# CELLO 🎻: Causal Evaluation of Large Vision-Language Models

**Meiqi Chen**[1,2*], **Bo Peng**[3,4], **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]Shanghai Jiao Tong University
[4]Shanghai Artificial Intelligence Laboratory
meiqichen@stu.pku.edu.cn, peng_bo2019@sjtu.edu.cn,
zhyzhy001@pku.edu.cn, luchaochao@pjlab.org.cn

## Abstract

Causal reasoning is fundamental to human intelligence and crucial for effective decision-making in real-world environments. Despite recent advancements in large vision-language models (LVLMs), their ability to comprehend causality remains unclear. Previous work typically focuses on commonsense causality between events and/or actions, which is insufficient for applications like embodied agents and lacks the explicitly defined causal graphs required for formal causal reasoning. To overcome these limitations, we introduce a fine-grained and unified definition of causality involving interactions between humans and/or objects. Building on the definition, we construct a novel dataset, CELLO, consisting of 14,094 causal questions across all four levels of causality: discovery, association, intervention, and counterfactual. This dataset surpasses traditional commonsense causality by including explicit causal graphs that detail the interactions between humans and objects. Extensive experiments on CELLO reveal that current LVLMs still struggle with causal reasoning tasks, but they can benefit significantly from our proposed CELLO-CoT, a causally inspired chain-of-thought prompting strategy. Both quantitative and qualitative analyses from this study provide valuable insights for future research. Our project page is at https://github.com/OpenCausaLab/CELLO.

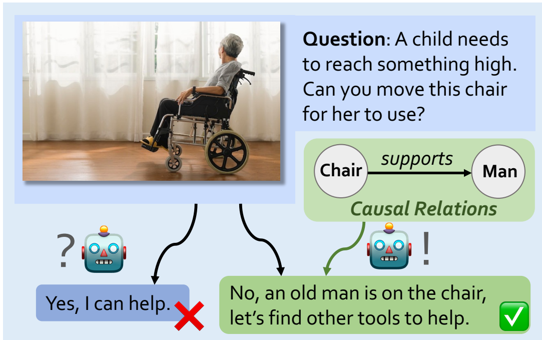



Figure 1: An example of causal reasoning in the vision-language context. LVLMs (e.g., GPT-4o) might generate inappropriate responses due to a limited understanding of causal relationships.

## 1 Introduction

Causal reasoning is recognized as a fundamental component of human intelligence (Penn and Povinelli, 2007; Harari, 2014). Recent advances in large language models (LLMs) have promoted a surge of research successfully adapting LLMs to vision-language tasks, resulting in powerful large vision-language models (LVLMs) (OpenAI, 2023; Liu et al., 2023a). Despite these advancements, a critical question arises: *Do LVLMs really understand causality?*

Previous work has primarily focused on commonsense causality between events and/or actions in a vision-language context (Zellers et al., 2019; Park et al., 2020; Kim et al., 2022), but often neglects the fine-grained causal relationships between humans and objects, between humans, and between objects. This limits the effectiveness of decision-making in real-world environments, such as embodied intelligent agents (Cheong et al., 2024; Gupta et al., 2024) and autonomous driving systems (Ramanishka et al., 2018). For instance, as illustrated in Figure 1, a model might respond "*yes*" to the request, "*A child needs to reach something high. Can you move this chair for her to use?*" This response overlooks the critical human-object causal relationship that "*the chair supports an old man*"[1], which would lead to a more reasonable decision. Moreover, these studies typically do not explicitly define the underlying causal graphs for key entities, rendering it challenging to systematically investigate the formal causal reasoning ability of LVLMs.

*This work was done during her internship at Shanghai AI Laboratory.
†Corresponding author.

[1]The human-object causal relationship will be detailed in Section 3.



*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22353–22374
November 12-16, 2024 ©2024 Association for Computational Linguistics

| Datasets | Question Types Disc. | Assoc. | Interv. | CF. | Fine-grained Causality | Answer Type | Rationale | # Size |
|---|---|---|---|---|---|---|---|---|
| Visual7W (Zhu et al., 2016) | ✓ | ✗ | ✗ | ✗ | ✗ | Open-ended | ✗ | 8,884[1] |
| VQA (v2) (Goyal et al., 2017) | ✓ | ✗ | ✗ | ✗ | ✗ | Open-ended | ✗ | 1,952[1] |
| FVQA (Wang et al., 2017) | ✓ | ✗ | ✗ | ✗ | ✗ | Open-ended | ✔ | 17[1] |
| OKVQA (Marino et al., 2019) | ✓ | ✗ | ✗ | ✗ | ✗ | Open-ended | ✗ | 115[1] |
| VCR (Zellers et al., 2019) | ✓ | ✗ | ✗ | ✗ | ✗ | Multi-choice | ✔ | 9390[1] |
| VisualCOMET (Park et al., 2020) | ✓ | ✗ | ✗ | ✗ | ✗ | Open-ended | ✔ | 13,768[1] |
| BD2BB (Pezzelle et al., 2020) | ✗ | ✗ | ✓ | ✗ | ✗ | Multi-choice | ✗ | 10,000 |
| COSIM (Kim et al., 2022) | ✗ | ✗ | ✓ | ✗ | ✗ | Multi-choice | ✗ | 3,500 |
| NORMLENS (Han et al., 2023) | ✗ | ✗ | ✓ | ✗ | ✗ | Multi-choice | ✗ | 10,000 |
| **CELLO (Ours)** | ✔ | ✔ | ✔ | ✔ | ✔ | Multi-choice | ✔ | 14,094 |

Table 1: Comparison of CELLO with existing causality-related vision-language datasets. Under the "Question Types" column, the abbreviations "Disc.", "Assoc.", "Interv.", and "CF." represent the four causal levels: Discovery, Association, Intervention, and Counterfactual, respectively. "✗" denotes the absence of causality, "✓" denotes commonsense causality, and "✔" denotes both commonsense and formal causality (with causal graph).

To address this, we first introduce a fine-grained and unified definition of causality in the vision-language context, drawing inspiration from the concept of *causal dispositions* (Mumford and Anjum, 2011; Lopez-Paz et al., 2017). We define a causal relationship as existing when one entity inherently possesses the ability to influence the state of another entity. This relationship can be further clarified through counterfactual reasoning (Pearl, 2009; Peters et al., 2017): if the "*cause*" entity were absent, the "*effect*" entity would not sustain its current state. This includes interactions such as "support" and "hold", as well as spatial positioning between *humans and humans*, *humans and objects*, and *objects and objects*. Using this foundational definition, we extract corresponding causal graphs from scene graphs in existing vision-language datasets and formulate questions based on these graph types. This results in CELLO, a novel dataset consisting of 14,094 causal questions across all four causal rungs of the *Ladder of Causation*[2] (Pearl and Mackenzie, 2018; Bareinboim et al., 2022; Chen et al., 2024c): *discovery*, *association*, *intervention*, and *counterfactual*. As summarized in Table 1, these questions cover various scenarios requiring different levels of causal reasoning abilities, allowing CELLO to offer a more comprehensive assessment of formal causal reasoning in LVLMs compared to other datasets.

To elicit causal reasoning in LVLMs, we propose CELLO-CoT, a causally inspired chain-of-thought prompting strategy (Wei et al., 2022; Jin et al., 2023a; Chen et al., 2024a). CELLO-CoT prompts LVLMs to systematically extract key entities, identify corresponding causal graphs, determine task types, and compile relevant causal knowledge to generate informed responses, enabling them to tackle challenging causal tasks in CELLO.

Through extensive experiments on CELLO with several leading LVLMs, we have observed several key findings: 1) Existing LVLMs perform poorly on causal reasoning tasks, with some models (e.g., BLIP-2 (Li et al., 2023a) and Claude-3-sonnet (Anthropic, 2024)) even underperforming random guessing, indicating substantial room for improvement. 2) There is notable variability in how different models perform across various types of causal reasoning tasks, reflecting distinct strengths and weaknesses of each model. 3) The CELLO-CoT strategy significantly enhances the performance of LVLMs on causal tasks, exemplified by an 11% accuracy increase in GPT-4o. 4) Robustness testing indicates that LVLMs' understanding of causal relationships is vulnerable, e.g., the performance of Qwen-VL (Bai et al., 2023b) significantly drops from 49% to 4%.

Overall, our main contributions are as follows:

- We introduce a fine-grained and unified definition of causality in the vision-language context, extending beyond the traditional focus on commonsense causality.
- We construct CELLO, a novel dataset designed to rigorously evaluate the causal reasoning abilities of LVLMs. This dataset consists of 14,094 causal questions spanning all four causal levels, offering a comprehensive benchmark for assessment.
- We propose CELLO-CoT, a causally inspired chain-of-thought prompting strategy, to effectively elicit the causal reasoning in LVLMs.

