Insight Over Sight: Exploring the Vision-Knowledge Conflicts in Multimodal LLMs

Xiaoyuan Liu[1,2*] **Wenxuan Wang**[3] **Youliang Yuan**[1,2] **Jen-tse Huang**[4]
Qiuzhi Liu[2] **Pinjia He**[1†] **Zhaopeng Tu**[2]
[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen
[2]Tencent [3]Renmin University of China [4]Johns Hopkins University
[1]xiaoyuanliu@link.cuhk.edu.cn [1]hepinjia@cuhk.edu.cn [2]zptu@tencent.com

Abstract

This paper explores the problem of commonsense level vision-knowledge conflict in Multimodal Large Language Models (MLLMs), where visual information contradicts model's internal commonsense knowledge. To study this issue, we introduce an automated framework, augmented with human-in-the-loop quality control, to generate inputs designed to simulate and evaluate these conflicts in MLLMs. Using this framework, we have crafted a diagnostic benchmark consisting of 374 original images and 1,122 high-quality question-answer (QA) pairs. The benchmark covers two aspects of conflict and three question types, providing a thorough assessment tool. We apply this benchmark to assess the conflict-resolution capabilities of nine representative MLLMs from various model families. Our results indicate an evident over-reliance on parametric knowledge for approximately 20% of all queries, especially among Yes-No and action-related problems. Based on these findings, we evaluate the effectiveness of existing approaches to mitigating the conflicts and compare them to our "Focus-on-Vision" prompting strategy. Despite some improvement, the vision-knowledge conflict remains unresolved and can be further scaled through our data construction framework. Our proposed framework, benchmark, and analysis contribute to the understanding and mitigation of vision-knowledge conflicts in MLLMs.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2022; Touvron et al., 2023; OpenAI, 2023a; Meta AI, 2024) have reshaped the landscape of deep learning for their comprehensive capabilities in language understanding, reasoning, and generation (Wei et al., 2022a,b; Chen et al., 2023). This evolution has paved the way for the emergence of Multimodal Large Language Models (MLLMs) (Li et al., 2022, 2023b; Lyu et al., 2023; Dai et al., 2023; Liu et al., 2024d; Zhu et al., 2023; Bai et al., 2023; Liu et al., 2024b; OpenAI, 2023b; Liu et al., 2024c; Tong et al., 2024; Gemini Team, 2024; OpenAI, 2024b; Meta AI, 2025), which integrate a vision model with the LLM to process visual information. Modern MLLMs like `GPT-4o` (OpenAI, 2024b) and `LLaVA-NeXT` (Liu et al., 2024c) have exhibited remarkable proficiency across various vision-language tasks such as image captioning (Chen et al., 2015), visual question answering (VQA) (Antol et al., 2015), and visual reasoning (Johnson et al., 2017; Yue et al., 2024).

Figure 1: Illustration of the vision-knowledge conflict, where the visual input contradicts MLLM's inherent knowledge. The MLLM's response over-relies on its inherent commonsense knowledge.

However, the persistence of knowledge conflicts in LLMs remains a significant challenge for MLLMs. Introducing visual information into MLLMs generates a novel form of discrepancy, a phenomenon we term "**vision-knowledge conflict**", in which the visual data contradicts the model's pre-existing parametric knowledge. Previous studies have demonstrated that LLMs can exhibit complex behavior when confronted with knowledge conflicts, oscillating between rigid adherence to their parametric knowledge and excessive sensitivity to contextual cues (Xie et al., 2024).

[*] Work was done when Xiaoyuan, Wenxuan, Youliang, and Jen-tse were interning at Tencent.
[†] Pinjia He is the corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17825–17846

This behavior brings substantial risks, especially when external information is critical for decision-making. Given the growing need for trustworthiness, real-time accuracy, and robustness in MLLM systems, it is imperative to further investigate and resolve these vision-knowledge conflicts, which are believed to be a primary cause of hallucinations (Liu et al., 2023a; Guan et al., 2024). While recent research has simulated these conflicts by manually crafting counterfactual images (Guan et al., 2024), there remains significant room for improvement in conflict categorization, question diversity, and image naturalness. Additionally, manually crafted benchmarks are limited in sample size and scalability, underscoring the need for a more automated pipeline to enable broader analysis.

To this end, we propose an automated framework with human-in-the-loop to develop a benchmark for simulating and analyzing vision-knowledge conflicts in MLLMs at the commonsense level [1]. Illustrated in Figure 2, our framework consists of 3 key modules: (1) knowledge component extraction, (2) counter-commonsense query construction, and (3) image and question-answer (QA) pair synthesis. This framework streamlines the generation of counter-commonsense inputs from scratch and is modularly designed to facilitate the addition of new conflict categories and QA formats in the future. We demonstrate the application of our framework by developing the ConflictVis benchmark, focusing on the aspects of Subject, Action, and Place. The benchmark consists of **374** original images with **1,122** high-quality QA pairs spanning two conflict targets and three question types, all manually verified.

Using the crafted benchmark, we evaluate nine representative MLLMs from five model families, providing insights into model behaviors, the causes of conflicts, and effective approaches to mitigate the negative effects of conflicts. Notably, when facing knowledge conflicts, MLLMs tend to over-rely on their parametric knowledge for the answer, especially among Yes-No questions and action-related problems. For instance, the top commercial model, `Claude-3.5-Sonnet`, exhibits a memorization ratio on parametric knowledge of 43.6% on Yes-No questions, substantially higher than results on more complex Open-Ended questions. Regarding conflict type, the average memorization ratio for counter-commonsense action problems is 23.8%, 10.4 percentage points higher than that for the place problems. Our detailed analysis of the failure cases reveals that MLLMs generally underutilize visual information, relying on parametric knowledge to infer the answer based on textual clues. Drawing on these observations, we assess several existing improvement methods to enhance the impact of visual context in answer generation. Interestingly, although Chain-of-Thought prompting improves the reasoning abilities, it guides MLLMs to utilize parametric knowledge more during rationalization, often resulting in contradictory conclusions or refusals. In response, we propose "Focus-on-Vision" (FoV) prompting to directly instruct MLLMs to prioritize visual information, which markedly improves the model's performance. Despite advancements by various mitigation approaches, the vision-knowledge conflict remains a persistent challenge.

Our main contributions are summarized below:

- We introduce an innovative framework to automatically construct counter-commonsense benchmarks from the ground up. This framework allows the flexible definition of conflict categories and QA formats, facilitating the large-scale creation of conflict samples with minimal human effort.
- We present ConflictVis, a pioneering diagnostic benchmark specifically designed to evaluate the commonsense level vision-knowledge conflicts in MLLMs. The benchmark is meticulously validated by human experts to guarantee the quality of data.
- We benchmark nine representative MLLMs and evaluate the effectiveness of several improvement methods in addressing conflicts. Through this analysis, we demonstrate the significance of the vision-knowledge conflict.

2 Related Work

Knowledge Conflicts. Knowledge conflicts in LLMs can be divided into three categories (Xu et al., 2024c): within the retrieved context (Chen et al., 2022), within model's parametric knowledge (Huang et al., 2023), and between the context and the model's parametric knowledge (Xie et al., 2024; Wu et al., 2024a; Su et al., 2024). These conflicts can lead to incorrect or inconsistent responses, undermining the model's trustworthiness (Xu et al., 2024c; Xie et al., 2024). In MLLMs, these conflict

[1] For example, a commonsense knowledge statement is "babies cry when they are hungry", as opposed to a factual knowledge statement like "the Eiffel Tower is 330 meters tall."

types expand to multimodal inputs, where conflicts can occur between the input image and the text instruction (Liu et al., 2024e; Han et al., 2024), or when the image contains counterfactual information that contradicts the model's parametric knowledge (Liu et al., 2024e; Guan et al., 2024).

To evaluate the conflicts between visual information and the parametric knowledge in MLLMs, HallusionBench (Guan et al., 2024) consists of manually edited informational graphics, forming (normal, counterfactual) image pairs used to assess model consistency through Yes-No questions. AutoHallusion (Wu et al., 2024b) introduces an automated approach to generate counterfactual scenarios by altering correlated objects in images, utilizing Yes-No questions to probe object existence and spatial relationships. In the context of commonsense knowledge, PhD (Liu et al., 2024e) generates counter-commonsense images through manual collection and synthesis, employing short-answer questions for assessment. In contrast, our proposed benchmark features an automated framework and a broad array of question types and conflict targets, enabling a more scalable and comprehensive evaluation of MLLMs.

