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

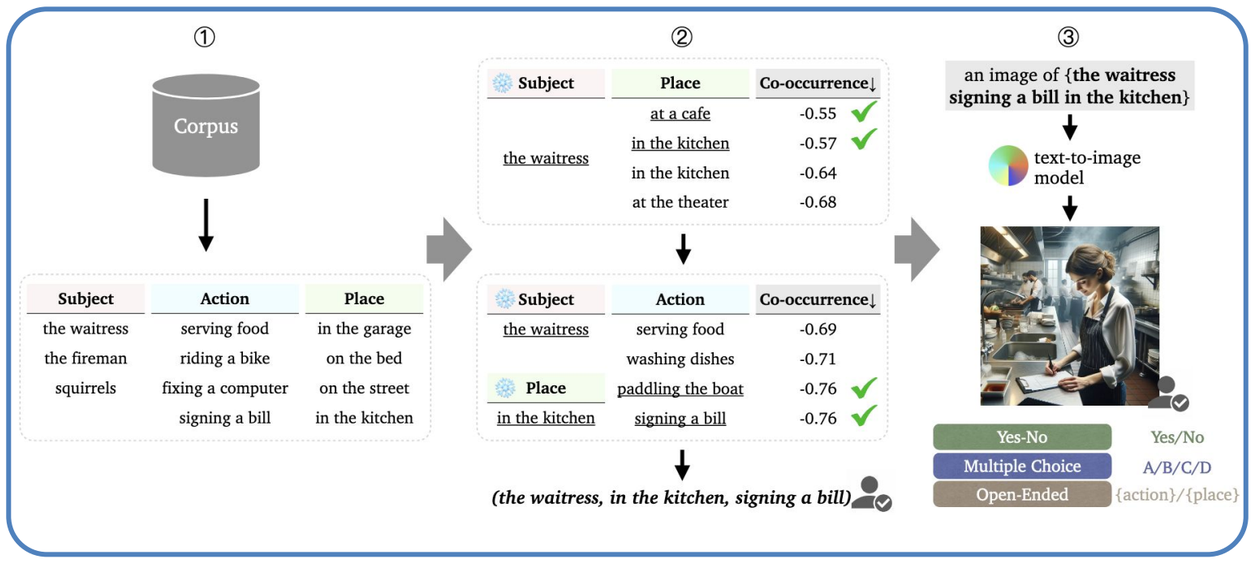
Xiaoyuan Liu[1,2], Wenxuan Wang[3], Youliang Yuan[1,2], Jen-tse Huang[4], Qiuzhi Liu[2], Pinjia He[1†], Zhaopeng Tu[2]

[1]The Chinese University of Hong Kong, Shenzhen [2]Tencent [3]Renmin University of China [4]Johns Hopkins University

Motivation

The problem of knowledge conflicts in LLMs remains a significant challenge for MLLMs. Introducing the vision modality creates a novel form of discrepancy we term **vision-knowledge conflict**, where **the visual inputs contradicts the model's pre-trained parametric knowledge**. Given the growing demand for accurate, trustworthy multimodal systems, we believe it is imperative to study such cases to better understand and improve the MLLMs.

Methodology

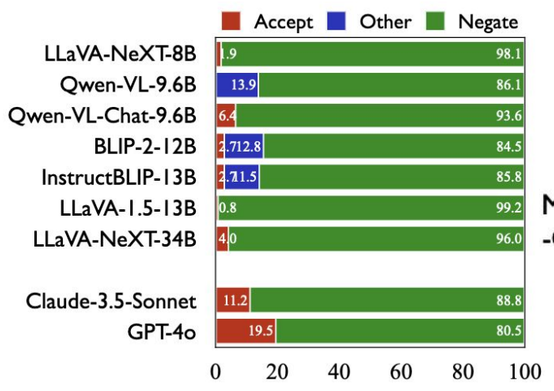


Automated Framework for ConflictVIS Benchmark Construction

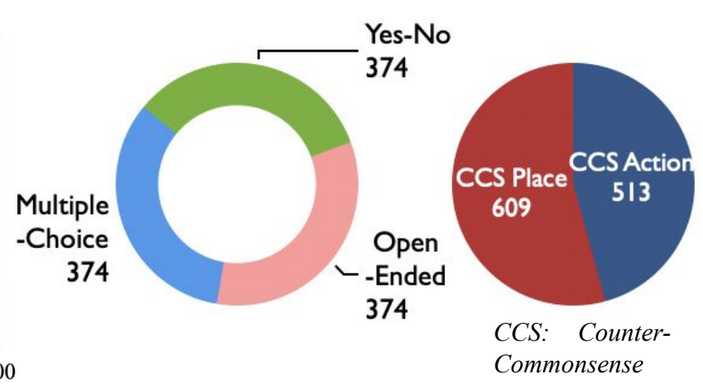
- Extract Knowledge Components
  Leveraging the syntactic labels including dependency (DEP), Part-of-Speech (POS), and Named Entity (NE) annotated, we use pre-defined rules extract the *Subject*, *Action*, and *Place* phrases from the input corpus as the atomic knowledge components.
- Construct Counter-commonsense Query
  We aim to construct scenes with a single anomalous component (i.e., conflict target) that seldom co-occurs with the others (i.e., context pair). To this end, we calculate the *Normalized Pointwise Mutual Information* (NPMI) for different combinations and select the context pair with a high NPMI score and a target with a low NPMI score given the selected context to form the raw input triplets (S, A, P).
- Generate Multimodal Inputs
  We apply predefined prompt template to generate synthesized images using text-to-image models, and define linguistic rules to construct different types of question: *Yes-No*, *Multiple-Choice*, *Open-Ended*, and different conflicting targets: *Action* and *Place*.