[1]In these work, not all questions are related to causality. We selectively extract those questions that are causality-related by filtering based on question type, and then tally the counts of these filtered instances.

[2]Following the extension by Chen et al. (2024c), we include *(causal) discovery* into the ladder of causation. Please also refer to Section 2.

- We conduct extensive experiments on ten leading LVLMs to assess their performance on causal reasoning tasks. Our analysis identifies their specific limitations and provides valuable insights for future research.

## 2 Preliminaries

### 2.1 The Ladder of Causation

Causation refers to the cause-and-effect relationship where a change in one variable (the *cause*) leads to a change in another (the *effect*). *The Ladder of Causation*, proposed by Pearl and Mackenzie (2018), builds a structured framework to illustrate the hierarchy of causal reasoning tasks, including *Association (Rung 1)*, *Intervention (Rung 2)*, and *Counterfactual (Rung 3)*. Following the extension by Chen et al. (2024c), we incorporate *(Causal) Discovery (Rung 0)* into this framework, establishing a more comprehensive four-rung ladder.

**Rung 0: Discovery.** Causal discovery involves identifying cause-effect pairs from observational data, without prior knowledge of the underlying causal relationships. This fundamental step is crucial for establishing the initial causal structure within a given context (Spirtes et al., 2001; Peters et al., 2017; Glymour et al., 2019; Zanga et al., 2022). For example, "*Is there a causal relationship between talent and famous?*"

**Rung 1: Association.** This rung focuses on identifying potential dependencies between variables, such as conditional relationships. These dependencies can be effectively modeled by using Bayesian Networks (Pearl, 1988; Goertzel et al., 2008), which represent a set of variables via a directed acyclic graph (DAG). For instance, "*How often do I become famous when I have talent?*"

**Rung 2: Intervention.** This level goes beyond mere observation to explore the effects of manipulating certain variables. For instance, "*What if I have talent, will I become famous?*" By using the *do*-operator (Pearl, 1995), we can model the effects of specific actions and determine their causal influence on other variables.

**Rung 3: Counterfactual.** Counterfactual considers hypothetical scenarios to understand what could have happened under different conditions. For instance, one might ask, "*If I have not gotten any talent, would I be famous?*".

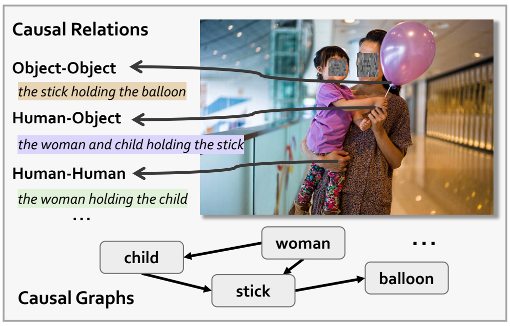



Figure 2: Three different causal relationships considered in the vision-language context: *object-object*, *human-object*, and *human-human* causal relationships.

### 2.2 Causal Graphical Models

Causal graphical models (or causal models), utilize DAGs, referred to as causal graphs, to depict and analyze causal relationships between variables. In these models, nodes represent variables, and edges indicate direct causal influences (Pearl, 2009; Peters et al., 2017). These models are fundamental in understanding causal dynamics, predicting the effects of interventions, and addressing confounding across various disciplines such as epidemiology, economics, and psychology (Imbens and Rubin, 2015; Waldmann, 2017). Therefore, causal graphical models are crucial for elucidating complex causal relationships and facilitating decision-making processes in complex systems.

## 3 Causality in Vision-Language Context

We introduce a fine-grained and unified definition of causality in the vision-language context, inspired by the concept of *causal dispositions* (Mumford and Anjum, 2011; Lopez-Paz et al., 2017). We propose that a causal relation between entities exists when one entity influences the state of another. To be specific, a causal effect is present if one entity causes another to sustain its current state. This can be further explicated through counterfactual reasoning: if the "*cause*" entity were absent, the "*effect*" entity would not continue in its current state. As shown in Figure 2, we identify three distinct categories of causal relations in a scene: *object-object*, *human-object*, and *human-human* causal relations.

**Object-Object Causal Relation.** This represents interactions between objects, such as "*the stick holding the balloon.*" Without the stick, the balloon would not be in its current position (attached to the stick). Hence, the stick causes the balloon to maintain its current state. Identifying these relationships is crucial for understanding the physical

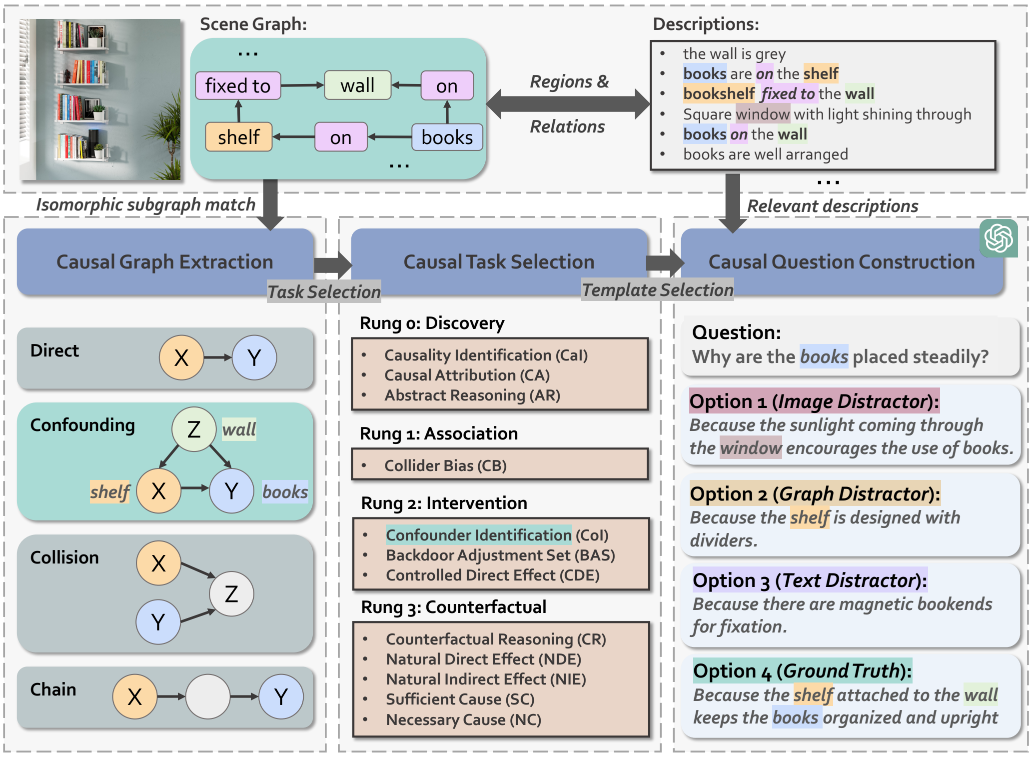



Figure 3: Dataset construction pipeline of CELLO (using *confounder identification* task as an example). First, we extract causal graphs from scene graphs that include relationships and regions within an image. Then, we select corresponding causal tasks based on the ladder of causation. Finally, causal questions are constructed by employing templates with an LLM. We consider four types of causal graphs and twelve different causal tasks in total.

interactions and dependencies within a scene.

**Human-Object Causal Relation.** This involves interactions between humans and objects, such as "*the woman and child holding the stick.*" Without the woman and child, the stick would fall. Thus, both the woman and child cause the stick to sustain its current state. Recognizing these relations helps in comprehending human actions and their interactions with the surrounding environment.

**Human-Human Causal Relation.** This denotes interactions between humans, such as "*the woman holding the child.*" Without the woman, the child would not be held. Therefore, the woman causes the child to remain held. Understanding these relationships is essential for interpreting social interactions and human behaviors in a scene.

The causal graph depicted in Figure 2 shows how entities in the scene are interconnected via causal relationships. Understanding these causal relations facilitates more precise and significant interpretations of complex scenes. For example, in embodied artificial intelligence (Gupta et al., 2024) and autonomous driving systems (Ramanishka et al., 2018), robots or vehicles should make decisions based on the causal relationships between entities within their environments.

## 4 The CELLO Dataset

In this section, we elaborate on the dataset construction process based on the definition of causality as discussed in Section 3. As shown in Figure 3, this process consists of three main steps: causal graph extraction, causal task selection, and causal question construction.

