

Benchmarking Multimodal Knowledge Conflict for Large Multimodal Models

Yifan Jia^{1,*}, Yuntao Du^{1,2,*†}, Kailin Jiang^{3,*}, Yuyang Liang¹, Qihan Ren⁴, Yi Xin², Rui Yang¹, Fenze Feng¹, MingCai Chen⁵, Hengyang Lu⁶, Haozhe Wang⁷, Qian Li¹, Xiaoye Qu⁸, Dongrui Liu⁸

¹ Joint SDU-NTU Centre for Artificial Intelligence Research & School of Software, Shandong University

² State Key Lab. for Novel Software Technology, Nanjing University, P.R. China

³ University of Science and Technology of China

⁴ Shanghai Jiaotong University

⁵ Nanjing University of Posts and Telecommunications

⁶ School of Artificial Intelligence and Computer Science, Jiangnan University

⁷ The Hong Kong University of Science and Technology

⁸ Shanghai AI Laboratory

Abstract

Large Multimodal Models (LMMs) face notable challenges when encountering multimodal knowledge conflicts, particularly under retrieval-augmented generation (RAG) frameworks, where the contextual information from external sources may contradict the model’s internal parametric knowledge, leading to unreliable outputs. However, existing benchmarks fail to reflect such realistic conflict scenarios. Most focus solely on intra-memory conflicts, while context-memory and inter-context conflicts remain largely unaddressed. Furthermore, commonly used factual knowledge-based evaluations are often overlooked, and existing datasets lack a thorough investigation into conflict detection capabilities. To bridge this gap, we propose **MMKC-Bench**, a benchmark designed to evaluate factual knowledge conflicts in both context-memory and inter-context scenarios. MMKC-Bench encompasses four types of multimodal knowledge conflicts and includes 1,593 knowledge instances and 2,701 images across 23 broad types, collected through automated pipelines with human verification. We evaluate four representative series of LMMs on both model behavior analysis and conflict detection tasks. Our findings show that while current LMMs are capable of recognizing knowledge conflicts, they tend to favor internal parametric knowledge over external evidence. We hope MMKC-Bench will foster further research in multimodal knowledge conflict and enhance the development of multimodal RAG systems.

Code — <https://github.com/MLLMKCBENCH/MLLMKC>

Introduction

The rapid advancement of large multimodal models (LMMs) and large language models (LLMs) has led to remarkable performance across a wide range of multimodal understanding, generation, and reasoning tasks (Bai et al. 2025; Chen et al. 2024b; Liu et al. 2023; Cui et al. 2024; Su et al. 2025a). Despite their impressive capabilities, static

LMMs and LLMs often suffer from limitations such as outdated or incorrect knowledge and hallucinations (Du et al. 2025a; Jiang et al. 2025a,b). To address these issues, retrieval-augmented generation (RAG) techniques have been introduced (Fan et al. 2024a; Mei et al. 2025), which enhance model outputs by incorporating up-to-date information from external sources. However, this paradigm introduces the challenge of knowledge conflict, where the retrieved contextual knowledge may contradict the model’s internal (parametric) knowledge (Xu et al. 2024). Recent studies have demonstrated that such conflicts can undermine the trustworthiness and reliability of model predictions (Wang et al. 2024), highlighting the need for a deeper understanding of model behavior under conflicting knowledge scenarios. To facilitate this, several benchmark datasets have been developed to study knowledge conflict both in textual contexts (Hou et al. 2024; Su et al. 2024; Wang et al. 2024) and in multimodal domains (Liu et al. 2024; Shao et al. 2024; Zhu et al. 2024).