Hallucination in MLLMs. Hallucination in MLLMs refers to the situation where the model generates descriptions that conflict with the given visual context. MLLMs' hallucinations are typically categorized based on the type of incorrect information, such as non-existent objects, incorrect object attributes, and inaccurate object relations (Liu et al., 2024a). Relevant research has mainly focused on two key areas: developing benchmarks and metrics to detect hallucinations (Rohrbach et al., 2018; Li et al., 2023c; Liu et al., 2023a; Wang et al., 2023) and proposing strategies to mitigate them (Liu et al., 2023a; Leng et al., 2024; Huang et al., 2024; Liu et al., 2024f).

In the context of knowledge conflicts, hallucination occurs when the model gives precedence to its intrinsic knowledge rather than the visual context (Liu et al., 2023a; Guan et al., 2024; Liu et al., 2024e,a). Our study explores this issue by simulating knowledge conflicts and evaluating the effectiveness of various hallucination-mitigation approaches in resolving the conflicts.

Benchmarks for MLLMs. Traditional vision-language benchmarks are designed to assess independent skills, including image captioning (Chen et al., 2015), visual grounding (Rohrbach et al., 2016), and visual question answering (Antol et al., 2015; Hudson and Manning, 2019). However, with the emergence of MLLMs, there is a growing need for more comprehensive and tailored benchmarks. The strong zero-shot abilities and advanced language generation skills exhibited by MLLMs make traditional benchmarks insufficient, as they may not account for the diversity of responses or the full range of MLLM capabilities. To address these limitations, researchers have developed more complex benchmarks to evaluate MLLMs across a wider range of tasks (Yue et al., 2024; Fu et al., 2023; Li et al., 2023a; Liu et al., 2023c). Meanwhile, diagnostic benchmarks have also been built to evaluate specific challenges or traits in MLLMs, such as hallucination (Li et al., 2023c; Liu et al., 2023a), social bias (Howard et al., 2024), and model safety (Liu et al., 2023b). CONFLICTVIS is a pioneering analytical benchmark that presents conflicting visual contexts to challenge the model's commonsense knowledge, enabling deeper investigations into model performance in the presence of vision-knowledge conflicts.

3 CONFLICTVIS Benchmark

This section outlines the framework and the data utilized to construct the CONFLICTVIS benchmark.

3.1 Automated Framework

Commonsense knowledge refers to the information generally accepted by the majority of people about everyday life, encompassing practical principles on how the world functions (Singh et al., 2002). Based on the widely-used image captioning dataset (Chen et al., 2015), we explore a specific type of commonsense knowledge encapsulated as a triplet $\langle \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{p} \rangle$, where s is the Subject, a is the Action, and p is the Place of the action or subject. For example, the statement "the waitress (Subject) washing dishes (Action) in the kitchen (Place)" illustrates this format. In this structure, the Subject outlines the main object's appearance, the Action describes the primary activity and its interactions with other relevant objects, and the Place highlights the background objects and setting. This three-part framework efficiently captures the vital details that are accurately reflected in an image.

We next describe our approach for generating images and corresponding QA pairs that challenge the commonsense knowledge in MLLMs. In general, we construct triplets of low co-occurring concepts

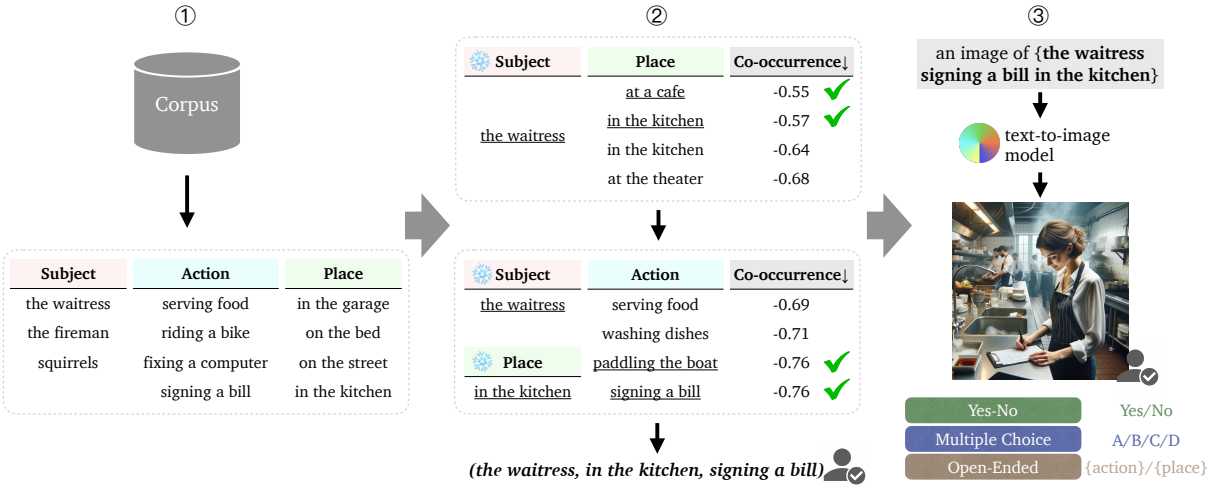


Figure 2: An automated framework with human quality control to construct images and question-answer pairs that conflict with commonsense knowledge.

to serve as counter-commonsense queries for the multimodal input generation. Our framework includes three stages, as depicted in Figure 2. We detail each stage below.

Extract Knowledge Components Drawing inspiration from (Li et al., 2021) on the creation of a compositional generalization test set, we identify Subject, Action, and Place phrases as the most frequent elements in the corpus and build compounds based on them. To this end, we first employ a transformer model pipeline (Honnibal et al., 2020) to extensively annotate the syntactic labels, including dependency (DEP), Part-of-Speech (POS), and Named Entity (NE). Then, we extract Subject, Action, and Place phrases from the corpus based on the predefined linguistic rules (See Appendix A.1). To ensure the data quality, our framework selects the top N phrases from each category based on the frequency. It further refines the data by removing named entities from Subject and Place phrases and consolidating similar expressions (e.g., "a doctor" and "the doctor") by retaining the most frequently occurring variant. This process results in three curated phrase lists for subsequent processing.

Construct Counter-commonsense Query We aim to construct scenes with one single anomalous component (i.e., target) that seldom co-occurs with the others (i.e., the context). This objective can be formalized into two key requirements: (1) The context components should exhibit a high co-occurrence level, acting as a common background, and (2) The target component should display a low co-occurrence with the given context, representing the anomaly. To identify strongly co-occurring context pairs, our framework first groups context components by the target category, e.g., $\textbf{Context}_{Action}$ consists of (Subject, Place) pairs. These groupings help to develop the focus of the questions in the next section. We omit $\textbf{Context}_{Subject}$ to prevent ambiguity and mitigate potential ethical concerns related to subject identity. Next, our framework enumerate all the context combinations within the each group and computes their Normalized Pointwise Mutual Information (NPMI) scores.

$$\text{PMI}(C_X; C_Y) \equiv \log_2 \frac{P(C_X, C_Y)}{P(C_X)P(C_Y)} \quad (1)$$

$$\text{NPMI}(C_X; C_Y) \equiv \frac{\text{PMI}(C_X; C_Y)}{-\log_2 P(C_X, C_Y)} \quad (2)$$

A higher NPMI score indicates a stronger co-occurrence between the two components. For each context group, the framework selects Top-K context pairs with the highest NPMI scores to form a candidate pool. For each context pair in the pool, the framework generates counter-commonsense triplets by selecting a target component that is unusual given the context. Concretely, the framework evaluates the co-occurrence between each target T and the context pair $\mathbf{C}$ using the NPMI score (i.e. $\text{NPMI}(T; \mathbf{C})$). For each context pair, the Top-M targets with the lowest NPMI scores are retained to construct counter-commonsense queries. To better align the model knowledge, the framework queries an LLM trained on large-scale web data to estimate the probability $P(\cdot)$, and a manual review is conducted after the query generation to ensure quality (See Appendix A.2 for more details).

Question	Content	Answer
Yes-No	Is the waitress in the kitchen signing a bill?	Yes
Multiple-Choice	Question: What is the waitress doing in the kitchen? Options: (A) washing dishes. (B) riding a bike. (C) starting a fire. (D) signing a bill.	D
Open-Ended	What is the waitress doing in the kitchen?	signing a bill

Table 1: Example question-answer pairs generated by our framework.