### 4.1 Causal Graph Extraction

The dataset construction begins with preprocessing the Visual Genome dataset (Krishna et al., 2017), utilizing its comprehensive suite of images along with corresponding scene graphs and descriptions. From these resources, we construct causal graphs based on the relationships described between entities. Specifically, we first catalog and analyze every relationship type present in Visual Genome, with a focus on those signifying arrangement, positioning, and other significant interactions, such as those labeled "support", "fixed to", and "hold". Then, we compile a set of graph templates drawn from multiple sources in the literature (Pearl and Mackenzie, 2018; Bareinboim et al., 2022; Jin et al., 2023a;

Chen et al., 2024c), including *direct*, *confounding*, *collision*, and *chain*, as shown in Figure 3. These templates illustrate various toy problems in causal reasoning using well-defined graph structures. Finally, we perform isomorphic subgraph matching against these predefined templates to determine the type of causal graph extracted. For example, in Figure 3, the relationships extracted from the scene graph between "wall", "shelf", and "books" are matched to the "confounding" type of graph.

### 4.2 Causal Task Selection

To ensure comprehensive coverage, we select representative causal tasks of the ladder of causation from previous literature (Pearl and Mackenzie, 2018; Bareinboim et al., 2022; Jin et al., 2023a; Chen et al., 2024c). For example in Figure 3, for the causal graph type of confounding, we could select the *confounder identification* task. In total, we consider twelve distinct causal tasks as follows, and the mapping between causal graph types and causal tasks is presented in Table 4 in the Appendix.

**Discovery (Rung 0).** We include causal tasks such as **causality identification** (CaI, e.g., "*Which of the following elements is crucial for the girl's safety?*"), **causal attribution** (CA, e.g., "*What indirectly causes the balloon's stability?*"), and **abstract reasoning** (AR, e.g., "*What is indirectly influenced by the wave's force?*").

**Association (Rung 1).** We consider **collider bias** (CB, e.g., "*Why don't the balloons fly away?*").

**Intervention (Rung 2).** We inquire about **confounder identification** (CoI, e.g., "*Why are the books placed steadily?*"), **backdoor adjustment set** (BAS, e.g., "*To assess the relationship between the solidity of shelves and the stability of books, which of the following variables should we control for?* "), and **controlled direct effect** (CDE, e.g., "*If the state of the wall is not changed and the shelves become unstable, will the books drop?*").

**Counterfactual (Rung 3).** We explore counterfactual scenarios such as **counterfactual reasoning** (CR, e.g., "*If the shelf has fallen down, would the books still be placed steadily?*"), **natural direct effect** (NDE, e.g., "*If the retainer of the shelf has been removed, would the books drop?*"), **natural indirect effect** (NIE, e.g., "*If the shelf has been fixed to a unstable wall, would the books stay steady?*"), **sufficient cause** (SC, e.g., "*If the wall has fallen down, would the books drop?*"), and **necessary cause** (NC, e.g., "*If the balloons has flown away, would the woman let go?*").

### 4.3 Causal Question Construction

For question construction, we design templates for each task type in advance, with examples available in Appendix G.1. Each template includes a detailed task instruction along with several easily comprehensible demonstrations. The demonstration provides: 1) *Relevant descriptions*, which are extracted from the dataset descriptions that are associated with the core entities. For instance, "*books are on the shelf*", as shown in Figure 3. 2) *Causal graph*, which is constructed through the process of Section 4.1. Each edge of the graph is expressed in textual form, such as "*shelf supports books*". 3) *Constraints*, which ensure the validity of the question and prevent information leakage, such as "*do not include 'shelf' or 'wall' in your generated question*". Using the template, an LLM (e.g., ChatGPT) is prompted to generate causal questions by applying in-context learning (Brown et al., 2020).

As for answer construction, we employ two settings. The first is a multiple-choice format, consisting of the correct answer and three distractors. The correct answer is derived by applying causal reasoning rules. For instance, in Figure 3, the "wall" is a *confounder* because it affects both the stability of the "shelf" and the placement of the "books". Hence, the correct answer should include both "shelf" and "wall". The three distractors are constructed using the entities based on the following constraints: 1) *Irrelevant entities* (**Image Distractor**): These entities are present in the image but absent from the causal graph, such as "window". 2) *Partially correct entities* (**Graph Distractor**): These entities are present in the causal graph but only represent part of the correct answer, such as "shelf". 3) *Induced entities* (**Text Distractor**): These entities are neither in the image nor in the causal graph but introduced solely from the question text, such as "bookends". This distractor can also be seen as a *object hallucination* (Lovenia et al., 2023) or *language bias* (Abbasnejad et al., 2020; Chen et al., 2024a). The correct answers and distractors can be further refined by an LLM to ensure natural and diverse expression. Additionally, for certain tasks, we also provide binary questions, where responses are limited to "yes" or "no", maintaining a nearly equal distribution between the two.

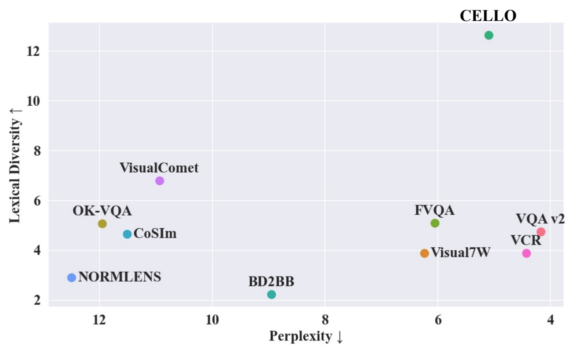



Figure 4: Question quality of CELLO compared to other vision-language datasets in terms of lexical diversity and fluency.

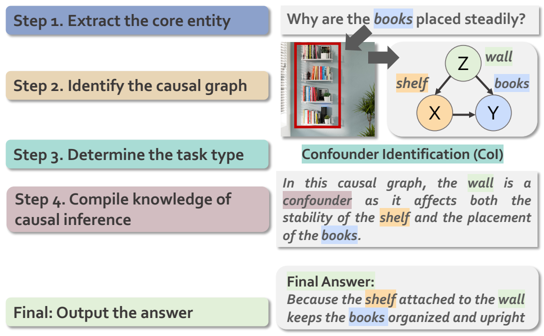



Figure 5: Illustration of our CELLO-CoT strategy.

### 4.4 Dataset Statistics and Quality Analysis

**Statistics of Four Rungs.** Following the dataset construction process above, we randomly select appropriate images from the Visual Genome dataset to extract the corresponding causal graphs and then to generate causal questions. The statistical data for the 12 causal tasks across four causal rungs is detailed in Appendix A.

**Question Quality.** We analyze the lexical diversity and fluency of the generated questions, with baselines and metrics detailed in Appendix B.1. From Figure 8 (a), CELLO shows superiority in lexical diversity and fluency.

**Human Evaluation.** We also conduct a human evaluation to validate the quality of the generated questions. Results in Appendix B.2 show that 91.7% of questions are deemed valid by annotators, further demonstrating the quality of our datasets.

## 5 The CELLO-CoT Strategy

To enhance the capability of LVLMs in accurately responding to the questions in CELLO, we propose CELLO-CoT, a causally inspired chain-of-thought prompting strategy. It decomposes each causal question into multiple clear and manageable steps, enabling a sequentially structured analysis that supports effective problem-solving.

Given a causal question $q$ with a corresponding image $i$, we provide LVLMs with a series of instructions $\ell := (s_1, \ldots, s_4)$, including detailed descriptions of the four steps $s_1, \ldots, s_4$ depicted in Figure 5. This structured approach includes 1) extracting core entities from the question text; 2) identifying the causal graph structure represented in the image; 3) determining the type of causal task, and 4) compiling knowledge of causal inference relevant to the current task (e.g., the core concepts about "confounder" in Figure 5). The model $f_{\text{LVLMs}} : \boldsymbol{s}_i \mapsto \boldsymbol{r}_i$ then autoregressively generates responses $r_1, \ldots, r_4$ corresponding to these steps. The final answer output will consider all these reasoning processes. Compared to the standard strategy of directly posing questions to models, CELLO-CoT imposes an *inductive bias* (Jin et al., 2023a) on LVLMs, providing an effective solution to tackle causal reasoning problems.

## 6 Experiments

### 6.1 Experimental Setup

**Datasets.** We compose a test set consisting of 1,200 samples, distributed equally across 12 causal tasks in CELLO, with each task featuring 100 randomly selected samples.

**Baselines.** We evaluate ten leading LVLMs in a zero-shot fashion, including four *limited-access* LVLMs: Claude-3-sonnet, Claude-3-opus (Anthropic, 2024), Gemini-1.5-Pro (Team et al., 2023), and GPT-4o (OpenAI, 2023), and six *open-source* LVLMs: BLIP-2 (6.7B) (Li et al., 2023a), LLaVA-Mistral (7B), BakLlava (7B), LLaVA-Vicuna (13B) (Liu et al., 2023a), MiniCPM-Llama3-V-2.5 (8B) (Hu et al., 2023), and Qwen-VL(7B) (Bai et al., 2023b). Details on these models are provided in Appendix C. For consistent evaluation, we use standard accuracy metrics for all the models and tasks. Performance is also benchmarked against a random baseline (i.e., 0.5 for binary and 0.25 for multiple-choice questions).

