions are resolved. By incorporating external knowledge, the final verified output integrates information from the entire reasoning process to provide a correct answer. A detailed illustration of `CAVE` and the prompt template is provided in Appendix E.

5 Experiments

5.1 Experimental Setup

Datasets We evaluate all test data from the `MORE` dataset using the **Multi-choice** settings, where MLLMs select answers from four provided options, with a random baseline accuracy of 25%.

Baselines We evaluate various leading MLLMs on our `MORE` dataset in a zero-shot fashion, including three limited-access MLLMs: GPT-4v, GPT-4o (OpenAI, 2023), and Gemini Pro Vision (Team et al., 2023), and four open-source MLLMs: InstructBLIP (13B) (Dai et al., 2023), mPLUG-Owl (7B) (Ye et al., 2023), LLaVA (v1.5, 13B) (Liu

et al., 2023a), and Qwen-VL (7B) (Bai et al., 2023b) the details are in Appendix D. For consistent evaluation, we use standard accuracy metrics for all the models. We further quantify the causal effects of images and questions on model predictions following Section 5.3.

5.2 Evaluation Results

The results of MLLMs on `MORE` are shown in Table 2 and further exemplified in Appendix F. We observe that:

1) All baselines perform poorly on `MORE` (e.g., even the best-performing model, Gemini Pro Vision, achieves only 22.3% accuracy, which does not surpass the random baseline), indicating MLLMs' vulnerability to biases.

2) Most models perform better on two-hop data than on three-hop data, suggesting that MLLMs' reasoning capabilities are challenged when the problems become more complex.

3) GPT-4v falls short versus Gemini Pro Vision, possibly because we use homologous GPT-generated distractors when constructing the language bias options in Section 3.2, which particularly challenges GPT-4v's judgment. This point is further analyzed in Section 5.3.

4) Our proposed `CAVE` significantly enhances GPT-4o and Gemini Pro Vision on `MORE`, validating the effectiveness of our method. However, the relatively low absolute values indicate ongoing challenges related to biases, suggesting the need for further research efforts.

5.3 Causal Analysis of VQA Biases

In this subsection, we select several representative MLLMs and analyze their performance through a causal lens.

Option Distribution In Figure 5, we show the option distribution of selected MLLMs. Here we mainly utilize language bias options generated by GPT-4v and Gemini Pro Vision for comparison because they explicitly provide corresponding text-only versions. The observations are as follows:

1) *Severe unimodal biases.* More than 40% of the options show either language or vision bias in all models, underscoring the prevalence of unimodal biases.

2) *Limited understanding ability.* Models' selection of semantically misleading options indicates some ability to combine visual and textual information, though not fully grasping the problem. This

Model	LLM	# Params	Two-Hop, acc (%)	Three-Hop, acc (%)	Overall, acc (%)
Random	/	/	25.0	25.0	25.0
InstructBlip	Vicuna	13B	17.0	16.2	16.6
mPLUG-Owl	Llama	7B	12.4	11.4	11.9
LLaVA	Llama	13B	20.8	13.6	17.5
Qwen-VL	Qwen	7B	17.4	15.6	16.5
GPT-4v	-	-	17.3	16.0	16.5
GPT-4o	-	-	18.0	17.0	17.5
GPT-4o + CAVE	-	-	33.7	27.3	30.5
Gemini Pro (V)	-	-	25.4	20.7	22.3
Gemini Pro (V) + CAVE	-	-	**35.6**	**28.8**	**33.2**

Table 2: MLLMs' results on the test set of MORE. We report the VQA accuracy (%) under the multi-choice settings on two-hop, three-hop, and all data, respectively."-" denotes not released information.
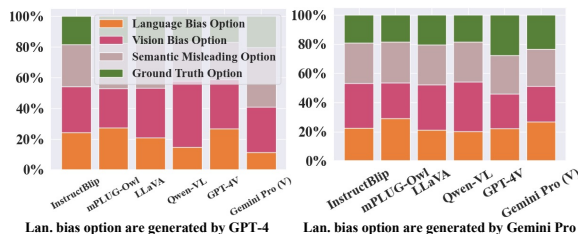

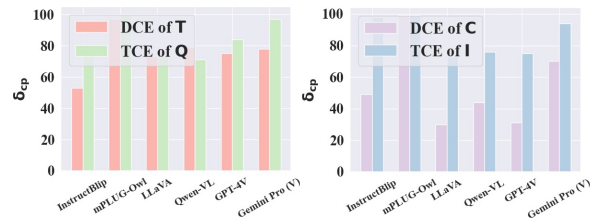Figure 5: Option distribution of MLLMs.