Generate Multimodal Inputs Building on the constructed queries, our framework generates three types of questions, Yes-No, Multiple-Choice, and Open-Ended, along with their corresponding answers. To achieve this, the framework employs predefined question templates and fills in the relevant components accordingly. Example question-answer pairs are provided in Table 1. To generate corresponding images, our framework concatenates the triplet components into a caption-like expression and employs a prompt template to query a text-to-image model, as illustrated in Figure 2. Following image generation, human annotators perform quality control by filtering out low-quality images, such as those that appear distorted or misaligned with the input prompt. Detailed image filtering guidelines, along with illustrative examples, are provided in the Appendix B.3.

3.2 CONFLICTVIS Benchmark Construction

We present how we use the automated framework to build CONFLICTVIS. Our input corpus consists of the top 100K sentences from the Open Mind Common Sense (OMCS) dataset (Singh et al., 2002). From this dataset, we extract and retain the 100 most frequent Subject phrases ($N_S = 100$) and the 150 most frequent Action and Place phrases ($N_A = N_P = 150$). Using feedback from the LLM `Vicuna-1.5-13b` (lmsys, 2024), we select the top 3 phrases ($K = 3$) with the highest NPMI scores to create context pairs for each subject. Next, for each context pair, we choose the top 3 targets ($M = 3$) with the lowest NPMI scores to assemble the candidate set of counter-commonsense triplets. After manually filtering out unexpected combinations, the remaining triplets are used to query `DALL·E 3` (OpenAI, 2024a) for image generation. Subsequently, human annotators review and remove low-quality images (See Appendix B). Following this two-stage generation and filtering process, the statistics of the final benchmark are summarized in Table. 2. In total, CONFLICTVIS comprises 1,122 test samples, spanning two conflict targets and three question types, all verified by human experts.

Target	#Triplets	#Images	#QAs
Action	188	171	513
Place	156	203	609
Total	344	**374**	**1122**

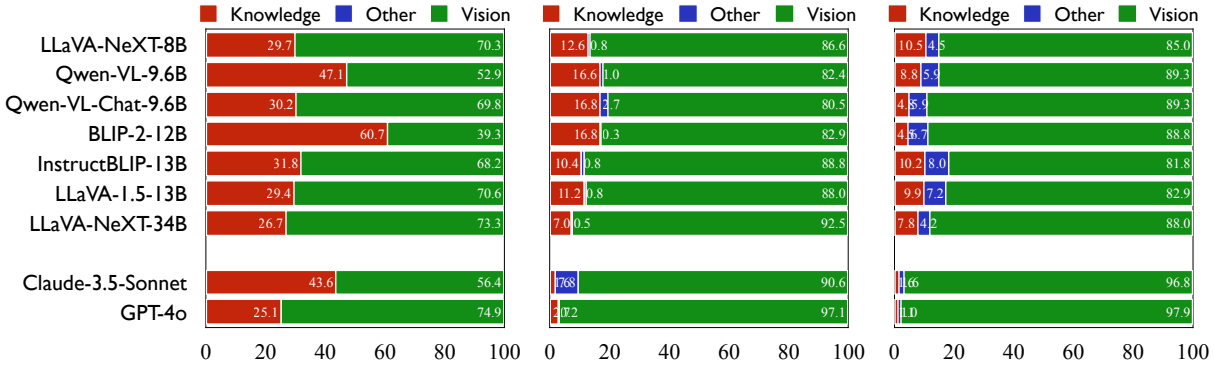
Table 2: Statistics of the constructed benchmark.

4 Experiment

4.1 Setup

Models To explore the behavior of MLLMs when encountering vision-knowledge conflicts, we perform a comprehensive evaluation on 9 MLLMs including 7 representative open-source MLLMs ranging from 8B to 34B, and 2 state-of-the-art commercial MLLMs. This evaluation covers the following five model series: `LLaVA` (8B, 13B, 34B) (Liu et al., 2024b,c), `BLIP-2` (12.1B, 13B) (Li et al., 2023b; Dai et al., 2023), `Qwen-VL` (9.6B) (Bai et al., 2023), `GPT-4o` (OpenAI, 2024b) and `Claude-3.5-Sonnet` (Anthropic, 2024). The diversity of model architectures and parameter sizes helps to enhance the generalizability of experiment results.

Evaluation We use Accuracy and Memorization Ratio (MR) (Longpre et al., 2021) as the main evaluation metrics in our experiment. Both metrics require classifying the model's responses into different categories. For Yes-No and Multiple-Choice questions, where there is a unique answer, we use exact matching for classification. For Open-Ended questions, due to the complexity of the task (i.e., evaluating both textual and visual relevance and correctness) and the instability of LLM evaluation (Stureborg et al., 2024), we rely on human annotators to perform the classification.



(a) Yes-No Questions (b) Multi-Choice Questions (c) Open-Ended Questions

Figure 3: MLLM response distributions on different question types.

4.2 Sanity Test

Do the instances in ConflictVis really conflict with MLLMs' commonsense knowledge? Before evaluating MLLMs' performance on ConflictVIS, we first verify whether the counter-commonsense cases in ConflictVis genuinely create conflicts with the models' commonsense knowledge. To this end, we use the textual inputs in ConflictVis and query the model with the following prompt format:

```
Based on common sense, is it possible
for [context] [target]?
```

where the context and target are filled with the specific phrases from the input. For example, the query for the instance in Figure 2 is "*Based on common sense, is it possible for the waitress in the kitchen signing a bill?*" The model's responses ("Yes", "No", and others) are categorized as "accept" (indicating no conflict), "negate" (indicating a conflict), and "other" (where the model does not give a direct answer). The results, shown in Figure 4, demonstrate that the vast majority of cases in ConflictVis indeed present valid knowledge conflicts that can be identified by MLLMs, with conflict rates ranging from 80.5% for `GPT-4o` to 99.2% for `LLaVA-1.5-13B`.

4.3 Benchmarking MLLMs

In this section, we assess how well the MLLMs handle conflicts between visual information and parametric knowledge using our ConflictVis benchmark. Following Longpre et al. (2021), we compare model predictions (i.e., answers) with and without the inclusion of the image input and quantify the extent to which the model's predictions are influenced by the presence of the image. Specifically, we compare the predictions in the current experiment, where the image is provided, to those obtained without the image using the sanity test prompts, and classify the current predictions as:

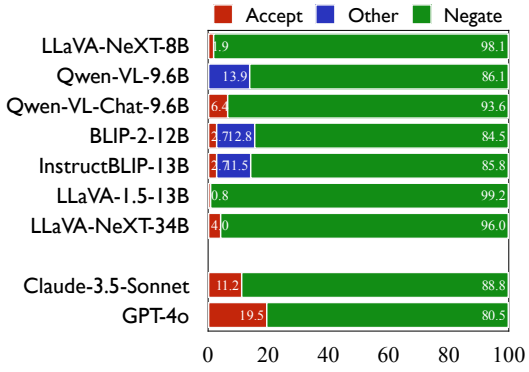


Figure 4: MLLM responses to sanity test inputs.

- Aligned with the prediction without the given image (**Knowledge**, P_K): This outcome suggests that the model generates its answer mainly based on its parametric knowledge rather than the visual information presented.

- Aligned with the visual information (**Vision**, P_V): This indicates that the model is capable of adapting to new information, adjusting its output to match the visual input. Such outputs are categorized as the correct answer.

- A different answer altogether (**Other**): This demonstrates that the model can modify its output based on visual input, though the result does not perfectly align with the visual information.

Figure 3 presents the results. Ideally, the model should **only** output the Vision answer, supported by visual information, rather than the Knowledge answer derived from textual training, or any Other

answers. However, both open-source and commercial MLLMs revert to producing the Knowledge answer, ignoring the visual information, to varying degrees. In general, commercial MLLMs perform better than the open-source counterparts, particularly in handling open-ended questions. Among the open-source models, `LLaVA-NeXT-34B` achieves the highest prediction accuracy at 84.6%, while the top commercial model, `GPT-4o`, reaches 89.9%. Moreover, our analysis reveals that MLLMs' uncertainty in answer generation is significantly higher on our CONFLICTVIS compared to the classic VQA benchmark, indicating a greater challenge to the model performance (See Appendix C.2).

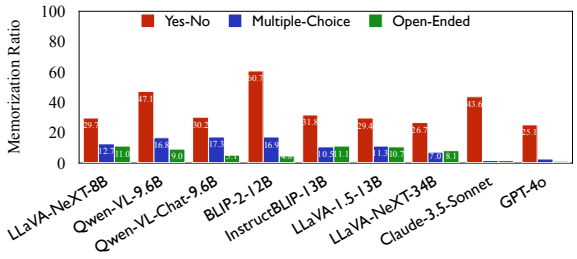


Figure 5: Memorization ratio of MLLMs on different question types.