As summarized in (Xu et al. 2024), knowledge conflicts in LLMs fall into three types: intra-memory, context-memory, and inter-context conflicts. Intra-memory conflict stems from contradictions within parametric knowledge caused by noisy or inconsistent pretraining data. Context-memory and inter-context conflicts arise during inference, with the former referring to contradictions between external context (e.g., prompts or retrieved documents) and internal knowledge, and the latter denoting inconsistencies among contextual sources, issues that are prominent in RAG settings. Existing multimodal conflict datasets target commonsense-based context-memory conflicts (Liu et al. 2024), cognition-perception intra-memory conflicts (Shao et al. 2024), and cross-modality conflicts via textual/multimodal question framing (Zhu et al. 2024). However, they fail to reflect real-world RAG retrieval scenarios because they (1) emphasize intra-memory over context-memory and inter-context conflicts, (2) lack external factual knowledge requirements (e.g., object recognition), and (3) focus on describing model behavior rather than evaluating conflict detection.

*Equal contribution.

† Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

	 Entity Recognition Conflict	 Entity Knowledge Conflict	 Visual Semantic Conflict
Original Knowledge	 This image is the Sydney Opera House . Sydney Opera House is an Australian landmark building	 The person in the image is Musk. Musk was born in 1971 . Musk is an American. and Musk is a businessman.	 This is a gesture of OK .
Conflict Knowledge 1	 This image is the White house . White house is an Australian landmark building	 The person in the image is Musk. Musk was born in 1974 . Musk is an American. and Musk is a businessman.	 This is a gesture of rejection .
Conflict Knowledge 2	 This image is the The Eiffel Tower . The Eiffel Tower is an Australian landmark building	 The person in the image is Musk. Musk was born in 1979 . Musk is an American, and Musk is a businessman.	 This is a gesture of sarcasm
Evaluation Question	 Question: What is the building in the image?	 Question: In Which year was the person in the image born?	 Question: What is the gesture in the image?

Figure 1: Three types of multimodal knowledge conflict in MMKC-Bench. It is noted that the original knowledge is shown to help understand what the conflict is, and is not contained in the dataset.

To address these gaps, we propose MMKC-Bench, a **multimodal knowledge conflict benchmark aimed at evaluating real-world knowledge conflict under both context-memory and inter-context scenarios**. As illustrated in Fig. 1, MMKC-Bench focuses on three representative types of multimodal conflicts: entity recognition conflict, entity knowledge conflict, and visual semantic conflict. The entity recognition and visual semantic conflicts target inconsistencies in entity identification and complex visual understanding, while the entity knowledge conflicts emphasize factual inconsistencies related to specific attributes or quantitative data. In addition to analyzing model behavior in the presence of conflict, MMKC-Bench also investigates whether models can detect and perceive conflicts at both coarse-grained (given whole evidence) and fine-grained (given a subset of the evidence) levels.

To construct the benchmark, we first curate original multimodal knowledge from a variety of sources, including Wikipedia, Google Images, and existing datasets. We then employ large language models (LLMs) to generate conflicting knowledge through counterfactual editing, which involves modifying the entity name, semantic content, or entity-related factual information. Based on the constructed knowledge pairs, we use LLMs to generate both multiple-choice and open-ended evaluation questions, along with candidate answers. All generated questions, answers, and the preceding data construction steps undergo rigorous human verification, with samples being filtered or revised as needed to ensure quality and accuracy. The final MMKC-Bench dataset comprises 1,593 knowledge instances and 2,701 images, spanning 23 broad knowledge categories.

We conduct experiments on ten representative LLMs from four prominent model families known for their

strong performance in multi-image reasoning: Qwen2.5-VL (Bai et al. 2025), InternVL3 (Zhu et al. 2025), GPT-4o mini (Hurst et al. 2024), and Gemini-2.5-pro. Our evaluation covers both model behavior analysis and conflict detection tasks. The experimental findings reveal several key insights: (1) LLMs tend to rely more heavily on internal (parametric) knowledge than on external evidence, a behavior that contrasts with previously observed trends in LLMs (Su et al. 2024; Hou et al. 2024); (2) LLMs are more sensitive to knowledge-level conflicts (e.g., entity knowledge) than to recognition-level conflicts (e.g., entity identification); (3) Larger models show a stronger promoting effect across all conflict types; (4) LLMs are capable of accurately identifying the presence of conflict in both coarse-grained and fine-grained scenarios.