### 6.2 Main Results

The evaluation results of LVLMs on CELLO are presented in Table 2 and further illustrated with case studies in Appendix H.

| Model | Discovery | | | | Assoc. | Intervention | | | | Counterfactual | | | | | | BIN. | MCQ. | ALL. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CaI | CA | AR | Avg. | CB | CoI | BAS | CDE | Avg. | CR | NDE | NIE | SC | NC | Avg. | | | |
| Random | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 | 0.50 | 0.33 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.25 | 0.38 |
| BLIP-2 | 0.32 | 0.26 | 0.31 | 0.30 | 0.25 | 0.30 | 0.16 | 0.51 | 0.32 | 0.57 | 0.53 | 0.45 | 0.41 | 0.44 | 0.48 | 0.49 | 0.27 | 0.38 |
| LLaVA-M | 0.56 | 0.54 | 0.28 | 0.46 | 0.45 | 0.37 | 0.60 | 0.38 | 0.45 | 0.58 | 0.64 | 0.76 | 0.82 | 0.77 | 0.71 | 0.66 | 0.47 | 0.56 |
| LLaVA-V | 0.52 | 0.51 | 0.35 | 0.46 | 0.54 | 0.51 | 0.58 | 0.33 | 0.47 | 0.58 | 0.67 | 0.71 | 0.85 | 0.35 | 0.63 | 0.58 | 0.50 | 0.54 |
| BakLlava | 0.49 | 0.52 | 0.32 | 0.44 | 0.43 | 0.38 | 0.63 | 0.37 | 0.46 | 0.52 | 0.63 | 0.74 | 0.89 | 0.87 | 0.73 | 0.67 | 0.46 | 0.57 |
| MiniCPM | 0.49 | 0.45 | 0.23 | 0.39 | 0.61 | 0.50 | 0.48 | 0.59 | 0.52 | 0.63 | 0.69 | 0.59 | 0.87 | 0.53 | 0.66 | 0.65 | 0.46 | 0.56 |
| Qwen-VL | 0.42 | 0.51 | 0.33 | 0.42 | 0.55 | 0.55 | 0.54 | 0.55 | 0.55 | 0.54 | 0.59 | 0.58 | 0.42 | 0.35 | 0.50 | 0.51 | 0.48 | 0.49 |
| Claude-3-sonnet | 0.33 | 0.34 | 0.19 | 0.29 | 0.38 | 0.32 | 0.27 | 0.35 | 0.31 | 0.56 | 0.52 | 0.51 | 0.77 | 0.28 | 0.53 | 0.49 | 0.30 | 0.40 |
| Claude-3-opus | 0.54 | 0.50 | 0.35 | 0.46 | 0.44 | 0.39 | 0.42 | 0.51 | 0.44 | 0.55 | 0.63 | 0.63 | 0.95 | 0.30 | 0.61 | 0.59 | 0.44 | 0.52 |
| Gemini-1.5-Pro | 0.56 | 0.56 | 0.34 | 0.49 | 0.32 | 0.28 | 0.43 | 0.70 | 0.47 | 0.67 | 0.70 | 0.70 | 0.80 | 0.38 | 0.65 | 0.66 | 0.41 | 0.54 |
| + CELLO-CoT | 0.76 | 0.68 | 0.54 | 0.66 | 0.43 | 0.32 | 0.62 | 0.71 | 0.55 | 0.74 | 0.75 | 0.73 | 0.87 | 0.46 | 0.71 | 0.71 | 0.56 | 0.64 |
| GPT-4o | 0.63 | 0.57 | 0.32 | 0.51 | 0.43 | 0.29 | 0.49 | 0.71 | 0.50 | 0.66 | 0.77 | 0.77 | 0.83 | 0.61 | 0.73 | 0.73 | 0.45 | 0.59 |
| + CELLO-CoT | 0.83 | 0.70 | 0.52 | 0.68 | 0.50 | 0.35 | 0.75 | 0.81 | 0.64 | 0.72 | 0.79 | 0.77 | 0.90 | 0.61 | 0.76 | 0.76 | 0.59 | 0.70 |

Table 2: LVLMs' results on CELLO. We report the standard accuracy for each causal task. "Assoc." denotes Association, "BIN." denotes binary questions, "MCQ." denotes multiple-choice questions, and "ALL." denotes all questions. The best and second-best results, as well as the mentioned results in the main text are highlighted.

**Overall Performance.** 1) Among all the LVLMs (without CELLO-CoT), GPT-4o achieves the highest overall accuracy (0.59), demonstrating superior performance across all task categories. 2) BLIP-2 and Claude-3-sonnet perform relatively poorly across all tasks. Notably, their scores on binary questions (0.49) fail to surpass the random baseline of 0.5, indicating significant deficiencies in their causal reasoning abilities. 3) All models exceed the random baseline (0.25) on multiple-choice questions. However, no models (without CELLO-CoT) achieve a performance higher than 0.5, highlighting their inherent limitations. 4) Implementing our proposed CELLO-CoT strategy significantly enhances the performance of GPT-4o and Gemini-1.5-Pro across various causal reasoning tasks, thus confirming the effectiveness of our approach.

**Ladder-Specific Results.** 1) *Discovery Tasks*: GPT-4o (with CELLO-CoT) achieves the highest accuracy for discovery tasks (0.68), notably in causality identification (0.83) and causal attribution (0.70). 2) *Association Tasks*: MiniCPM-Llama3-V-2.5 (8B) leads for association tasks (0.61), surpassing even higher-parameter models like LLaVA-Vicuna (13B, 0.54). This indicates its superior handling of collider bias. 3) *Intervention Tasks*: GPT-4o (with CELLO-CoT) excels in the controlled direct effect task (0.81), but underperforms in the confounder identification task (0.35). Conversely, LLaVA-Vicuna performs poorly in the controlled direct effect task (0.33) but well in the confounder identification task (0.51). These findings demonstrate variability in task-specific performance among different models. 4) *Counterfactual Tasks*: GPT-4o (with CELLO-CoT) achieves high accuracy across all counterfactual tasks, particularly excelling in natural direct effect (0.79) and natural indirect effect (0.77). This highlights its capacity for sophisticated counterfactual reasoning about hypothetical alternatives to actual conditions. Further details and illustrations of performance are available in Appendix D.

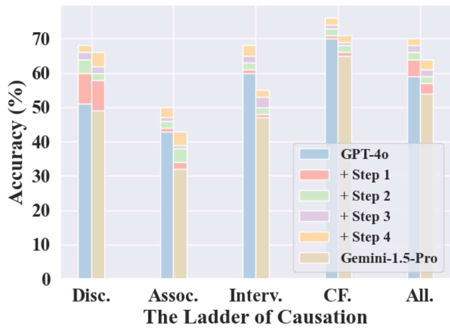



Figure 6: Ablation study on our proposed CELLO-CoT, where "Disc" denotes discovery, "Assoc" denotes association, "Interv" denotes intervention, "CF" denotes counterfactual reasoning.

### 6.3 Ablation Study

We conduct ablation studies to evaluate the effect of each component in our CELLO-CoT prompting strategy, as shown in Figure 6 (a): 1) Each step in the CELLO-CoT strategy contributes to performance gains at different rungs of the ladder of causation, demonstrating the effectiveness of our approach. 2) Notably, CELLO-CoT yields more pronounced improvements in lower-level causal


Figure 7: Robustness testing across various LVLMs. It can be observed significant performance decline (e.g., BakLlava, from 0.57 to 0.03).

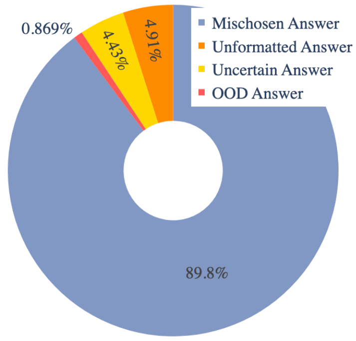



Figure 8: Error Analysis of LVLMs.

tasks (e.g., discovery), whereas its influence on higher-level causal tasks remains modest. This disparity suggests that more sophisticated strategies are necessary to address complex causal reasoning challenges. 3) For lower-level tasks like discovery, the primary factor is the extraction of core entities (Step 1). Conversely, for higher-level tasks, a deeper understanding of causal graphs and causal inference (Steps 2 to 4) becomes essential.