MLLMs are more likely to overly rely on parametric knowledge for Yes-No questions. One interesting observation from Figure 3 reveals that MLLMs exhibit poorer performance on straightforward Yes-No questions. Notably, the open-source `BLIP-2-12B` achieves a low accuracy of 39.3%, while the commercial `Claude-3.5-Sonnet` manages only 56.4%.

We leverage Memorization Ratio (MR) (Longpre et al., 2021) as a metric to empirically estimate the the extent to which the model adheres to its parametric knowledge in the presence of conflicting information.

$$MR = \frac{P_K}{P_K + P_V}. \quad (3)$$

Figure 5 illustrates the MR across different question types. Notably, the MR values for Yes-No questions are significantly higher than the other two question types across all MLLMs, indicating that MLLMs tend to overly rely on parametric knowledge when answering Yes-No questions. We hypothesize that this could be due to the format of Yes-No questions, which directly present counter-commonsense expressions (e.g., "Is the waitress in the kitchen signing a bill?") to the MLLMs, while other question types have a less direct phrasing. This counter-commonsense query may discourage the model from thoroughly analyzing the visual input, causing it to immediately output a negation.

Counter-commonsense actions are more challenging for MLLMs. Our CONFLICTVIS evaluates two distinct types of conflict targets: counter-commonsense actions (e.g., the waitress in the kitchen signing a bill) and places (e.g., a doctor conducting an experiment at the theater). As shown in Figure 6, counter-commonsense action problems pose more of a challenge than places do. Specifically, MLLMs consistently exhibit a lower accuracy on counter-commonsense action problems, with an average output accuracy of 73.9%, notably lower than the 85.2% recorded for place problems. Similarly, the average Memorization Ratio for action problems is markedly higher at 23.8%, compared to 13.4% for place problems. This discrepancy could be attributed to the richer visual context typically available for identifying places (e.g., numerous background elements suggesting the location "at the theater"), in contrast to the relatively sparse and fine-grained visual cues required to recognize actions (e.g., subtle hand posture for "signing a bill"). Detailed results for each model can be found in the Appendix C.4.

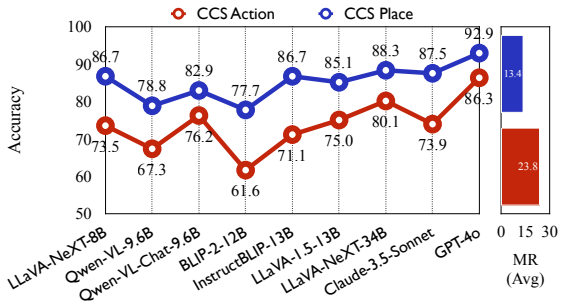
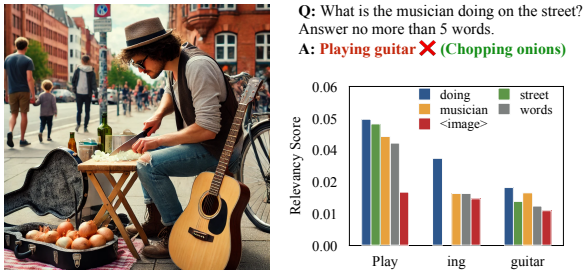


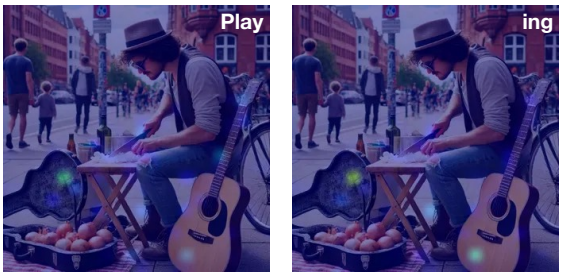
Figure 6: MLLM accuracy on Counter-Commonsense (CCS) Action problems and CCS Place problems.

4.4 Mitigating Vision-Knowledge Conflicts

Underutilization of visual information Our primary experiment demonstrates that when MLLMs encounter visual inputs that contradict their pre-existing knowledge, around 18% of the responses generated are not aligned with the visual context but rather reflect the model's prior knowledge. To further investigate this discrepancy, we leverage the tool proposed by Stan et al. (2024) to analyze input-output relevancy scores, which identify the most relevant parts of the input to the generated output



(a) Input-output relevancy score bar plot



(b) Input image-output token relevancy map

Figure 7: Failure case analysis of input-output relevancy. The visual information is underutilized.

based on model attention. Figure 7(a) illustrates a case in which the model `LLaVA-1.5-13B` fails to accurately answer an open-ended question, representing a common failure pattern. In general, the model assigns greater weight to the textual input than to the visual context, resulting in an inaccurate response. In this case, the most attended tokens, "musician", "doing", "street"}, guide the model to generate "playing guitar" based on its parametric knowledge, while the image tokens, denoted by the <image> symbol in the bar plot, play a less significant role in answer generation. The image-output relevancy maps in Figure 7(b) further indicate that the model may not effectively leverage visual information during answer generation. In such cases, the model is more susceptible to vision-knowledge conflicts, as the visual input is largely underutilized. (More cases can be found in Appendix D.1.)

Effectiveness of Existing Methods Based on the above observations, a straightforward strategy for improvement is to enhance the influence of visual context during the answer generation process. We consider several improvement approaches:

- *Visual Contrastive Decoding (VCD)* (Leng et al., 2024) contrasts the output distributions derived from original and noisy visuals to ensure the generated content adheres to the visual inputs.
- *Pay Attention to Image (PAI)* (Liu et al., 2024f) adaptively adjusts and amplifies the attention weights assigned to image tokens, thereby giving greater prominence to visual elements.
- *Vision-Centric Reasoning (VR)* employs structured intermediate reasoning steps to thoroughly analyze visual inputs, achieved by Chain-of-Thought (CoT) prompting (Wei et al., 2022b) or supervised finetuning on the long CoT multimodal reasoning dataset (Xu et al., 2024b).

Furthermore, we propose a direct prompting technique, termed **Focus-on-Vision** (FoV) prompting, which explicitly instructs MLLMs to prioritize visual information. This can be seamlessly integrated into current MLLMs using the following template: `[textual query] Please focus on the visual information.`

Model	YN	MC	OE	Avg.
LLaVA-1.5-13B	70.6	88.0	82.9	80.5
+ VCD	72.7	89.3	84.2	82.1
+ PAI	85.6	88.8	86.1	**86.8**
+ VR (CoT)	38.0	89.8	76.7	68.2
+ VR (SFT)	64.0	87.4	88.5	80.0
+ FoV	82.9	89.0	81.8	84.6
Qwen-VL-Chat	69.8	80.5	89.3	79.9
+ VCD	82.4	79.9	85.6	82.6
+ VR (CoT)	79.7	65.8	77.8	74.4
+ VR (SFT)	69.3	87.4	88.0	81.6
+ FoV	82.4	83.2	87.4	**84.3**
LLaVA-NeXT-34B	73.3	92.5	88.0	84.6
+ VR (CoT)	43.6	87.2	72.5	67.7
+ FoV	85.8	92.5	89.8	**89.4**
GPT-4o	74.9	97.1	97.9	89.9
+ VR (CoT)	66.0	98.7	93.6	86.1
+ FoV	75.9	96.5	98.9	**90.5**

Table 3: MLLM accuracy under different improvement methods. YN: Yes-No, MC: Multiple-Choice, OE: Open-Ended.

Table 3 presents the results of the evaluated improvement methods. For VCD and PAI, we follow the methodologies outlined in the original papers and apply each technique to the corresponding applicable MLLMs, and observe performance improvements for both methods. To implement Vision-Centric Reasoning (VR), we explore two approaches: (1) applying Chain-of-Thought (CoT) prompting, and (2) fine-tuning models on the multimodal reasoning dataset introduced by Xu

et al. (2024b). However, the results show minimal performance gains from finetuning, and vanilla CoT prompting can even degrade model accuracy. Upon analyzing failure cases, we find that teaching MLLMs to reason in a straightforward, step-by-step manner can lead to increased reliance on textual knowledge at test time. For example, our case study (in Appendix D.2) suggests that during CoT reasoning, models tend to rationalize the inputs using their existing commonsense knowledge. However, counter-commonsense inputs are inherently difficult to rationalize or explain. As the output text accumulates, it may become increasingly difficult for the model to accurately interpret the visual input given its own generated explanations. This often results in self-contradictory answers or refusals to answer. In contrast, our Focus-on-Vision (FoV) prompting mitigates this issue by explicitly directing the model's attention to the visual input and consistently boosts model performance. Nevertheless, none of the approaches can entirely resolve the vision-knowledge conflict, especially for open-source models. For instance, `LLaVA-1.5-13B` and `Qwen-VL-Chat` still exhibit notable error rates of 13.2% and 15.7%, respectively, even when equipped with the most effective improvement method. This highlights the profound challenge posed by vision-knowledge conflict.