Figure 6: Comparison of direct and total effects of image and question on prediction for MLLMs.

highlights the challenge our MORE dataset poses to current MLLMs.

3) *Obvious selection tendency*. GPT-4v often incorrectly chooses language bias options generated from GPT-4 (i.e., the text-only version of GPT-4v). Switching to Gemini Pro (i.e., the text-only version of Gemini Pro Vision) shifts this trend, with GPT-4v's language bias selections decreasing and Gemini Pro's increasing. These observations align with our prior analysis of GPT-4v.

Please note that discrepancies may exist between the proportions of ground truth options presented here and the accuracy values reported in Table 2, as some models' outputs may not align with the provided option format (e.g., mPLUG-Owl), thus affecting the count of valid answers.

Causal Effects of Images and Questions To further analyze the impact of unimodal biases on the model predictions, we assess the causal effects based on the discussion in Section 3.4. As shown in Figure 6, we select several representative MLLMs:

1) Current MLLMs exhibit high sensitivity (high TCE), a possible reason is that instruction tuning makes models sensitive to variations in in-

put (Stolfo et al., 2023).

2) However, the model's robustness is relatively low, as indicated by a high DCE, meaning its predictions fluctuate even when the ground truth remains constant. This phenomenon becomes particularly pronounced when irrelevant text surface forms are introduced. This suggests that, although MLLMs can adapt well to answers, they exhibit a lack of robustness when exposed to interference, relying on spurious correlations rather than genuine causal features.

6 Related Work

Multimodal Large Language Models (MLLMs) Recent advances in large language models (LLMs) have led to the emergence of MLLMs, which demonstrate exceptional performance in multimodal tasks (OpenAI, 2023; Team et al., 2023; Liu et al., 2023a; Hu et al., 2023; Yu et al., 2024). However, in contrast to the extensive evaluation of reasoning capabilities in LLMs (Wei et al., 2022; Chen et al., 2024a), currently, the evaluation of MLLMs primarily emphasizes basic visual tasks (Liu et al., 2023b; Fu et al., 2023; Lu et al., 2024;

Chen et al., 2024b), with limited investigation into their reasoning capabilities.

Knowledge-based VQA Datasets Existing VQA datasets (Wang et al., 2017; Marino et al., 2019; Chen et al., 2023) are limited by their focus on image-related information, lack of multi-hop reasoning, open-ended answers, and reasoning rationales. They also fail to measure the effect of language and vision biases. Our MORE dataset, outlined in Table 1, addresses these shortcomings by providing a more comprehensive assessment.

Language and Vision Biases in VQA Research reveals that some VQA models primarily depend on statistical priors from training data instead of genuinely comprehending image content (Agrawal et al., 2018). These models exhibit language and vision biases; the former arises from strong correlations between specific questions and answers (Abbasnejad et al., 2020; Zhu et al., 2020), and the latter from frequent co-occurrences of textual and visual elements in the dataset (Si et al., 2022; Gupta et al., 2022). Recent efforts to address these biases mostly involve data augmentation (Niu et al., 2021). Besides, several studies (Rohrbach et al., 2018; Parcalabescu et al., 2022; Parcalabescu and Frank, 2023) use examples derived from language biases to assess object hallucination in models. However, these constructed examples are typically counterfactual, with the answer types limited to yes/no, essentially judging whether the constructed statements are correct or incorrect.

7 Conclusion

This paper presents a comprehensive approach to quantifying and mitigating the unimodal biases in MLLMs. Through our causal inference framework, we provide an in-depth analysis to assess the causal effects of such biases on the model's prediction in VQA problems. The introduced MORE dataset challenges MLLMs to engage in multi-hop reasoning and to overcome language and vision biases, thereby pushing the boundaries of their reasoning capabilities. Our proposed CAVE method demonstrates significant potential in enhancing the reasoning abilities of MLLMs.