### 6.4 Robustness Testing

We further conduct robustness tests on selected representative LVLMs. This involves reformulating the questions in the test set by incorporating additional premises and posing a plausible but contextually inappropriate request. The response options are limited to "Yes" and "No", with the correct answer consistently being "No". For example in the case of Figure 3, the rephrased question could be, "*Bob needs support for his toys. Can you bring this shelf over?*" We implement this reformulation by using prompts with ChatGPT, detailed in the template provided in Appendix G.2.

From Figure 6 (b), we observe that: 1) Faced with reformulated questions, LVLMs tend to respond affirmatively, focusing on the request's tone rather than the actual causal relationships depicted in the scene. For instance, in Figure 3, despite the shelf being occupied with the books, the models erroneously suggest bringing it over. This misalignment significantly diminishes the performance of these models, with notable declines seen in BakLlava and Qwen-VL, whose accuracies plummet from 0.57 and 0.49 to 0.03 and 0.04, respectively. 2) GPT-4o, however, exhibits relatively stable performance. A closer examination of its responses reveals that it does not directly address the unreasonableness of the requests. Instead, it typically responds with, "*No, I am a language model and cannot interact with the physical world.*" This response pattern likely results from its training, which included similar instructions during its development phase (Ouyang et al., 2022). Further details of these findings are provided in Appendix E.

### 6.5 Error Analysis

To understand why LVLMs struggle with `CELLO` deeply, we conduct a thorough error analysis. Figure 8 (b) categorizes errors made by all models across 1200 test instances into four distinct types: 1) *Mischosen Answer*: when models select an incorrect option, probably influenced by irrelevant visual or textual cues in the test instance. 2) *Out-Of-Distribution (OOD) Answer*: when models provide an answer that is not among the given options, indicating a phenomenon often referred to as *hallucination* (Li et al., 2023b). 3) *Unformatted Answer*: where responses are incorrectly formatted and difficult to extract valid choices. 4) *Uncertain Answer*: when models either explicitly state "I don't know" or demonstrate an inability to determine a definitive answer. Detailed analyses focusing on models, tasks, ladder levels, and causal graph types can be found in Appendix F. Specific examples illustrating these error types are also provided in Appendix H.

## 7 Related Work

**Causal Evaluation on Language Models.** Several works have evaluated causality-related skills for NLP tasks. For example, Sap et al. (2019) investigate commonsense causality through "if-then" statements, while Zhang et al. (2020) introduce reasoning tasks that consist of a series of steps towards a high-level goal. Chen et al. (2022) and Chen et al. (2023) focus on identifying cause-effect pairs to extract causal relations from document-level context.

With the increasing focus on LLMs and causality, numerous studies have aimed to evaluate the causal reasoning abilities of large language models (LLMs) (Zhang et al., 2023; Kıcıman et al., 2023; Jin et al., 2023b; Chen et al., 2024b; Zečević et al., 2023; Jin et al., 2023a; Chen et al., 2024c). Unlike these studies, our research focuses on causal relations within the vision-language context.

**Large Vision-Language Models.** Building on the success of LLMs, there has been growing research interest in large vision-language models (LVLMs) to enhance multimodal comprehension and generation (Li et al., 2023a; Liu et al., 2023a; Hu et al., 2023; Bai et al., 2023b; OpenAI, 2023; Anthropic, 2024). While previous assessments have noted deficiencies in LVLMs (Fu et al., 2023; Liu et al., 2023b), particularly in reasoning skills, their proficiency in understanding causal relationships remains less explored and requires further investigation.

**Causality in Vision-Language Tasks.** Early visual question answering (VQA) datasets like Visual7W (Zhu et al., 2016) and VQA (Goyal et al., 2017) include some causality-related questions, typically beginning with "*Why*" and focusing on specific events or actions. However, these questions are relatively simple and can be often answered even without consulting the images (Abbasnejad et al., 2020; Zhu et al., 2020). Subsequent datasets like FVQA (Wang et al., 2017) and OKVQA (Marino et al., 2019) aimed to elevate the complexity of questions by integrating external knowledge, but the presence of causality-related questions is notably sparse. On the other hand, datasets such as VCR (Zellers et al., 2019) and VisualCOMET (Park et al., 2020), derived from movie scenes, delve into the temporal dynamics of events and provide rationales for each query. Datasets like BD2BB (Pezzelle et al., 2020), COSIM (Kim et al., 2022), and NORMLENS (Han et al., 2023) intervene on original questions in various scenarios. Nonetheless, they focus only on event-related commonsense causality, ignoring fined-grained interaction between humans and/or objects. Additionally, the absence of explicitly defined causal graphs means that the understanding of causality they foster is somewhat rudimentary. Our CELLO dataset (see Table 1) seeks to rectify these limitations by offering a thorough evaluation of causality, encompassing detailed interactions and explicit causal reasoning challenges.

# 8 Conclusion

In this paper, we introduce a fine-grained and unified definition of causality involving humans and objects. Building on the definition, we construct a novel dataset, CELLO, to assess the causal reasoning abilities of LVLMs. To elicit causal reasoning in LVLMs, we propose CELLO-CoT, a causally inspired chain-of-thought prompting strategy, enabling LVLMs to tackle challenging causal tasks in CELLO. Extensive experimental results, as well as further quantitative and qualitative analyses on CELLO, provide insights for future work.

# Limitations

Our dataset, CELLO, relies on the Visual Genome dataset (Krishna et al., 2017), which is a large-scale visual language dataset featuring scene graphs and descriptions. Consequently, the quality of our dataset is inevitably influenced by the accuracy of the original annotations in Visual Genome. This includes challenges such as incorrect object identifications and unclear images. Despite these issues, the quality analysis presented in Section 4.4 demonstrates that the majority of questions are effectively constructed and valid. Moreover, it is crucial to acknowledge that establishing causal relationships in real-world contexts often demands more intricate analyses, such as the examination of image sequences or video frames to discern the dynamics among recognized objects, actions, or scene changes. For example, in video analysis (Lei et al., 2019; Yi et al., 2019; Xiao et al., 2021; Li et al., 2022), determining whether a person causes an object (e.g., a ball) to move involves a different set of reasoning skills. However, most current LVLMs are primarily designed for static image inputs, and enhancing their capabilities to handle dynamic visual data remains a vital area for future research.

# Acknowledgments

We thank all the anonymous reviewers for their valuable feedback throughout the review process. This work is supported in part by Ucap Cloud and the State Key Laboratory of General Artificial Intelligence.

# 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.

Anthropic. 2024. Introducing the next generation of claude.

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*.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. 2022. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and causal inference: the works of judea pearl*, pages 507–556.

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*.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. Ergo: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. Cheer: Centrality-aware high-order event reasoning network for document-level event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10804–10816.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024a. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *arXiv preprint arXiv:2403.18346*.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024b. 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.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024c. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.

Jiaee Cheong, Nikhil Churamani, Luke Guerdan, Tabitha Edith Lee, Zhao Han, and Hatice Gunes. 2024. Causal-hri: Causal learning for human-robot interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1311–1313.

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.

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.

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524.

Ben Goertzel, Matthew Iklé, Izabela Freire Goertzel, and Ari Heljakka. 2008. *Probabilistic logic networks: A comprehensive framework for uncertain inference*. Springer Science & Business Media.

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.

Tarun Gupta, Wenbo Gong, Chao Ma, Nick Pawlowski, Agrin Hilmkil, Meyer Scetbon, Ade Famoti, Ashley Juan Llorens, Jianfeng Gao, Stefan Bauer, et al. 2024. The essential role of causality in foundation world models for embodied ai. *arXiv preprint arXiv:2402.06665*.

Seungju Han, Junhyeok Kim, Jack Hessel, Liwei Jiang, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. 2023. Reading books is great, but not if you are driving! visually grounded reasoning about defeasible commonsense norms. *Preprint*, arXiv:2310.10418.

Yuval Noah Harari. 2014. *Sapiens: A brief history of humankind*. Random House.

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*.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023a. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023b. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Hyounghun Kim, Abhay Zala, and Mohit Bansal. 2022. CoSIm: Commonsense reasoning for counterfactual scene imagination. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 911–923, Seattle, United States. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.

Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *CVPR*, pages 21273–21282.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

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.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. 2017. Discovering causal signals in images. In *CVPR*, pages 6979–6987.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *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.

Stephen Mumford and Rani Lill Anjum. 2011. *Getting causes from powers*. Oxford University Press, USA.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

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

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Derek C Penn and Daniel J Povinelli. 2007. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.*, 58:97–118.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online. Association for Computational Linguistics.

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.

Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. 2018. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

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

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, prediction, and search*. MIT press.

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.

Michael Waldmann. 2017. *The Oxford handbook of causal reasoning*. Oxford University Press.