The contribution of this work is summarized as follows,

- We propose MMKC-Bench, a multimodal knowledge conflict benchmark focusing on factual knowledge conflict under both context-memory and inter-context scenarios.
- We propose a novel pipeline to construct the benchmark that collects original knowledge, generates conflict knowledge, and produces evaluation question.
- Extensive experiments with various models under both context-memory and inter-context for behavior understanding and conflict detection are conducted, revealing several characteristics of existing LLMs.

Related Work

Large Multimodal Model

Modern LLMs (Su et al. 2025b) integrate visual and textual information through three core components: a language en-

coder (e.g., LLaMA (Grattafiori et al. 2024; Touvron et al. 2023), Qwen (Yang et al. 2024)), a vision encoder (e.g., ViT (Dosovitskiy et al. 2020)), and cross-modality alignment modules (Caffagni et al. 2024). This architecture has spawned advanced models like Qwen2.5-VL (Bai et al. 2025), InternVL2.5 (Chen et al. 2024a), and LLaVA (Liu et al. 2023), each employing distinct training strategies for cross-modal alignment. These developments and following agent-based method (Fan et al. 2024b; Gao et al. 2024) have significantly improved performance across multimodal tasks (Kil et al. 2024; Huang and Zhang 2024), demonstrating rapid progress in LMM design (Huang et al. 2025).

Knowledge Conflict in LLMs

Knowledge conflict arises from discrepancies between contextual inputs and a model’s parametric knowledge (Chen, Zhang, and Choi 2022; Xie et al. 2023), categorized into context-memory, inter-context, and intra-memory conflicts. While RAG research focuses mainly on context-memory conflicts, existing datasets are synthetic, created via entity substitution (Longpre et al. 2021; Wang et al. 2024) or LLM-generated contradictions (Ying et al. 2024). Current benchmarks like ConflictBank (Su et al. 2024) (binary mismatches) and WikiConflict (Hou et al. 2024) (Wikipedia contradictions) attempt to capture real-world complexity. Related work also explores conflict detection (Wang et al. 2024; Li, Raheja, and Kumar 2024) and mitigation (Gekhman et al. 2023; Chuang et al. 2024; Cheung and Lam 2023) for improved LLM reliability.

Knowledge Conflict in LMMs

Recent studies have explored LMM knowledge conflicts from various perspectives, including: commonsense conflicts in context-memory (Liu et al. 2024), cognition-perception conflicts via multi-capability evaluation (Shao et al. 2024), and cross-modality conflicts through textual-/multimodal comparison (Zhu et al. 2024). However, these benchmarks inadequately represent real-world RAG scenarios, as they: (1) focus primarily on intra-memory conflicts, neglecting context-memory and inter-context cases; (2) underrepresent critical external knowledge conflicts (e.g., visual object identification); and (3) prioritize behavior observation over conflict acknowledgment evaluation.

Problem Definition

Original and Conflict Knowledge Representation

MMKC-Bench focuses on factual knowledge based knowledge conflict and encompasses three types: entity recognition conflict, entity knowledge conflict, and visual semantic conflict. For these types, each original piece of original knowledge is represented in a unified format $k = (i, d)$, where i denotes an image of the entity or semantic action, and d is the corresponding textual description. To construct conflicting instances, the original knowledge is modified to form k_c . It takes the form $k_c = (i_c, d_c)$, where i_c is another image of the same entity or action, and d_c is a conflicting description. The examples of each type are shown in Fig 1.

Multimodal Knowledge Conflict Types

To comprehensively reflect real-world scenarios, MMKC-Bench includes three types of multimodal knowledge conflicts: entity recognition conflict, entity knowledge conflict, and visual semantic conflict.

Entity Recognition Conflict simulates cognitive inconsistencies where different sources identify the same entity differently. This is achieved by keeping the entity image unchanged while replacing the entity name in the description with that of another entity of the same type. For example, as shown in Fig 1, describing the Sydney Opera House image using “Eiffel Tower” or “White House” as the entity name.