Results

Sanity Test Results

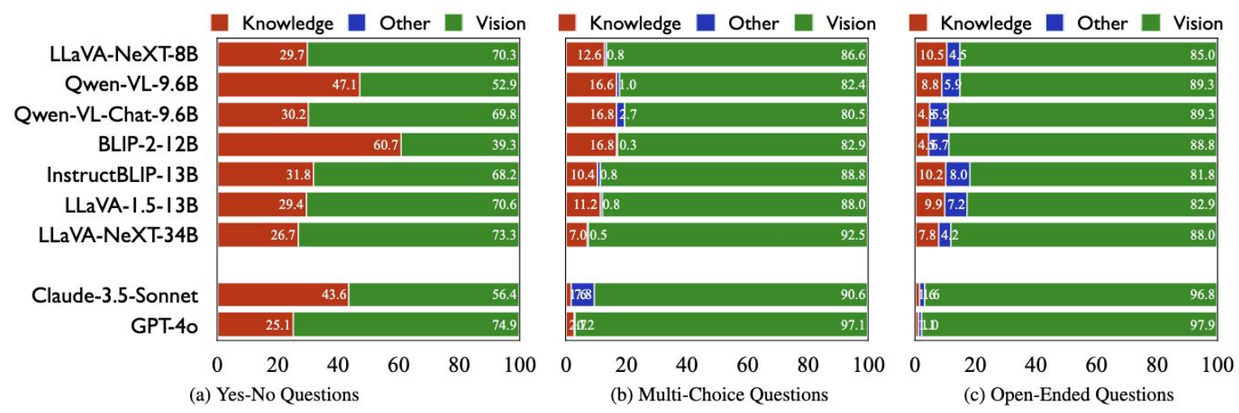


Data Distribution



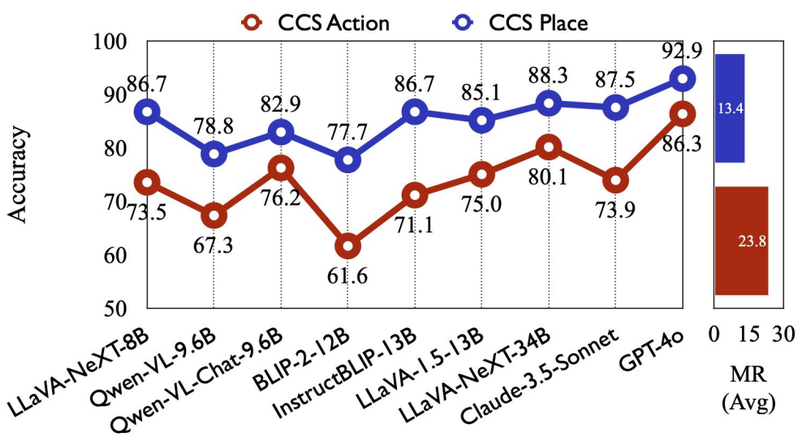
CCS: Counter-Commonsense

Benchmark Results (Question Type)



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

Benchmark Results (Conflict Type)

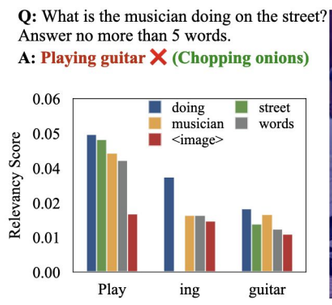


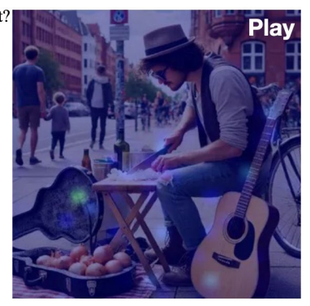
Knowledge: answer based on the model's parametric knowledge
Vision: answer based on the visual input
Other: Not grounded in either

$$MR = \frac{P_K}{P_K + P_V}$$

(Memorization Ratio)

Analysis & Improvements





Model	Yes-No	Multi-Choice	Open-Ended	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**

VCD: Visual Contrastive Decoding [CVPR'24]; **PAI:** Pay more Attention to Image [ECCV'24]; **VR**: Vision-Centric Reasoning [Arxiv'24]; **FoV (Ours)**: Focus-on-Vision Prompting.

- ConflictVIS presents **valid** vision-knowledge conflicts for MLLMs.
- Around **20%** of responses rely on prior knowledge over visual input.
- MLLMs are especially prone to errors in **Yes-No** questions and **Counter-Commonsense Action** inputs.

- Upon conflicts, MLLMs tend to **assign more weight to textual input** than to visual context, often leading to incorrect answers.
- Existing hallucination mitigation methods and our FoV prompting **improve MLLM accuracy but cannot fully resolve the issue.**

Dataset Available at: xiaoyuanliu/conflict_vis