Limitations

Our current generation of rationales is based on heuristic rules. Previous works have demonstrated the effectiveness of incorporating rationales into instructions (Wei et al., 2022). Therefore, we believe that refining and polishing these rationales with an LLM (e.g., ChatGPT) could be beneficial. Besides, the Wikidata5M dataset we employed was released in 2021, and some information in the knowledge graph may be outdated. Although we have made efforts to manually verify the test set and try to ensure it does not contain incorrect information, it is still inevitable that errors may occur within the extensive training data.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. 2020. Counterfactual vision and language learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10041–10051. IEEE.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. IEEE Computer Society.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024a. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. 2024b. Cello: Causal evaluation of large vision-language models. *arXiv preprint arXiv:2406.19131*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint*, abs/2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Tianrui Guan, Fuxiao Liu, Xiyang Wu Ruiqi Xian Zongxia Li, Xiaoyu Liu Xijun Wang, Lichang Chen Furong Huang Yaser Yacoob, and Dinesh Manocha Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv e-prints*, pages arXiv–2310.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan L. Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5068–5078. IEEE.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, et al. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint*, abs/2307.06281.

16458

Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, et al. 2024. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities. *arXiv preprint arXiv:2401.15071*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. 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*, pages 3195–3204. Computer Vision Foundation / IEEE.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. 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*, pages 12700–12710. Computer Vision Foundation / IEEE.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280, Dublin, Ireland. Association for Computational Linguistics.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Lucas Shen. 2022. Lexicalrichness: A small module to compute textual lexical richness.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3698–3712, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561, Toronto, Canada. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *ArXiv preprint*, abs/2312.11805.

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. 2023a. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

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

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. PaperRobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *ArXiv preprint*, abs/2309.17421.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv preprint*, abs/2304.14178.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1083–1089. ijcai.org.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

Question Generation

```
Task Description:
Provide a question according to the starting entity and path (relations split by ',') in a knowledge graph.

Template:
Starting entity: <HEAD_ENTITY>
Path: <RELATION_PATH_IN_KG>

Examples:

Starting entity: Coca-Cola
Path: discoverer or inventor, place of birth
Generated Question: Which city is the birthplace of the inventor of Coca-Cola?

Starting entity: James Berkeley
Path: place of birth, owned by
Generated Question: Who is the owner of the building where James Berkeley was born?

Starting entity: Conwy Castle
Path: country, highest point, material used
Generated Question: What material is used in the highest point of the country where Conwy Castle is located?
```

Figure 7: Prompt template of multi-hop question generation.

Language Bias Option Generation

```
Task Description:
Given a question, provide a specific answer.

Examples:

Question: What is the parent taxon of the main food source of this animal?
Answer: Animalia

Question: What is the heritage designation of the burial place of the person who commissioned this building?
Answer: UNESCO World Heritage Site.
```

Figure 8: Prompt template of language bias option generation.

Dataset	#I, Q, A	Len of Q / A	# Ent
MORE-train	10,000	14.3 / 2.1	1,261
- 2-hop	4,134	11.6 / 2.0	886
- 3-hop	5,866	16.1 / 2.2	686
MORE-dev	1,000	13.8 / 2.3	118
- 2-hop	548	12.2 / 2.2	71
- 3-hop	452	15.8 / 2.5	73
MORE-test	1,000	13.9 / 2.4	251
- 2-hop	500	12.3 / 2.2	153
- 3-hop	500	15.6 / 2.6	143

Table 3: Dataset statistics of different hops.

A Prompt Templates

A.1 Question Generation

We present the prompt template for generating language bias options of Section 3.2 in Figure 7.

A.2 Language Bias Option Generation

We present the prompt template for generating language bias options of Section 3.2 in Figure 8.

B Question Distribution

In Figure 15, we categorize the generated questions into distinct types, based on their starting n-grams. The dataset MORE showcases an extensive lexical diversity in the questions generated. This diversity is evidenced by variations in the introductory interrogative words (e.g., "what", "who", "where", etc.), exemplified by phrases like "What is the...", "In which country...", and more. Such lexical richness is crucial for mitigating the vulnerability of MLLMs to linguistic variations.

C Quality Analysis Details

C.1 Question Quality

To ensure the quality of the comprising datasets, we analyze the lexical diversity and the fluency of the generated questions, which are useful for conducting a robust evaluation using questions that are linguistically diverse and coherent.

Baselines We select extensive VQA datasets for comparison, including Visual7W (Zhu et al., 2016), VQA (v2) (Goyal et al., 2017), FVQA (Wang et al., 2017), OK-VQA (Marino et al., 2019),
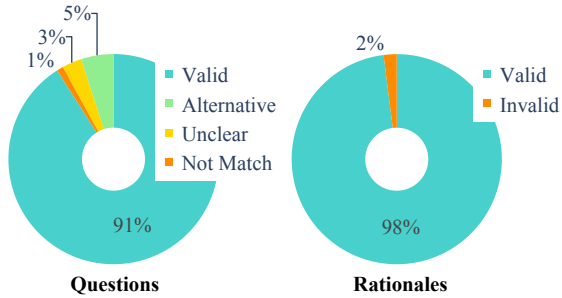
Figure 9: Human evaluation results of MORE.