5 Conclusion

In this paper, we develop an automated framework and construct the ConflictVis benchmark to systematically evaluating and understanding the nuances of vision-knowledge conflicts in MLLMs at the commonsense level. Our experiments across nine representative MLLMs reveal a tendency toward over-reliance on parametric knowledge, especially among Yes-No and action-related questions. We further evaluate three improvement methods and design a new prompting technique to mitigate the conflicts, but the problem remains notable. Collectively, our insights lay the groundwork for future research to develop more reliable MLLMs.

Acknowledgment

This paper was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515010145) and the Shenzhen Science and Technology Program (Shenzhen Key Laboratory Grant No. ZDSYS20230626091302006).

Limitations

In this study, we focused on evaluating the reliably of multimodal large language models (MLLMs) under vision-knowledge conflict. While our work provides insights into the model performance and effectiveness of various mitigation approaches, it's important to acknowledge certain limitations: (1) Limited root cause analysis: while we conducted failure case analysis using relevancy maps, the fundamental causes of the unexpected behaviors exhibited by MLLMs under vision-knowledge conflicts remain underexplored. We hypothesize that these behaviors may stem from biases in the training data, where MLLMs are predominantly aligned with image-text pairs that conform to commonsense expectations encoded during text pretraining, with limited exposure to counter-commonsense scenarios. To further validate this hypothesis, future work could involve controlled experiments comparing models trained with and without curated datasets designed for counter-commonsense alignment, in order to investigate the resulting behavior differences. (2) Unique probability model: due to budget constraints, our framework relies on `Vicuna-1.5-13B` to estimate co-occurrence probabilities, which may introduce model-specific features into the benchmark construction. Although our quality control processes help mitigate this issue, future work can explore more balanced probability estimation approaches, such as averaging probabilities across multiple models.

Ethics Statements

While our benchmark is designed to be safe and useful for research purposes, the developed framework for generating counter-commonsense inputs warrants careful consideration regarding its potential applications. In certain contexts, the framework may produce images that deviate from conventional norms, and if misused, could result in cases that some users may perceive as harmful or offensive. To mitigate such risks, we recommend implementing precautionary measures, including the use of a carefully curated and pre-filtered input corpus, as well as incorporating a manual verification process. Additionally, we encourage the use of text-to-image generative models through their official platforms and in accordance with their terms of use, which can help effectively leverage built-in safety mechanisms and reduce the likelihood of generating inappropriate or harmful content.

References

Anthropic. 2024. claude-3-5-sonnet. https://www.anthropic.com/news/claude-3-5-sonnet. Accessed: 2024-08-01.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341, Singapore. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Accessed: 2024-06-15.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in mllms. *arXiv preprint arXiv:2402.03757*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Phillip Howard, Anahita Bhiwandiwalla, Kathleen C Fraser, and Svetlana Kiritchenko. 2024. Uncovering bias in large vision-language models with counterfactuals. *arXiv preprint arXiv:2404.00166*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing.

2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. 2021. On compositional generalization of neural machine translation. In *ACL*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 317–324.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024c. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024e. Phd: A prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024f. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.

X Liu, Y Zhu, J Gu, Y Lan, C Yang, and Y Qiao. 2023b. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. *arXiv preprint arXiv:2311.17600*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

lmsys. 2024. Vicuna-1.5. Accessed: 2024-06-24.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *EMNLP*.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *Preprint*, arXiv:2306.09093.

Meta AI. 2024. Introducing meta llama 3. Accessed: 2024-09-18.

Meta AI. 2025. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Accessed: 2025-02-05.

OpenAI. 2022. Introducing chatgpt. Accessed: 2024-09-18.

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

OpenAI. 2023b. Gpt-4v(ision). Accessed: 2024-09-18.

OpenAI. 2024a. Dall-e 3. Accessed: 2024-04-30.

OpenAI. 2024b. Gpt-4o. Accessed: 2024-06-15.

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.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 817–834. Springer.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings*, pages 1223–1237. Springer.

Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kevin Wu, Eric Wu, and James Zou. 2024a. How faithful are rag models? quantifying the tug-of-war between rag and llms' internal prior. *arXiv preprint arXiv:2404.10198*.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, and 1 others. 2024b. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. *arXiv preprint arXiv:2406.10900*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and 1 others. 2024a. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024b. Llava-cot: Let vision language models reason step-by-step. *Preprint*, arXiv:2411.10440.

Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024c. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Benchmark Construction Details

A.1 Linguistic rules to extract knowledge components

To systematically extract key knowledge components from textual inputs, our framework categorizes phrases into three main types, Subject, Action, and Place, based on syntactic and semantic annotations. Specifically, the framework first utilizes the transformer pipeline `en_core_web_trf` from spaCy (Honnibal et al., 2020) to annotate the sentences, and then performs the extraction based on the following procedures:

- **Subject Phrases**: Our framework examines noun chunks, selecting those where the head noun functions as the nominal subject, indicated by dependency labels "nsubj" or "nsubjpass".

- **Action Phrases**: Our framework identifies the verb token ("VERB") and then examines its dependents for direct objects marked as "dobj". Upon finding such objects, the framework includes any determiners or compound modifiers present to form the complete action phrase. Verbs are converted to the "VBG" form to align with the common practice in image captions.

- **Place Phrases**: Distinguishing place phrases from other prepositional phrases relies on more than just syntactic labels. Accordingly, the framework employs semantic role labeling techniques by using the model `structured-prediction-srl` from AllenNLP (Gardner et al., 2017) to annotate each word's semantic role. The framework then extracts the phrases with a location tag ("ARGM-LOC") that contains 3 to 4 words.

In Table 4, we summarize the main linguistic rules employed to identify specific components.

A.2 Probability Calculation

To calculate the joint probabilities for multiple co-occurring components (e.g. $P(C_X, C_Y)$ or $P(C_X, C_Y, T)$), our framework forms concatenated expressions of the concept phrases (e.g., "the waitress in the kitchen" or "the waitress in the kitchen signing a bill") and inputs them into a pre-trained LLM to obtain the generative probability. Utilizing an LLM here offers an advantage over frequency counting for addressing the issue of zero frequencies for rare concept compositions, which often arises with limited dataset size. Since large language models are trained on large-scale data to learn the statistical tendencies of human language, their generative probability provides an approximation of the likelihood of certain phrase appearing in reality. For the probability of each individual component, we use its relative frequency within its category (Lin, 1999).

A.3 Question Generation Details

With the constructed counter-commonsense triplets, our framework generates three types of questions:

- **Yes-No Question**: For counter-commonsense Actions and Places, our framework utilizes templates "Is/are [Subject Place] [Action]?" and "Is/are [Subject Action] [Place]?", respectively. The correct response to this type of question is typically "Yes".

- **Multiple-Choice Question**: For these questions, the templates are "What is/are [Subject] doing [Place]?" and "Where is/are [Subject] [Action]?" for counter-commonsense Action and Place, respectively. For a predetermined option count m, our framework selects $(m-1)$ seemingly relevant but incorrect targets from the candidates. To this end, the framework partitions candidates into $m-1$ bins $\{B_1, ..., B_{m-1}\}$ based on their NPMI scores to the context pair. We randomly sample one candidate from each bin, together with the counter-commonsense target, to formulate the question's multiple-choice options. We randomly shuffle the option list to eliminate any positional bias.

- **Open-Ended Question**: The question text in the multiple-choice question is reused, but no list of options is included. To guide the model toward generating a brief and specific response, we add the constraints like "Answer with a single phrase", with slight adjustment according to the specific prompt for different MLLMs.

A.4 Image Generation Details

As described in Section 3.1, our framework converts each triplet into a caption-like expression (e.g., "a waitress signing a bill in the kitchen") and inserts it into the prompt template "An image of {expression}" to query the DALL-E 3 model. In our implementation, we used the DALL-E 3 API with the image resolution set to 1024×1024 and the quality parameter configured to "standard".