Visual Semantic Conflict addresses inconsistencies in interpreting complex visual semantics, such as gestures, body actions, or symbolic cues. Unlike Entity Recognition Conflict, Visual Semantic Conflict focuses more on the semantic information within the image rather than the identity of the subject. Here, the semantic meaning associated with an action is replaced with that of another action of the same type. For instance, as shown in Fig 1, the meaning of an “OK” gesture is altered to convey “rejection” or “sarcasm”, without requiring attention to the identity of the person performing the gesture.

Entity Knowledge Conflict centers on factual discrepancies surrounding entity attributes like nationality, occupation, or birth year. This is simulated by substituting the tail entity in a factual triple with another entity of the same type. For example, as shown in Fig 1, altering Elon Musk’s birth year from the correct value to 1974 or 1979.

MMKC-Bench

In this section, we present the pipeline for constructing MMKC-Bench, a QA-based benchmark comprising 1,593 instances that encompass three types of multimodal knowledge conflicts, as illustrated in Fig 2.

Original Knowledge Collection

We begin with original knowledge collection by first listing candidate entity types or visual semantics. Then, we collect the corresponding images and descriptions.

For entity recognition and entity knowledge conflicts, we manually define multiple candidate visual entity types (e.g., person, building). For each type, we use an LLM to generate a list of the most prominent entities (e.g., Messi under the “person” category). Once the entity list is obtained, we crawl their images from Google and retrieve entity descriptions from Wikipedia summary dumps¹, which are then summarized by an LLM to retain essential information. For visual semantic conflicts, we follow prior work on knowledge editing with visual semantic modifications (Du et al. 2025b), focusing on four categories of visual semantic knowledge: everyday gestures, human body actions, human emotions, and symbol identification. Since these types are already included in MMKC-Bench, we directly obtain the original visual semantic knowledge, consisting of paired images and their meanings, from the MMKC-Bench dataset, rather than collecting them manually.

¹<https://dumps.wikimedia.org>

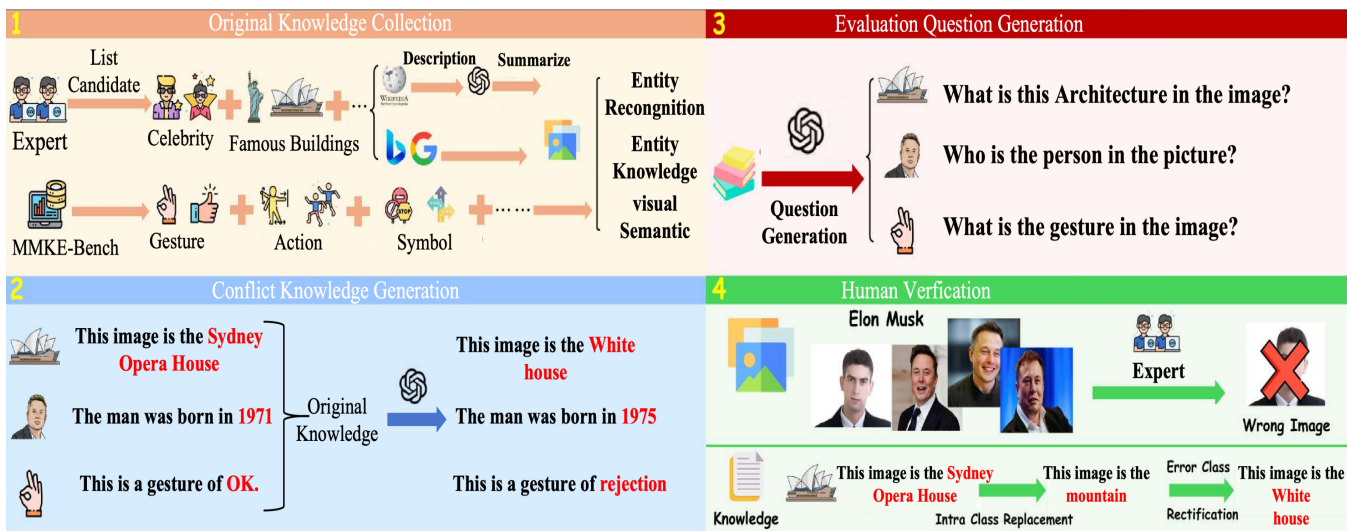


Figure 2: The construction pipeline of MMKC-Bench.