S3VQA (Jain et al., 2021), A-OKVQA (Schwenk et al., 2022), and INFOSEEK (Chen et al., 2023) (contains both automated generation version and human-annotated version).

Evaluation Metrics For lexical diversity, we utilize three metrics that are not dependent on length: moving average type-token ratio (MATTR) (Covington and McFall, 2010), measure of textual lexical diversity (MTLD) (McCarthy, 2005), and hypergeometric distribution diversity (HDD) (McCarthy and Jarvis, 2010). We average these three metrics for a unified assessment and employ the Lexical-Richness package (Shen, 2022) (version 0.5.03) for calculation. For fluency, we employ a pre-trained language model GPT2-large (Radford et al., 2019) with 774M parameters to compute the perplexity of the questions, which is often used as a measure by previous work (Wang et al., 2019; Cahyawijaya et al., 2021).

C.2 Human Evaluation

Questions We conduct a human evaluation of 100 questions randomly chosen from the MORE dataset to validate and assess the quality of the generated questions. This evaluation is carried out by three human annotators independently, who are provided with detailed guidelines and illustrative examples before starting the evaluation process. For each question, given the visual context and ground truth answer, we first ask two junior annotators to determine whether: 1) the question is valid, 2) the question allows for an alternative answer, 3) the question does not match the answer, or 4) the question is unclear or ambiguous. If the choices of the two annotators are inconsistent, a senior annotator checks the answers and makes the final decision. The average inter-annotator agreement is 88.6% (Cohen's kappa).

Rationales We also conduct a human evaluation of the causal rationales, following the same procedure as described above. The difference is that here we provide annotators with only two options: to assess whether the generated rationales are valid.

As shown in Figure 9, the results are encouraging, with 91% questions and 98% rationales being classified as valid by the annotators, further demonstrating the quality of our datasets.

D Baselines

For open-source MLLMs, we consider the following baselines:

1) InstructBLIP (Dai et al., 2023), an instruction-tuned version of BLIP-2 on various tasks including VQA. We employ its InstructBLIP-Vicuna (Chiang et al., 2023)-13B variant.

2) mPLUG-Owl (Ye et al., 2023), which proposes a new two-stage training method for aligning images and text. We employ its mPLUG-Owl-Llama (Touvron et al., 2023a)-7B variant.

3) LLaVA (Liu et al., 2023a), which translates images into texts of captions and bounding boxes, and prompts GPT-4 to generate a multimodal instruct-tuning dataset in the context of seed examples. We employ its LLaVA-Llama (Touvron et al., 2023a)-13B variant.

4) Qwen-VL (Bai et al., 2023b), which builds upon Qwen (Bai et al., 2023a) and employ multi-stage training pipeline. Qwen-VL facilitates grounding and text comprehension by aligning image-caption-box tuples. It processes inputs of images, text, and bounding boxes, and produces corresponding text and bounding boxes as outputs.

E Details of the CAVE Framework

In this section, we first present the overall framework as shown in Figure 10, and then we will go over each part of it in detail.

E.1 Overall Framework

Given a question Q and an image I, the VQA task demands the system to return an output A that concisely answers the question. As shown in Figure 10, we first initialize a question decomposer D to analyze Q and break it down into manageable subquestions. For each subquestion, we introduce a causality-enhanced reasoner C to evaluate the correctness of the decomposed subquestion. Then, we employ a verifier V to confirm the accuracy of the

Input

Question: Which country is hosting the next World Cup after this venue?

Question Decomposition

To answer the question, first, we need to identify *what this venue is*.

Causality-enhanced reflection

If this image is replaced with shot at **the same venue at a different time or from a different angle**, will the sub-questions I have decomposed change?

Verification

Image Retrieval: Original Image.

Question Decomposition

Based on the information provided, the venue is Allianz Arena. Due to the Allianz Arena hosted the 2006 World Cup, the question becomes *which country hosted the next World Cup after Allianz Arena*.

Causality-enhanced reflection

If the question becomes *"Which country is hosting the immediately following World Cup after this venue?"*, will the sub-question I have decomposed change?

Verification

Text Retrieval: Which country hosted the 2010 World Cup?