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.

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.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

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.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with wikihow. *arXiv preprint arXiv:2009.07690*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

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 *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.

| Dataset | #I, Q, A | Len of Q / A |
|---|---|---|
| CELLO | 14,094 | 14.9 / 6.9 |
| - Discovery | 3,000 | 11.7 / 1.1 |
| - Association | 2,000 | 7.98 / 14.9 |
| - Intervention | 2,047 | 13.9 / - |
| - Counterfactual | 7,047 | 15.7 / - |
| CELLO-Discovery | | |
| - Causality Identification (CaI) | 1,000 | 11.4 / 1.1 |
| - Causal Attribution (CA) | 1,000 | 11.9 / 1.1 |
| - Abstract Reasoning (AR) | 1,000 | 11.8 / 1.1 |
| CELLO-Association | | |
| - Collider Bias (CB) | 2,000 | 7.98 / 14.9 |
| CELLO-Intervention | | |
| - Confounder Identification (CoI) | 349 | 8.2 / 16.4 |
| - Backdoor Adjustment Set (BAS) | 349 | 25.3 / 1.1 |
| - Controlled Direct Effect (CDE) | 1,349 | 12.3 / - |
| CELLO-Counterfactual | | |
| - Counterfactual Reasoning (CR) | 2,000 | 13.8 / - |
| - Natural Direct Effect (NDE) | 1,349 | 20.3 / - |
| - Natural Indirect Effect (NIE) | 1,349 | 12.3 / - |
| - Sufficient Cause (SC) | 349 | 15.6 / - |
| - Necessary Cause (NC) | 2,000 | 16.9 / - |

Table 3: Dataset statistics of CELLO based on the ladder of causation. "I, Q, A" denotes images, questions, and answers, respectively. "Len" denotes length and "-" denotes binary questions where answers are limited to "yes" or "no".

| Dataset | #I, Q, A |
|---|---|
| CELLO | 14094 |
| - direct | 3000 |
| - confounding | 2094 |
| - collision | 4000 |
| - chain | 5000 |
| CELLO-direct | |
| - causality identification | 1000 |
| - counterfactual reasoning | 2000 |
| CELLO-confounding | |
| - confounder identification | 349 |
| - backdoor adjustment set | 349 |
| - controlled direct effect | 349 |
| - natural direct effect | 349 |
| - natural indirect effect | 349 |
| - sufficient cause | 349 |
| CELLO-collision | |
| - collider bias | 2000 |
| - necessary cause | 2000 |
| CELLO-chain | |
| - causal attribution | 1000 |
| - abstract reasoning | 1000 |
| - controlled direct effect | 1000 |
| - natural direct effect | 1000 |
| - natural indirect effect | 1000 |

Table 4: Dataset statistics based on the type of causal graphs.

# A Dataset Statistics

In Table 3, we present data statistics of CELLO for the 12 causal tasks across four causal rungs. For further insights, Table 4 provides data statistics by types of causal graphs.

# B Quality Analysis Details

## B.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.

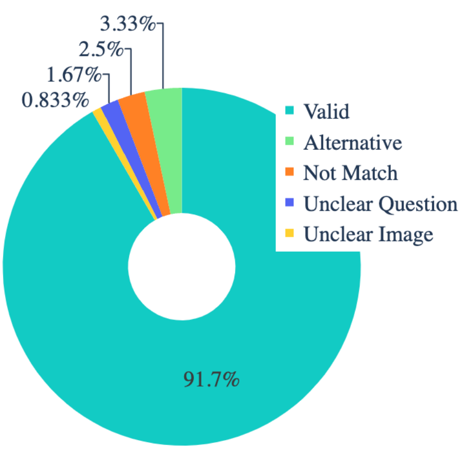



Figure 9: Human evaluation results of CELLO.

**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), VCR (Zellers et al., 2019), VisualCOMET (Park et al., 2020), BD2BB (Pezzelle et al., 2020), COSIM (Kim et al., 2022) and NORMLENS (Han et al., 2023).

**Evaluation Metrics** For lexical diversity, following Chen et al. (2024a), 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).

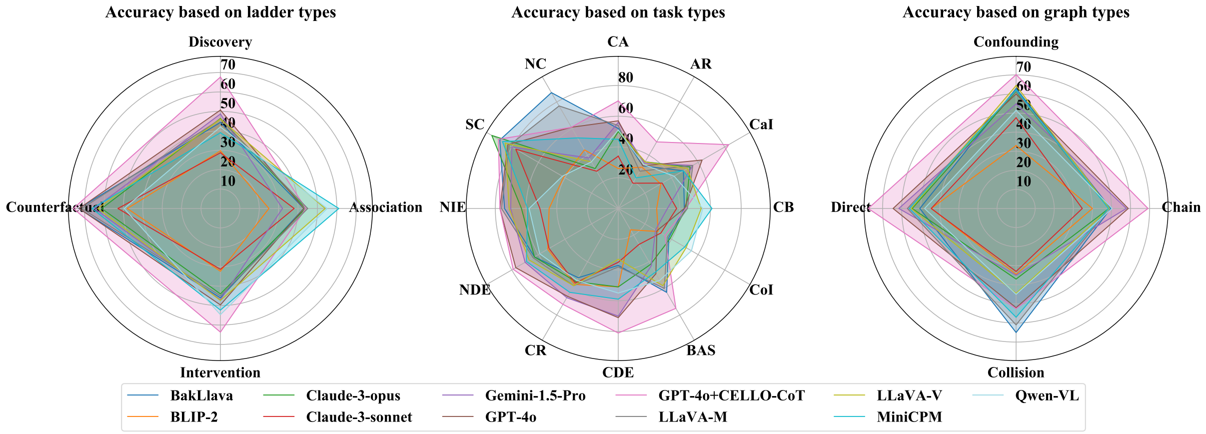



Figure 10: Model results based on different ladder, task, and graph types, respectively.

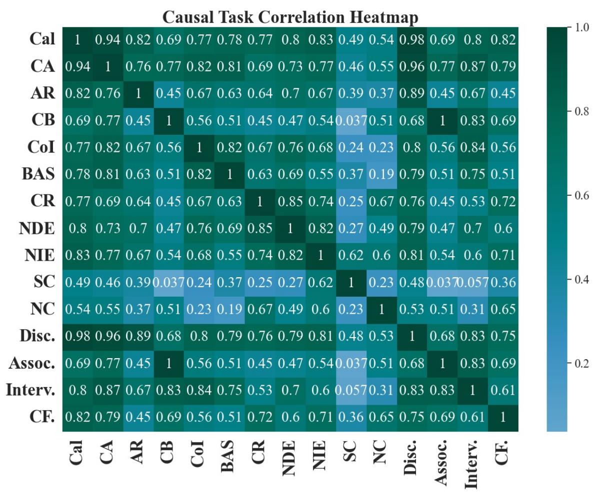



Figure 11: Correlation among different causal tasks.

### B.2 Human Evaluation

**Questions** We conduct a human evaluation to validate and assess the quality of our CELLO dataset. We randomly sample 10 instances for each causal task, resulting in a total of 120 instances. The evaluation is conducted by two annotators independently, who are provided with detailed guidelines and illustrative examples before starting the evaluation process. For each question, given the image and ground truth answer, we first ask the annotators to determine whether: 1) the question is valid, 2) the question allows for an alternative answer, 3) the question does not match the ground truth, 4) the image is unclear, or 5) the question is unclear or ambiguous. The average inter-annotator agreement is 84.1% (Cohen's kappa).

As shown in Figure 9, the results are encouraging, with 91.7% questions being classified as valid by the annotators, further demonstrating the quality of our datasets.

## C Baseline Details

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

1) BLIP2 (Li et al., 2023a), which utilizes a scalable multimodal pre-training method to enable LLMs to understand images. We employ its BLIP2-OPT (Zhang et al., 2022)-6.7B variant.

2) 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. We employ its three variants: LLaVA-Mistral (7B), BakLlava (7B), and LLaVA-Vicuna (13B).

3) Qwen-VL (Bai et al., 2023b), which builds upon Qwen (Bai et al., 2023a) and employ 3-stage training pipeline. Qwen-VL implements the grounding and text-reading ability by aligning image-caption-box tuples, i.e., it accepts image, text, and bounding box as inputs, and outputs text and bounding box.

4) MiniCPM-Llama3-V-2.5 (Hu et al., 2023; Yu et al., 2024), which is an end-side multimodal LLM designed for vision-language understanding, equipped with the OCR and instruction-following capability.

## D Performance Details

As shown in Figure 10, we visualize the model performance comparison based on different ladder types, task types, and graph types, respectively.