Category	Main Requirement	Example
Subject	`p in noun_chunks && p.root.dep in ["nsubj", "nsubjpass"]`	a baby, a chef
Action	`p[0].pos == "VERB" && p[-1].dep == "dobj"`	fixing a computer
Place	`p.srl = "ARGM-LOC" && 3 <= p.length <= 4`	at the bookstore

Table 4: Linguistic rules to extract different components in the triplet.

A.5 Examples in ConflictVis

In Figure 8, we demonstrate some samples generated by our framework. The conflict target is highlighted in red.

(a) the hunter in the forest taking a note

(b) the bear in the woods mailing a letter

(c) a chef making bread at the theater

(d) the knight riding a horse in a locker room

Figure 8: Example counter-commonsense queries and their corresponding images generated in ConflictVis benchmark.

A.6 Licenses

The Open Mind Common Sense database is under the Creative Commons license. The spaCy transformers pipeline `en_core_web_trf` is under the MIT license, and the AllenNLP model `structured-prediction-srl` is under the Apache License 2.0. The probability model `Vicuna-1.5-13b` is released under a non-commercial license, and the text-to-image model `Dall·E 3` API is subject to OpenAI's Terms of Use. We use these artifacts exclusively for non-commercial research purposes, in accordance with their intended use.

B Human Quality Control

This section outlines the quality control procedures within our automated framework. As shown in Figure 2, there are two required quality control points: (1) after generating knowledge triplets, and (2) after generating images. In practice, to reduce the workload for annotators at the first point, we subdivide it into three steps: (i) after extracting knowledge components, (ii) after generating context pairs, and (iii) after generating knowledge triplets. In this section, we first describe the quality control guidelines applied at each stage of the framework, followed by the principles used for annotating open-ended responses.

B.1 Human Annotators

Our annotation team comprises four Ph.D. students specializing in computer science, all of whom are proficient in English. All annotators were fairly compensated based on their total working hours. Prior to annotation, we conducted training sessions to inform them that their tasks might include offensive content or identifying information, and clarified that their contributions would be used solely for research purposes to study model behavior.

The annotation results were based on agreement among annotators. To ensure the reliability and objectivity of the evaluation process, each annotator independently performed the annotation tasks. Final labels were determined through majority voting, with group discussions held to resolve any ties.

B.2 Filtering Knowledge Triplets

Filtering knowledge components This step removes semantically abstract phrases in the Subject, Action, and Place categories, which are difficult to visualize. Annotators classify each phrase based on its concreteness:

- **Concrete (1)**. The phrase represents tangible objects or actions.
- **Abstract or Unsafe (0)**. The phrase represents

abstract concepts or contains offensive words or identifying information.

Category	Label	Phrases
Subject	0	the statement, the story, the stock market, words
Subject	1	a patient, kids, politicians, whales, a sailor
Action	0	understand the event, lose weight, have sex
Action	1	take a shower, use a computer, play chess
Place	0	in the event, in the newspaper, in your life
Place	1	on the ground, on the sidewalk, by the fire

Table 5: Labeling examples for knowledge components.

Table 5 provides some labeling examples at this step. In this step, the annotators perform labeling starting from the top 1000 most frequent phrases from each category in descending order, and keep the first 100 Subject phrases and the first 150 Action and Place phrases that meet the requirements.

Filtering context pairs The target of this step is to remove the unsatisfying phrases that remain in the candidate list after the automatic pre-filtering. To this end, our annotators conduct a binary classification on the context pairs based on their commonness in the real world. Specifically, the annotators are asked to examine the context tuples **without** the knowledge of their co-occurrence scores and label them as:

- **Common (1)**. The context tuple depicts a typical scene in the real world.
- **Uncommon (0)**. The context tuple depicts an unusual scene in the real world.

To better illustrate, we provide some labeling examples from our human annotators in Table 6:

In this step, our annotators thoroughly label candidate context pairs within the Top K range provided by the framework. Context pairs labeled as **1** are retained for the next stage of generation.

Filtering knowledge triplets After the previous filtering steps, the workload for this phase is significantly reduced. The goal here is to eliminate any unexpected triplets that are not captured in earlier

Category	Label	Phrases
(Subject, Action)	0	(the cat, walking the dog), (a butcher, playing cards)
(Subject, Action)	1	(the cat, climbing a tree), (a butcher, slaughtering a pig)
(Subject, Place)	0	(cows, on the beach), (a gardener, in the desert)
(Subject, Place)	1	(cows, on a farm), (a gardener, in a greenhouse)

Table 6: Labeling examples for context pairs.

filtering. Without access to the exact co-occurrence score between the context and target phrases, annotators perform a binary classification based on the commonness of the triplet's expression. Specifically, the annotators are asked to label each triplet in the candidate list as:

- **Common (0)**. The context tuple depicts a typical scene in the real world.
- **Uncommon (1)**. The context tuple depicts an unusual scene in the real world.

Category	Label	Phrases
((Subject, Place), Action)	0	((the waitress, at the bar), using a vcr)
((Subject, Place), Action)	1	((the waitress, at the bar), hitting a deer)
((Subject, Action), Place)	0	((politicians, drinking alcohol), at the theater)
((Subject, Action), Place)	1	((politicians, drinking alcohol), in the oven)

Table 7: Labeling examples for knowledge triplets.

Again, we provide some labeling examples from

our human annotators in Table 7 for a better illustration. In this step, our annotators label all the candidate context-target triplets in the Top M range provided by the framework, and triplets with label **1** are used for subsequent image generation.

B.3 Filtering Images

The objective of this step is to select high-quality images that align closely with the text prompt. Each annotator is provided with an image and its corresponding text query, and evaluates the image based on two criteria:

- **Alignment with text prompt (0 or 1)**. The image should contain the key objects, actions, and background described in the text. The main focus should clearly represent the scene without ambiguity or misinterpretation.
- **Image quality (0 or 1)**. The image should be clear, free from significant distortions, artifacts, or unnatural effects like warping, blurring, or pixelation.

With these two guidelines, the annotators extensively label all the generated images, and the images with a total score of **2** are kept for the subsequent stages. In Figure 9, we demonstrate some sample images that are below our quality standard. In this stage, we generate about 830 images in total and keep the best 374 images for the benchmark construction.

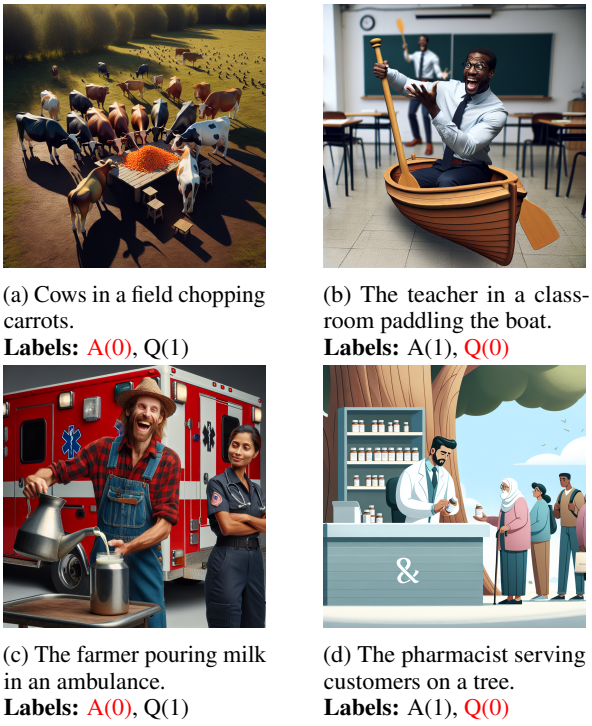
B.4 Guidelines for Manual Labeling Open-Ended Responses

To label the correctness of MLLMs responses for the open-ended questions, we design our evaluation criteria following (Xu et al., 2024a). Specifically, the annotators are asked to determine the correctness of MLLMs' responses based on two criteria:

- **Relevancy (0 or 1)**: The content of the response is reflected in the image and addresses the focus of the question.
- **Responsiveness (0 or 1)**: The response directly answers the question as instructed.

Responses with a total score of **2** are considered correct for accuracy calculation.

To classify responses as either knowledge-based or vision-based, annotators grade the semantic closeness between the current response and the reference responses on a scale from 0 to 2:



(a) Cows in a field chopping carrots.
Labels: A(0), Q(1)

(b) The teacher in a classroom paddling the boat.
Labels: A(1), Q(0)

(c) The farmer pouring milk in an ambulance.
Labels: A(0), Q(1)

(d) The pharmacist serving customers on a tree.
Labels: A(1), Q(0)

Figure 9: Examples of rejected images and their labels.**A**: Alignment with text prompt, **Q**: Quality of the image.