Final Output

Based on the information provided, **South Africa** hosted the 2010 World Cup.

Image Retrieval

Captions of retrieved images:
[1] Image of Allianz Arena.
[2] Allianz Arena Muenchen
[3] Allianz Arena Munich Germany
...

Text Retrieval

Contents of retrieved documents:
[1](Title: 2010 FIFA World Cup): 2010 FIFA World Cup The 2010 FIFA World Cup was the 19th FIFA World Cup, the world championship for men's national association football teams. It took place in South Africa from 11 June to 11 July 2010.
[2](Title: 2010 FIFA World Cup): that he and the other members of FIFA's executive committee were bribed in order to promote the South African 1998 and 2010 World Cup bids. Blazer stated, "I and others on the Fifa executive committee agreed to accept bribes in conjunction with the selection of South Africa as the host nation for the 2010 World Cup.
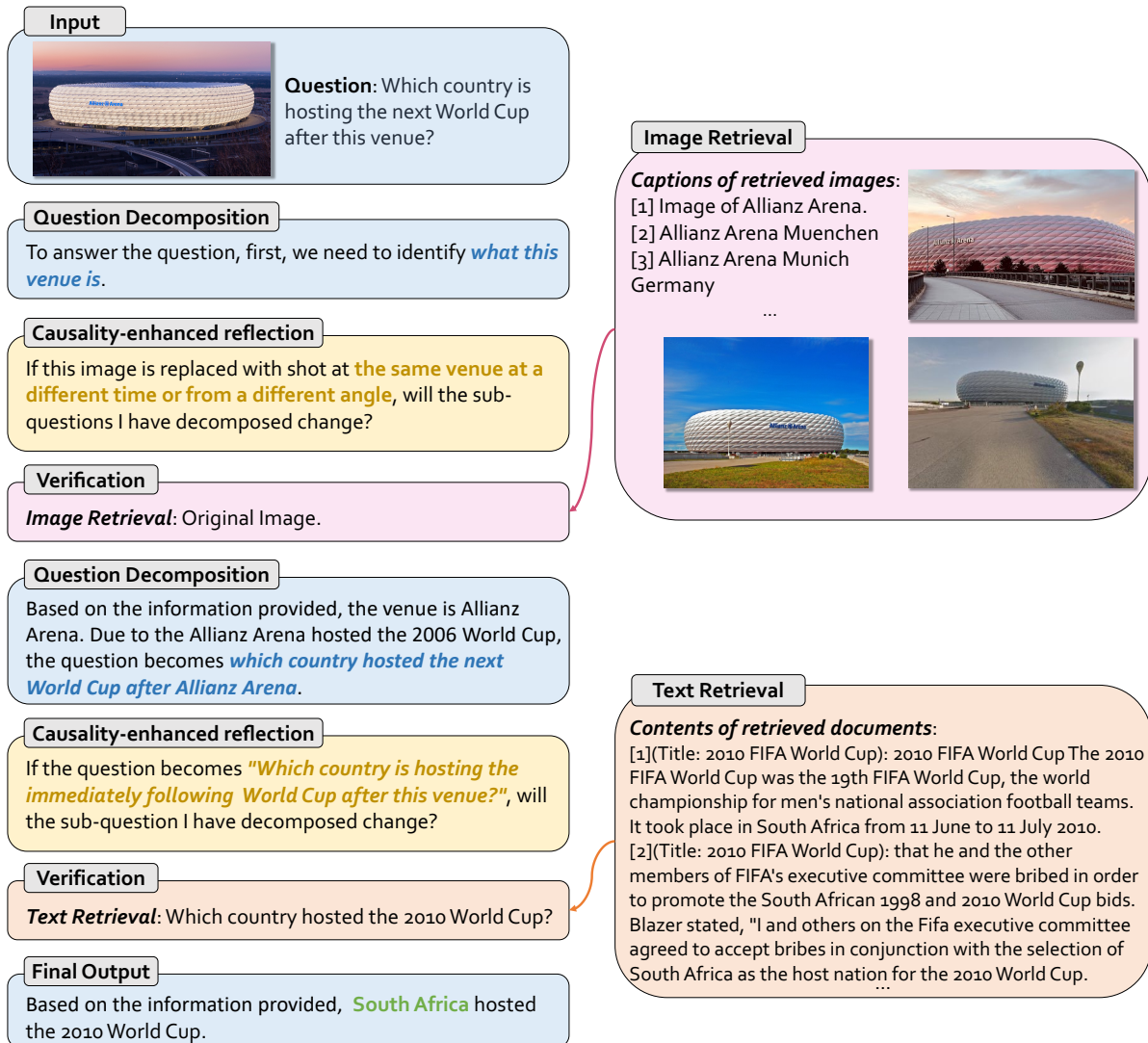...