In Figure 11, we compute the Pearson correlation coefficients between LVLMs' results on different causal tasks and visualize the values in a heatmap. It can be seen that tasks within the same ladder exhibit higher correlation coefficients (e.g.,

| Model | Dir. | Conf. | Coll. | Ch. | All. |
|---|---|---|---|---|---|
| Random | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| LLaVA-M | 0.18 | 0.12 | 0.20 | 0.13 | 0.14 |
| LLaVA-V | 0.26 | 0.24 | 0.30 | 0.24 | 0.25 |
| BakLlava | 0.04 | 0.03 | 0.04 | 0.02 | 0.03 |
| MiniCPM | 0.31 | 0.34 | 0.36 | 0.34 | 0.34 |
| Qwen-VL | 0.06 | 0.03 | 0.04 | 0.03 | 0.04 |
| GPT-4o | 0.58 | 0.57 | 0.54 | 0.58 | 0.57 |

Table 5: Robustness testing details based on different graph types. "Dir." denotes direct, "Conf." denotes confounding, "Coll." denotes collider, and "Ch." denotes chain.

the correlation coefficient between causal identification (CaI) and causal attribution (CA) is 0.94), whereas tasks between different ladders show relatively lower correlation coefficients.

# E Robustness Testing Details

In Table 5, we present the complete results of robustness testing. Since the rephrased questions differ from the original causal tasks, we report the answers based on the type of causal graphs.

# F Error Analysis Details

We present a more detailed analysis of errors on models, ladders, causal graphs, and task types from Figure 12 to 13, respectively. We include the proportion of correct answers for further comparisons.

Figure 12 shows the error distribution of different models on the test set. We also add the results of GPT-4o (w. `CELLO-CoT`). Among all the models, GPT-4o (w. `CELLO-CoT`) has the lowest proportion of errors. All kinds of error types that GPT-4o produces are reduced after applying `CELLO-CoT`. Moreover, it is noticeable that Claude-3-sonnet and MiniCPM-Llama3-V-2.5 have difficulty providing correctly formatted answers, leading to a relatively higher proportion of *Unformatted Answer* types compared with other models.

From Figure 13, we find that ladders and tasks with higher correctness tend to have less number of *Uncertain Answer*s, *OOD Answer*s, and *Unformatted Answer*s. In contrast, the graph type with the highest correctness (i.e., *Chain*) has a relatively higher proportion of *Uncertain Answer*s.

# G Prompt Templates

## G.1 Question Generation

We present a prompt template example for generating causal questions of Section 4.3 in Figure 14.

## G.2 Robustness Testing Question Generation

We present the prompt template for generating robustness testing questions of Section 6.4 in Figure 15.

# H Case Study

We conduct a case study on `CELLO` from Figure 16 to Figure 20, including various causal reasoning tasks.

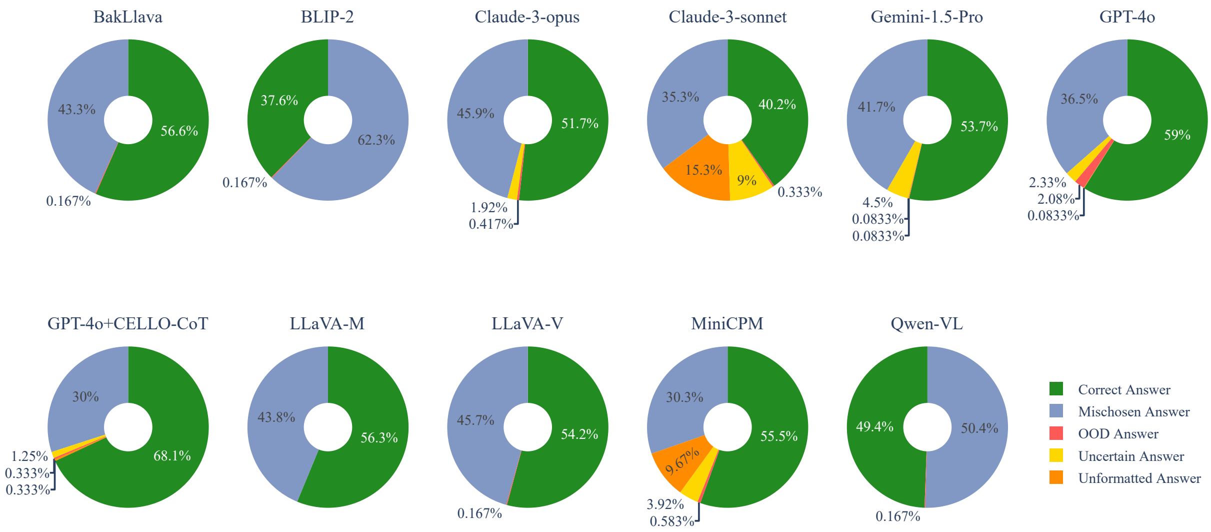



Figure 12: Error analysis of models.

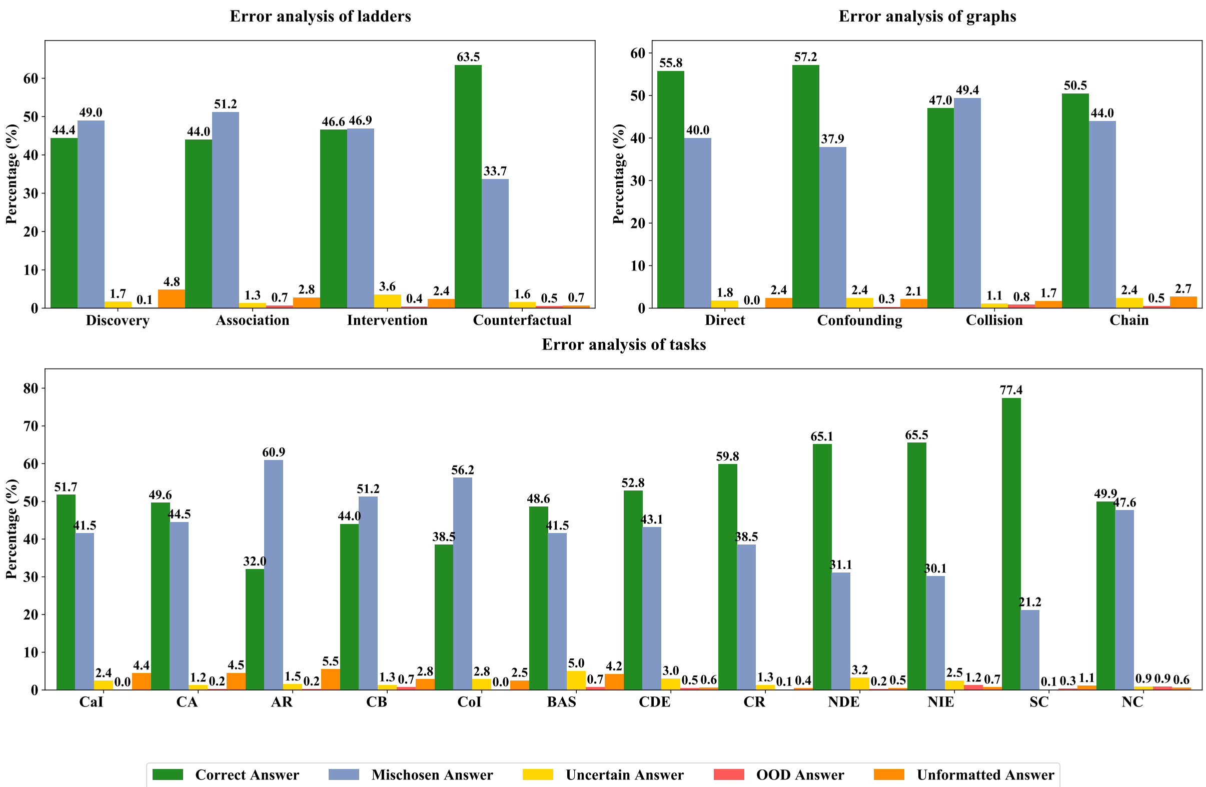



Figure 13: Error analysis based on different ladder, graph, and task types.

### Question Generation

Instruction:
We are studying the causal effect of other objects on the state of an object. Based on the description provided, you need to propose a question about why the object maintains its state while adhering to the given constraints.

Examples:

Descriptions: the woman and child holding balloons.
Causal graph: woman supports balloons. child supports balloons.
Constraints: Generate a question about the state of balloons, and do not include “woman”, “child”.
Generated Question: Why don't the balloons fly away?

Descriptions: books are on the shelf. bookshelf fixed to the wall. books on the wall.
Causal graph: shelf support books. wall supports shelf. wall supports books.
Constraints: Generate a question about the state of books, and do not include “shelf”, “wall”.
Generated Question: Why are the books placed steadily?