Conflict Knowledge Generation

Considering the multimodal nature of LMMs, we generate conflict knowledge by deliberately introducing misalignments across modalities, with the assistance of large language models (LLMs).

For entity recognition, visual semantics, and entity knowledge conflicts, we retain the original image while modifying the textual component. Specifically, we replace the entity name, the meaning of the action, or the tail entity in a factual triple with another instance of the same type. For example, as illustrated in Fig 1, the name “Empire State Building” is replaced with “Eiffel Tower” or “White House”, and Elon Musk’s birth year is altered to 1974 or 1979.

To simulate both context-memory conflicts and inter-context conflicts, we generate two conflicting versions for each piece of original knowledge following the above procedures. In the context-memory setting, one conflicting version is randomly selected as internal evidence. In the inter-context setting, both conflicting versions are provided as internal evidence.

Evaluation Question Generation

We adopt a visual question answering (VQA) format to construct evaluation questions and answers, leveraging LLMs for automatic generation. The specific prompts used are provided in the appendix (Jia et al. 2025).

We consider two types of questions: multiple-choice questions (MCQs) and open-ended questions (OQAs). For entity recognition and semantic recognition conflicts, the questions focus on identifying the entity name or interpreting the semantic meaning depicted in the image. For entity knowledge conflicts, the questions target fine-grained factual knowledge, such as querying a person’s occupation or age. In MCQs, each question includes four answer options: one answer within the models’ internal knowledge, two answers from the conflicting knowledge variants, and one unrelated distractor option.

Human Verification and Benchmark Statistics

During construction, we conducted manual collection, review, and filtering to ensure data quality. In the original knowledge collection stage, all images associated with each entity, semantic action, and piece of knowledge were manually reviewed to ensure their accuracy and relevance. Furthermore, following counterfactual editing and question generation, we performed additional manual verification, filtering out inappropriate samples, revising ambiguous or ill-formed questions, and correcting incorrect answers.

The statistical information of MMKC-Bench is shown in Tab 1. MMKC-Bench encompasses three types of conflict knowledge, containing 1,593 pieces of knowledge and 2,701 images. These knowledge spans 23 fine-grained types, highlighting the diversity of MMKC-Bench. Note that here, a single piece of knowledge corresponds to multiple images and an image is used for multiple knowledge, simulating the input scenarios of real-world RAG.

	#Types	#Instances	#Images
Visual Entity Conflict	13	757	2,271
Entity Knowledge Conflict	6	669	669
Visual Semantic Conflict	4	147	441

Table 1: The statistics of MMKC-Bench.

Experiment

Setup

Models In multimodal knowledge conflict scenarios, the model input consists of multiple interleaved images and texts. Therefore, we select LMMs that perform well in multi-image understanding. Specifically, we conduct a comprehensive evaluation on 10 LMMs across 4 model series, with sizes ranging from 3B to 72B. The selected models include:

Qwen2.5-VL (3B, 7B, 32B, 72B) (Bai et al. 2025), **InternVL3** (8B, 14B, 38B, 78B) (Zhu et al. 2025), **GPT-4o mini** (Hurst et al. 2024), and **Gemini-2.5-Pro** (Comanici et al. 2025).

Settings We consider two conflict-related tasks: conflict behavior analysis and conflict detection. The former investigates how models behave under conflicting scenarios, while the latter evaluates whether models can correctly detect the presence of conflict.