Figure 10: An overview of our proposed CAVE Framework.

original answer to each subquestion. The generation verifier typically involves active information-seeking and answer-verification, which acquires the necessary context or information needed to investigate and revise the answers. This includes two optional operations: image retrieval to seek images similar to I and determine their titles, or text retrieval with a specific query to fetch pertinent documents and summarize their content. This iterative process of answering and verifying will continue until we resolve each subquestion. Finally, the verified answer A is output following the aforementioned reasoning process and retrieved information.

E.2 Question Decomposer

For a given VQA problem, MLLMs may simply exploit a spurious shortcut to make predictions. In order to alleviate this issue, motivated by Chain-of-Thought reasoning (Wei et al., 2022), we encourage MLLMs to decompose the question Q before outputting the answer, so as to gradually solve a complex question that requires multi-hop reasoning. As shown in Figure 10, for the question "*Which country hosted the next World Cup after this venue?*", our decomposer breaks it down into two subquestions :

1. "*What this venue is*?

2. "*Which country hosted the next World Cup after Allianz Arena*?

Such decomposition will explicitly constrain the model to comprehend and extract the truth semantics of the question, thus avoiding simply exploring a spurious path to give the answer.

E.3 Causality-enhanced Reasoner

The accuracy of sub-question decomposition plays a crucial role in the subsequent reasoning process. To address this, we introduce causality-enhanced self-reflection, which explicitly guides the model in validating the sub-question it has decomposed. Specifically, based on the proposed causal framework, we guide the model to assess whether the decomposed subquestion changes under two conditions: 1) rephrasing the question text, or 2) replacing the image with another captured by the same entity from a different angle or at a different time. If the model identifies that the decomposed subquestion needs to be altered, this indicates that its initial understanding is influenced by irrelevant factors or biases, failing to capture the true semantics. In such cases, we will enforce a re-decomposition of the question.

E.4 Verifier

Some works have found that vision illusion and language hallucination may appear in the process of MLLMs' response generation (Guan et al., 2023; Yang et al., 2023). To alleviate this issue, we adopt the retrieval-augmented generation approaches (Khandelwal et al., 2020; Chen et al., 2022). Specifically, we consider two different retrieval ways for the verifier to choose during each verification step: image retrieval and text retrieval.

Image Retrieval   Although our framework is applicable to any image retrieval method, in this paper, we mainly utilize Google Image Search to obtain a broad range of information related to the image as provided by Google Lens API[1]. This information encompasses various details, such as knowledge graph entities and captions of analogous or identical images. The availability of these details can vary based on the image input provided to Google Image Search. Then, the verifier gleans relevant information from captions associated with visually similar images, so as to verify the original answer and conduct the next round of reasoning.

Text Retrieval   Similarly, our framework is applicable to any text retrieval method, we explore a simple, off-the-shelf dense retriever for Wikipedia, GTR (Karpukhin et al., 2020), as our text retriever. First, the verifier constructs a query to perform text retrieval according to the currently generated context, and then the query is input into a GTR model to get related document titles and contents. Finally, the verifier will fetch pertinent documents and summarize their content to verify the immediate answer.

E.5 Final Verified Answer

Finally, the improved response that takes verification into account is generated. This is executed by a final prompt where the context takes into account all of the previous reasoning steps, the baseline response, and the verification question-answer pairs, so that the corrections can take place.

F Case Study

We conduct a case study on the development set of MORE in Figure 11~14, including both the "Open-ended" and "Multi-Choice" settings.

---

[1]Web interface available at https://images.google.com.

2-hop case

Chrysler PT Cruiser —Followed By→ Fiat 500X

Chrysler PT Cruiser —Brand→ The Chrysler Corporation

Fiat 500X —Brand→ Fiat Automobiles S.p.A.

Question: What brand follows this vehicle in its product line?

BLIP2 — Vision Bias

The chrysler pt cruiser

InstructBLIP — Vision Bias

chrysler

mPLUG-Owl — Visual Illusion

The vehicle follows the brand BMW in its product line.