- **0**: The answers are entirely unrelated.
- **1**: The answers share similar concepts but are not identical or synonymous.
- **2**: The answers are identical or synonymous.

The candidate response is assigned to the category with a higher score (e.g., Vision or Knowledge). For example, a response is classified under the Vision category if it receives a vision score of 2 and a knowledge score of 1. Responses with both scores equal to 0 are categorized as "Other".

B.5 Human Performance

In Table 8, we present the results of human annotators' performance on our CONFLICTVIS benchmark. The reported results are averaged across three annotators. The results indicate that human can achieve a higher accuracy than MLLMs.

Question Type	Accuracy
Yes-No (YN)	93.0%
Multiple-Choice (MC)	99.7%
Open-Ended (OE)	91.7%

Table 8: Human Performance on CONFLICTVIS .

C Experiment Details

C.1 MLLM Generation Configurations

For both open-source and closed-source models, we use the default generation configurations (e.g., temperature) to align with the typical usage.

Question prompts for different MLLMs are crafted based on the prompt templates used in each model's original evaluation. Specifically, we use the synthesized question text from the framework as the base and append a corresponding postfix tailored to each question type and model family. In practice, for the LLaVA series, GPT-4o, and Claude-3.5-Sonnet, the prompts are as follows:

Question Type	Prompt Format
Multiple-Choice	`{question text}` `Answer with the option's letter from the given choices directly.`
Yes-No	`{question text}` `Answer Yes or No directly.`
Open-Ended	`{question text}` `Answer no more than 5 words.`

For the BLIP-2 model family, the prompt postfixes are "`Answer:`" for MC and YN questions and "`Short Answer:`" for OE questions. For Qwen-VL, the postfixes include "`Choice Answer:`" for MC questions, while both YN and OE questions use "`Answer:`". By contrast, Qwen-VL-Chat uses "`Answer with the option's letter from the given choices directly.`" for MC questions, "`Answer Yes or No directly.`" for YN questions, and "`Answer with a single word or phrase.`" for OE questions.

C.2 Uncertainty Measurement

Unlike the traditional VQA benchmarks, where visual information typically aligns with the model knowledge, CONFLICTVIS presents conflicts between the two sources of knowledge, making it a more challenging benchmark for MLLMs. To empirically validate our claim, we conduct a comparative analysis of model uncertainty between our CONFLICTVIS and the conventional VQA benchmark (Antol et al., 2015). A higher degree of uncertainty in model predictions on a particular benchmark indicates that the benchmark poses a greater challenge for the model to navigate. Figure 10 shows the results, where a higher entropy value denotes more uncertainty. It is evident that the model's predictions are more uncertain on CONFLICTVIS than the standard VQA benchmark, highlighting the increased challenge posed by CONFLICTVIS.

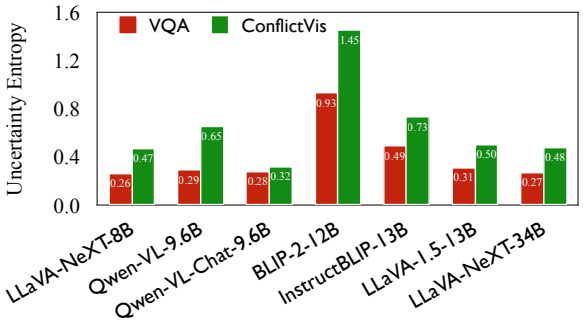


Figure 10: Model uncertainty on traditional VQA dataset and on CONFLICTVIS benchmark.

To measure the uncertainty in model's responses, we rely on the entropy. The average entropy H_{avg} for a language model's response y is defined as:

$$H_{\text{avg}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{v\in V} p_n(v)\log p_n(v)$$

where N is the number of tokens in the response y, V is the vocabulary and $p_n(v)$ is the softmax probability of token v in the vocabulary at position n in the response sequence.

To ensure a fair comparison, we first download the original VQA v1 dataset and extract an equal number of instances (374) for each question type (e.g., Multiple-Choice, Yes-No, and Open-Ended). Using the crafted subsets, we conduct controlled experiments under identical generation configurations to evaluate MLLMs on VQA v1 and our benchmark. The final uncertainty for each model is computed as the average entropy across all instances.

C.3 Extended Analysis on No-type Questions

To mitigate the potential dataset bias favoring "Yes" answers, we extend the Yes-No questions in our benchmark by constructing additional "No"-type questions and conduct further experiments. Specifically, for each original Yes-No query, we create a negative counterpart by replacing the counter-commonsense target with the most frequently co-occurring component from the corresponding option list. For example, given an image of a waitress signing a bill in the kitchen and the original

question “Is the waitress in the kitchen signing a bill?” (expected answer: “Yes”), we generate the counterpart question “Is the waitress in the kitchen washing dishes?” (expected answer: “No”).

We then reevaluate the MLLMs on these negative queries, with results summarized in Table 9. To compute the Memorization Ratio (MR), we follow the procedures outlined in Section 4.2: we first obtain the model's responses without the image input and identify those accepted by the model, then evaluate these cases using the No-type question inputs, and finally compute MR according to Eq. 3. The results indicate that these No-type questions still trigger a substantial number of failures across models, consistent with our original findings.

Model	Acc. (↑)	MR (↓)
LLaVA-NeXT-8B	81.3	21.8
Qwen-VL-9.6B	93.0	11.2
Qwen-VL-Chat-9.6B	71.7	31.5
BLIP-2-12B	74.9	18.8
InstructBLIP-13B	85.0	19.0
LLaVA-1.5-13B	77.5	26.5
LLaVA-NeXT-34B	82.1	20.8
Claude-3.5-Sonnet	98.9	1.3
GPT-4o	95.7	4.6

Table 9: MLLM accuracy (higher is better) and Memorization Ratio (MR) (lower is better) on No-type Yes-No questions.

C.4 Detailed Results on Conflict Targets

As outlined in Section 4.3, here we provide detailed experimental results on counter-commonsense actions and places, as presented in Figure 11.

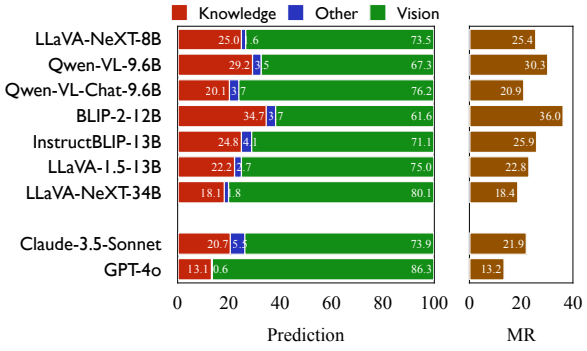
C.5 Improvement Methods Configurations

We present the detailed configurations for the improvement methods in Table 10.

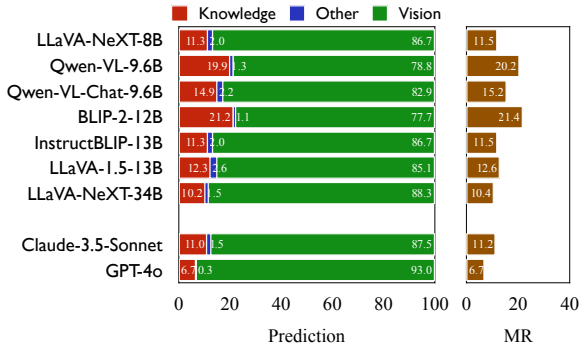
D Case Study

D.1 Extended Error Analysis

In Section 4.4, we identified a common failure pattern in which image tokens are underutilized during answer generation. In Figure 12, we present four additional examples from `LLaVA-1.5-13B`, spanning both counter-commonsense action and place categories. The first two are failure cases, while the bottom two are correct cases. In the failure cases, image tokens are significantly underutilized: in the



(a) Counter-commonsense **Action**



(b) Counter-commonsense **Place**

Figure 11: MLLM response distributions on two distinct categories of conflict targets.