For conflict behavior analysis, we consider two types of conflict scenarios: context-memory conflict and inter-context conflict. In context-memory conflict, one piece of conflicting external evidence, composed of an image and associated text, is provided as an in-context example. In inter-context conflict, two conflicting pieces of evidence about the same knowledge are provided as in-context examples. The model is then required to answer an evaluation question based on this context.

For conflict detection, we explore both coarse-grained and fine-grained conflict detection. In the coarse setting, a full piece of evidence (either conflicting or non-conflicting) is provided in context, and the model must determine whether a conflict exists by answering “yes” or “no”. Following previous work (Wang et al. 2024), the fine-grained setting involves providing only one sentence, which is the subset of full evidence, and the model must again judge whether a conflict is present.

Evaluation Metrics To understand the behavior of LMMs in conflicting scenarios, we first establish their baseline responses Q through unimpeded question answering, treating these answers as reflections of their internal parametric knowledge. When exposed to conflicting external evidence, the model generates revised responses Q' , which are categorized into three classes: (1) **OA**: answers Q' consistent with the original baseline Q , (2) **CA**: answers aligned with the external conflicting evidence, or (3) **IA**: irrelevant answers inconsistent with both. We quantify this behavior using three metrics: Original Answer Ratio(OAR), Counter Answer Ratio(CAR), and Irrelevant Answer Ratio(IAR), where $\text{OAR} + \text{CAR} + \text{IAR} = 1$. Higher OAR values indicate stronger adherence to internal knowledge, elevated CAR reflects greater susceptibility to external influence, and increased IAR suggests avoidance or uncertainty in conflicting contexts:

$$\text{OAR} = \frac{\text{Count}(\text{OA})}{N}, \text{CAR} = \frac{\text{Count}(\text{CA})}{N}, \text{IAR} = \frac{\text{Count}(\text{IA})}{N}, \quad (1)$$

Where N is sample number. For conflict detection, we treat this as a binary classification task. If a knowledge conflict exists, the model should output “yes”; otherwise, it should output “no”, in both coarse and fine-grained settings. We report the detection accuracy as the evaluation metric.

Model Behavior Analysis

The results under both context-memory and inter-context conflict scenarios, using multiple-choice and open-ended question formats, are presented in Table 2, Table 3, Table 4, Fig. 3, and Fig. 4. Based on these results, we draw the following observations:

1) LMMs are more receptive to internal knowledge

than to external evidence. As shown in Table 2 and Table 3, under context-memory conflicts, the average OAR exceeds CAR in all cases, indicating that LMMs tend to favor internal knowledge. Closed-source GPT-4o mini and Gemini2.5-Pro show consistent results with open-source models, suggesting that even advanced closed models are insensitive to external evidence. This differs from LLMs, which have shown high receptiveness to external knowledge (Su et al. 2024; Hou et al. 2024). One reason for this contrast is the difference in training data formats: LLMs are typically trained on long text contexts involving multiple information sources, while LMMs are mostly trained on isolated image-text pairs. This limits the exposure to multi-source contexts and reduces the ability to integrate external information at inference.

This finding is important for designing multimodal RAG systems, as it reveals that LMMs may not naturally leverage retrieved evidence and instead rely on parametric knowledge. Thus, improving LMMs’ ability to incorporate external information is important, which may require innovations in training paradigms and model architecture.

2) LMMs are more sensitive to knowledge-related conflicts and less sensitive to recognition-based conflicts.

We categorize the three conflict types into recognition-based (entity recognition, visual semantics) and knowledge-related (entity knowledge). LMMs exhibit lower OARs on knowledge-related conflicts than on recognition-based ones, showing higher sensitivity to factual inconsistencies; Although recognition conflicts often achieve the highest OARs, indicating reliance on internal knowledge for perception tasks, prior work (Shao et al. 2024) shows that perception and cognition are core LMM abilities. Recognition depends on visual-text alignment (perception), whereas knowledge-related tasks require cognitively driven factual reasoning. Since LMMs are primarily trained on perception-heavy tasks (e.g., VQA, grounding, captioning) with limited cognitively demanding data, they develop stronger perceptual than reasoning abilities. As a result, LMMs tend to rely on internal memory for recognition but may depend on external sources for knowledge-intensive tasks, highlighting the need for cognitively challenging training data to enhance reasoning.