LLaVA — Ground Truth

Fiat

GPT-4V — Semantic Misleading

The vehicle in the image is a Chrysler PT Cruiser Convertible. The Chrysler PT Cruiser is a model from the Chrysler brand, which is part of Stellantis, a multinational automotive manufacturing corporation.

Gemini Pro Vision — Semantic Misleading

The Chrysler PT Cruiser is a front-wheel drive, four-passenger convertible that was produced by Chrysler from 2001 to 2010. The PT Cruiser was based on the Chrysler Neon platform and was available in two trim levels: the base model and the GT. The PT Cruiser was powered by a 2.4-liter four-cylinder engine that produced 150 horsepower and 165 lb-ft of torque. The PT Cruiser was discontinued after the 2010 model year.

Figure 11: Two-hop case in the dev set of MORE. The wrong answers are marked in red and the correct is in green.

Chrysler PT Cruiser —Followed By→ **Fiat 500X**

Fiat 500X —Brand→ **Fiat Automobiles S.p.A.**

Chrysler PT Cruiser —Brand→ **The Chrysler Corporation**

Question: What brand follows this vehicle in its product line?
Choose from the following four options:
A. Chrysler PT Cruiser   B. Tesla
C. The Chrysler Corporation   D. Fiat Automobiles S.p.A.

BLIP2 — Vision Bias

The Chrysler PT Cruiser

InstructBLIP — Vision Bias

Chrysler PT Cruiser

mPLUG-Owl — Vision Bias

Chrysler PT Cruiser
The brand of the car that follows the model which this vehicle follows is Chrysler PT Cruiser. This is the correct answer.

LLaVA — Vision Bias

Chrysler PT Cruiser

GPT-4V — Semantic Misleading

The vehicle in the image is a Chrysler PT Cruiser, which is a model produced by Chrysler. Therefore, the brand that follows the model which this vehicle follows is Chrysler.

Gemini Pro Vision — Vision Bias

Chrysler PT Cruiser

Figure 12: Two-hop case (multi-choice setting) in the dev set of MORE. The wrong answers are marked in red.

3-hop case

Continent — **Salève** → **Europa** — Highest point

Mountain range ↓ → **Elbrus**

↓ Mountain range

French Prealpes   **Caucasus Major**

Question: What is the mountain range of the highest point on the continent where this mountain is located?

BLIP   **Visual Illusion**

The alaska range

InstructBLIP   **Visual Illusion**

rocky

---

mPLUG-Owl   **Visual Illusion**

The mountain of Mount McKinley, also known as Denali, is located in Alaska, USA. It is the tallest mountain in North America and the tallest mountain in the world when measured from the base of its sea-level summit. The mountain range that Mount McKinley is a part of is the Alaska Range.

LLaVA   **Language Bias**

The highest mountain range is the Alps, the highest peak is Mont Blanc.

GPT-4V   **Semantic Misleading**

The mountain in the image is the Salève. The Salève is a mountain located in the Prealps, near Geneva, Switzerland, and France.

Gemini Pro Vision   **Visual Illusion**

The mountain range is the Chugach Mountains, which are part of the larger Coast Mountains.

Figure 13: Three-hop case in the dev set of MORE. The wrong answers are marked in red.

3-hop case (Option)

Salève —Continent→ Europa —Highest point→ Elbrus
Salève —Mountain range→ French Prealpes
Elbrus —Mountain range→ Caucasus Major

Question: What is the mountain range of the highest point on the continent where this mountain is located?
Choose from the following four options:
A. Salève   B. the Himalayas
C. French Prealps   D. Caucasus Major

BLIP2   **Language Bias**

 The Himalayas

- -

InstructBLIP   **Semantic Misleading**

 French Prealps

- -

mPLUG-Owl   **Vision Bias**

The mountain range of the highest point on the continent where this mountain is located is Salève.
Salève, which is located in the Rhône-Alpes region of France, is a small mountain range that extends from the city of Geneva to the city of Annecy. The summit of this range, Mont Salève, is the highest point on the continent where this mountain is located. Salève is a popular destination for hiking and skiing, and it offers spectacular views of the surrounding Alps, including the Mont Blanc massif and the Swiss Alps. It is also a popular destination for rock climbing, and the steep cliffs and rocky faces of Salève provide a challenging climbing experience.

- -

LLaVA   **Ground Truth**

 Caucasus Major

- -

GPT-4V   **Language Bias**

Based on the options provided, the highest point on the continent where the Salève mountain is located would be Mount Everest, as Salève is part of the Prealps, a region in Europe, and Mount Everest is the highest point on the Asian continent. The Greater Caucasus Mountains are also in Asia but are not as high as Mount Everest. Therefore, the correct answer from the options given is the Himalayas.