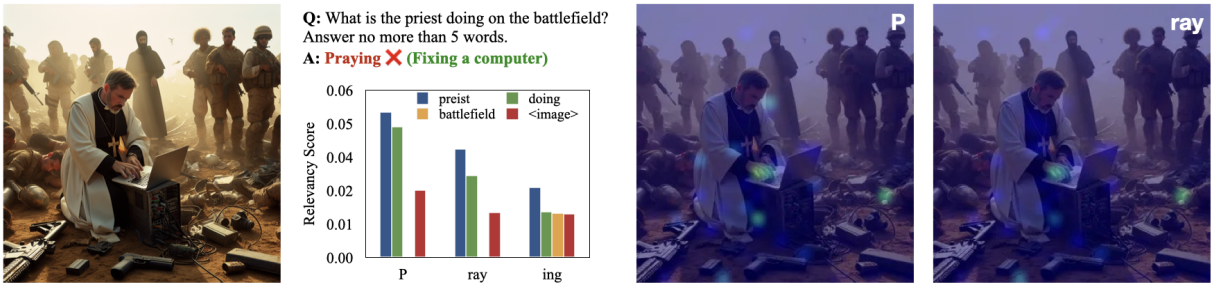
bar plot, the <image> token, representing the image token with the highest relevancy score, is often less related to the output tokens than other textual tokens from the question description. The relevancy heatmaps also display only a few weakly attended regions. In contrast, for the correct cases, the <image> token typically has the highest relevancy scores for visual-dependent output tokens, and the heatmaps exhibit more numerous and strongly attended areas. Together, these examples highlight that the underutilization of visual information is a key factor contributing to model failures under vision-knowledge conflict. To address this issue, a stronger alignment between the image and the generated output is needed, guiding the development of more targeted and vision-aware improvement methods, as explored in the subsequent section.

D.2 Failure Cases

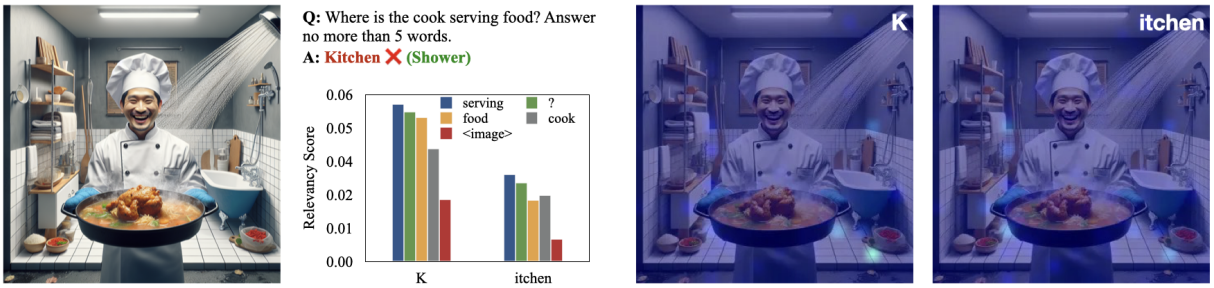
In this section, we present representative failure cases from our ConflictVis benchmark, spanning various models and question types. Specifically, Figures 13, 14, and 15 showcase failure cases for Yes-No, Multiple-Choice, and Open-Ended questions, respectively, while Figures 16 and 17 illustrate exacerbated failure cases when employing Chain-of-Thought prompting.

Method	Configuration	Method	Configuration
VCD	$\alpha = 1.0, \beta = 0.1,$ noise_step $= 500$	VR – CoT	prompt template: `{original prompt} Let's think step by step.`
PAI	$\alpha = 0.5, \gamma = 1.1,$ apply_layers $= (2, 32)$	VR – SFT	dataset: `LLaVA-CoT-100k`, batch_size $= 256$, learning_rate $= 1e-5$, lr_scheduler $= linear$, epochs $= 3$, max_seq_length $= 1024$

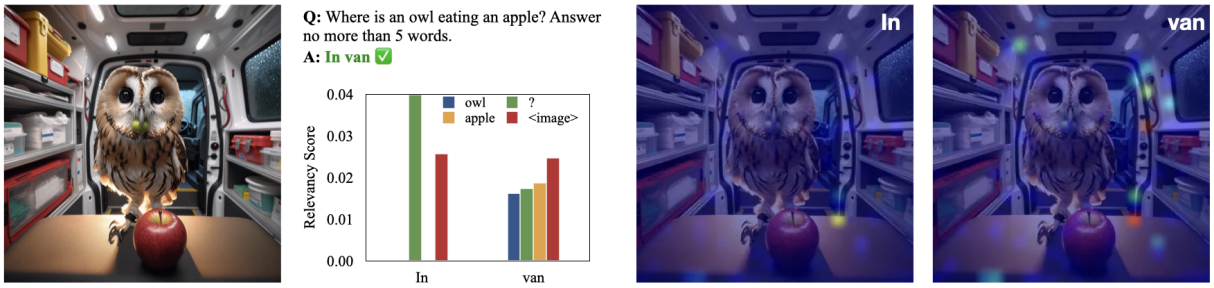
Table 10: Configurations for the improvement methods.



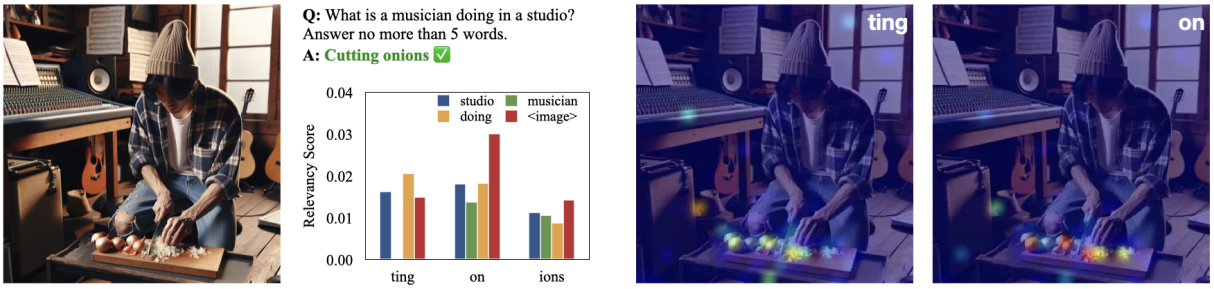
(a) Failure case: The priest fixing a computer on the battlefield.



(b) Failure case: The cook serving food in a shower.



(c) Correct case: An own eating an apple in an ambulance(van).



(d) Correct case: A musician chopping(cutting) some onions in a studio.

Figure 12: Case study of input-output relevancy. Image tokens are underutilized in failure cases (a)(b), but effectively used in correct cases (c)(d).

Failure Case - Yes-No

Triplet: (a carpenter, in a factory, chopping some onions)
Question: Is a carpenter in a factory chopping some onions?

Claude-3.5-Sonnet: No.

Figure 13: Failure case on yes-no question.

Failure Case - Multiple Choice

Triplet: (the janitor, in an office, signing a bill)
Question: What is a janitor doing in an office?
(A) sweeping the floor (B) eating the roadrunner
(C) honoring a knight (D) signing a bill

GPT-4o: (A) sweeping the floor

Figure 14: Failure case on multiple-choice question.

Failure Case - Subjective

Triplet: (a cook, serving food, in a shower)
Question: Where is a cook serving food?

LLaVA-NeXT-34B: Kitchen

Figure 15: Failure case on open-ended question.

CoT Failure Case - Knowledge Over-reliance

Triplet: (the fireman, powdering the baby, by the fire)
Question: What is the fireman doing by the fire?

Original (zero-shot):

Qwen-VL-Chat: The fireman is standing in front of the fire and holding a baby.

CoT:

Qwen-VL-Chat: ***Let's think step by step.*** We can see a fireman in the picture. We know that firemen put out fires, so the fire in the picture must be put out by the fireman. Therefore, the answer is "put out the fire".

Figure 16: CoT failure mode 1: The MLLM reasons based on commonsense knowledge to derive an incorrect answer.

CoT Failure Case - Denial

Triplet: (a student, taking an exam, in the fridge)
Question: Where is the student taking an exam?

Original (zero-shot):

LLaVA-NeXT-8B: Refrigerator.

CoT:

LLaVA-NeXT-8B: ***Let's think step by step.*** The student is in a kitchen, sitting inside an open refrigerator. The refrigerator is not a typical place for an exam, so let's consider the context. The student is likely taking a break from studying or perhaps trying to find a quiet and cool place to focus. However, the refrigerator is not a suitable environment for an exam, as it is cold and not designed for such activities. Therefore, the answer is "Not possible."

Figure 17: CoT failure mode 2: The MLLM incorporates commonsense knowledge and asserts that the answer is impossible.