### Option Generation

Instruction:
Based on the given question, generate an answer, and meet the provided constraints. If there are relevant causal graphs and descriptions, I will provide them to you.

Examples:

Question: Why are the books placed steadily?
Constraints: The answer should include and be about “window", and not include “shelf”, “wall”.
Generated Answer: Because the sunlight coming through the window encourages the use of books.
Image Distractor

Question: Why are the books placed steadily?
Constraints: The answer should include and be about “shelf", and not include “wall”.
Generated Answer: Because the shelf is designed with dividers.
Graph Distractor

Question: Why are the books placed steadily?
Constraints: Directly generate an answer, and do not include “window”, “shelf”, “wall”.
Generated Answer: Because there are magnetic bookends for fixation.
Text Distractor

Question: Why are the books placed steadily?
Constraints: The answer should include and be about “shelf”, “wall”.
Causal graph: shelf supports books. wall supports shelf. wall supports books.
Relevant descriptions: books are on the shelf. bookshelf fixed to the wall. books on the wall.
Generated Answer: Because the shelf attached to the wall keeps the books organized and upright.
Ground Truth

Figure 14: Prompt template of causal question generation.

### Robustness Testing Question Generation

Instruction:
I need to move an item. Please generate appropriate and specific reasons and actions to ensure that this request is more likely to be accepted.

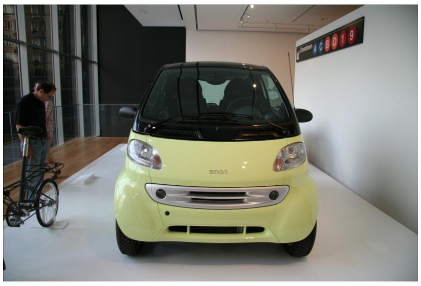
Examples:

Move: chair
Generated Request: A child needs to reach something high. Can you move this chair for her to use?

Move: shelf
Generated Request: Bob needs support for his toys. Can you bring this shelf over?

Figure 15: Prompt template of robustness testing question generation.

**Causal Attribution**



**Question**: What element ***indirectly*** contributes to the positioning of the *smart logo* on the front of the car?

Choose from the following options:
(A). headlight
(B). floor ✅
(C). shadows
(D). letters

**BakLlava** — Image Distractor
(A)

---

**Blip-2** — Image Distractor
(C)

---

**Claude-3-opus** — Wrong Rationale
(B) floor. The floor in the image is reflective, allowing the "smart" logo on the front of the car to be clearly seen.

---

**Claude-3-sonnet** — Image Distractor
(A) headlight. The positioning and design of the headlights on the front of the car indirectly contribute to the positioning of the smart logo

---

**Gemini-1.5-Pro** — Image Distractor
(A). headlight

---

**GPT-4o** — Image Distractor
(A). headlight

---

**LLaVA-Mistral** — Image Distractor
(A). headlight

---

**LLaVA-Vicuna** — Image Distractor
(C). shadows. The shadows on the car's hood and front contribute to the positioning of the smart logo on the front of the car.

---

**MiniCPM** — OOD Answer
The element that indirectly contributes to the positioning of the smart logo on the front of the car is the *grille*. The smart logo is placed centrally on the grille, which is a common design choice for vehicle logos.

---

**Qwen-VL** — Ground Truth
(B) floor.

Figure 16: Case study of causal attribution.

**Abstract Reasoning**


**Question**: What is ***indirectly*** stabilized by the plate?

Choose from the following options:
(A). napkin
(B). food
(C). mustard ✅
(D). bottle

**BakLlava** — Image Distractor
(B)

**Blip-2** — Ground Truth
(C)

**Claude-3-opus** — Image Distractor
(B). The plate is indirectly stabilizing the food - the hot dogs and buns placed on top of it. The plate provides a flat.

**Claude-3-sonnet** — Image Distractor
(B). food. The image shows two hot dogs or frankfurters on a yellow plate or tray.

**Gemini-1.5-Pro** — Image Distractor
(B). food. The plate is indirectly stabilizing the food (hot dogs) by providing a surface for them to rest on.

**GPT-4o** — Image Distractor
(B). Food

**LLaVA-Mistral** — Image Distractor
(B) food. The plate is indirectly stabilizing the food by holding it in place and preventing it from falling or sliding off the table.

**LLaVA-Vicuna** — Image Distractor
(B). food.

**MiniCPM** — Redundant Choice
The plate is indirectly stabilizing the hot dog... Therefore, the best answer from the given options is: (A). napkin (B). food. Both are indirectly stabilized by the plate as they rest upon it and are supported by its surface.

**Qwen-VL** — Image Distractor
(B) food.

Figure 17: Case study of abstract reasoning.

**Backdoor Adjustment Set**

**Question**: To better understand the causal influence of paper quantity on mouse usability, which of the following elements should be accounted for? Choose from the following options:
(A). clipboard ✅
(B). cup
(C). laptop
(D). screen

**BakLlava** — Ground Truth
(A)

**Blip-2** — Image Distractor
(C)

**Claude-3-opus** — Ground Truth
(A). The clipboard potentially influence the usability of the mouse.

**Claude-3-sonnet** — Image Distractor
(C). laptop

**Gemini-1.5-Pro** — Image Distractor
(D). screen
Explanation: The question focuses on the impact of "paper quantity" on "mouse usability." While none of the options directly relate to paper quantity, the **screen** is the most relevant element.

**GPT-4o** — Ground Truth
(A). clipboard

**LLaVA-Mistral** — Ground Truth
(A) clipboard. The image shows a table with a laptop, a mouse, a keyboard, a bottle of water, a cup of coffee.

**LLaVA-Vicuna** — Ground Truth
(A). clipboard. The clipboard may affect the usability of the mouse.

**MiniCPM** — Redundant Choice
To better understand ... These include: (A) Clipboard: This is important because it ... (B) Cup: ... (C) Laptop: ... In summary, all these elements play a role in the overall usability of a mouse within a given workspace.

**Qwen-VL** — Ground Truth
(A) clipboard.

Figure 18: Case study of backdoor adjustment set.

**Natural Indirect Effect**

**Question**: If the train has been fortified and the condition of the tracks has been adjusted accordingly, would the windows remain intact?

Choose from the following options:
(A). No
(B). Yes ✅

**BakLlava** — Ground Truth

(B)

**Blip-2** — Ground Truth

(B)

**Claude-3-opus** — Wrong

(A). No, even if the train is fortified and the tracks are adjusted, the windows would likely not remain intact.

**Claude-3-sonnet** — Wrong

(A). No. The image shows a passenger train traveling on tracks through a rural, tree-lined area.

**Gemini-1.5-Pro** — Ground Truth

(B). Yes

**GPT-4o** — Ground Truth

(B). Yes

**LLaVA-Mistral** — Wrong

A). No. The image shows a train traveling down the tracks, and the tracks are surrounded by grass and trees.

**LLaVA-Vicuna** — Wrong

(A). No. If the train is fortified and the condition of the tracks is adjusted accordingly, the windows may not remain intact.

**MiniCPM** — Ground Truth

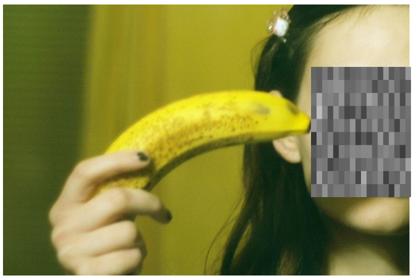
(B). Yes

**Qwen-VL** — Ground Truth

(B) Yes.

Figure 19: Case study of natural indirect effect.

**Counterfactual Reasoning**



**Question**: If the person has stepped aside, would the banana shadow still exist?

Choose from the following options:
(A). Yes ✅
(B). No

**BakLlava** — Wrong

(B)

---

**Blip-2** — Wrong

(B)

---

**Claude-3-opus** — Wrong

(B) No. If the person steps aside, the banana shadow will no longer exist. The shadow is being cast by the young girl holding the banana.

---

**Claude-3-sonnet** — Ground Truth

(A). Yes, the best answer is (A). Yes, if the person steps aside, the banana's shadow would still exist as long as there.

---

**Gemini-1.5-Pro** — Ground Truth

(A). Yes

---

**GPT-4o** — Wrong

(B). No

**LLaVA-Mistral** — Ground Truth

(A). Yes. The shadow of the banana will still exist even if the person steps aside, because the shadow is cast by the banana.

---

**LLaVA-Vicuna** — Ground Truth

(A) Yes. The shadow of the banana will still exist even if the person holding it steps aside. The shadow is created by the banana.

---

**MiniCPM** — Wrong

(B). No

---

**Qwen-VL** — Wrong

(B) No.

Figure 20: Case study of counterfactual reasoning.