- -

Gemini Pro Vision   **Semantic Misleading**

The highest point of the mountain range is Mont Blanc, which is located in the French Prealps. The French Prealps are located in Europe.

Figure 14: Three-hop case (multi-choice setting) in the dev set of MORE. The wrong answers are marked in red and the correct is in green.
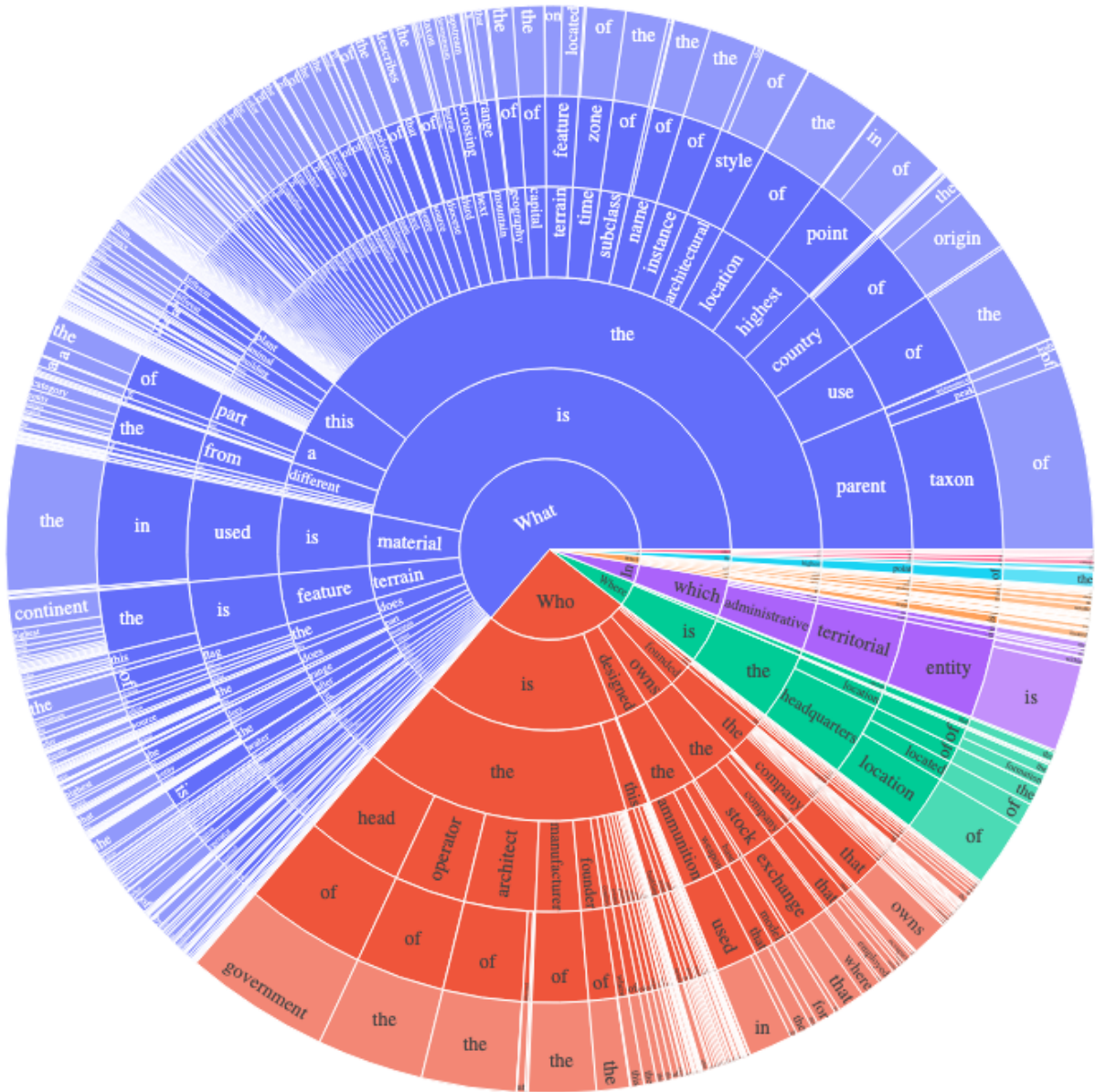
Figure 15: Question prefix distribution in MORE. The arc length is proportional to the number of questions containing the word.