3) When provided with more external evidence, LMMs exhibit greater alignment with external information, though the improvement remains limited. Compared to context-memory conflict scenarios, models generally achieve higher CARs under inter-context conflicts, suggesting a slight increase in reliance on external evidence. This is because, given more internal information, the model output would be affected more. However, the overall improvement is limited: the largest increase in CAR is 21% on average. These results reaffirm that LMMs predominantly rely on their internal parametric knowledge, even when presented with multiple external sources.

4) Larger models exhibit a stronger promoting effect across all conflict types. As illustrated in Fig.3 and Fig.4, the Overall Agreement Rate (OAR) generally increases with model size within the Qwen2.5-VL series. Specifically, the OAR improves progressively as the model scales from 3B

	Qwen2.5-VL-7B				InternVL3-8B				GPT-4o mini				Gemini-2.5-Pro			
	ER	EK	VS	Avg.	ER	EK	VS	Avg.	ER	EK	VS	Avg.	ER	EK	VS	Avg.
Context-Memory Conflict																
OAR	0.81	0.50	0.71	0.67	0.48	0.46	0.70	0.49	0.78	0.54	0.61	0.66	0.64	0.56	0.76	0.65
CAR	0.15	0.48	0.21	0.30	0.41	0.45	0.26	0.41	0.11	0.43	0.30	0.27	0.22	0.41	0.18	0.27
IAR	0.04	0.02	0.08	0.03	0.11	0.09	0.04	0.09	0.11	0.03	0.09	0.07	0.12	0.03	0.06	0.08
Inter-Context Conflict																
OAR	0.87	0.51	0.72	0.70	0.41	0.47	0.66	0.46	0.76	0.47	0.57	0.62	0.71	0.54	0.64	0.63
CAR	0.10	0.48	0.25	0.28	0.53	0.51	0.30	0.50	0.19	0.50	0.40	0.34	0.21	0.41	0.19	0.27
IAR	0.03	0.01	0.03	0.02	0.07	0.02	0.05	0.04	0.05	0.03	0.03	0.04	0.08	0.05	0.10	0.10

Table 2: Results of context-memory and inter-context conflicts on MMKC-Bench with multiple-choice question format.

	Qwen2.5-VL-7B				InternVL3-8B				GPT-4o mini				Gemini-2.5-Pro			
	ER	EK	VS	Avg.	ER	EK	VS	Avg.	ER	EK	VS	Avg.	ER	EK	VS	Avg.
Context-Memory Conflict																
OAR	0.66	0.26	0.13	0.44	0.60	0.27	0.27	0.43	0.76	0.36	0.45	0.56	0.68	0.46	0.32	0.49
CAR	0.20	0.62	0.48	0.40	0.19	0.53	0.15	0.33	0.02	0.47	0.05	0.21	0.21	0.35	0.48	0.35
IAR	0.14	0.10	0.39	0.15	0.22	0.20	0.58	0.24	0.22	0.17	0.50	0.22	0.11	0.19	0.20	0.16
Inter-Context Conflict																
OAR	0.65	0.14	0.05	0.38	0.46	0.15	0.18	0.30	0.82	0.37	0.37	0.58	0.73	0.36	0.43	0.51
CAR	0.21	0.76	0.77	0.50	0.40	0.72	0.52	0.54	0.06	0.47	0.07	0.24	0.14	0.49	0.13	0.25
IAR	0.14	0.10	0.18	0.12	0.14	0.14	0.30	0.15	0.12	0.16	0.56	0.18	0.13	0.15	0.44	0.24

Table 3: Results of context-memory and inter-context conflicts on MMKC-Bench with open question answering format.

	Qwen2.5-VL-7B						InternVL3-8B						GPT-4o mini						Gemini-2.5					
	PT	PL	PC	LT	LC	LO	PT	PL	PC	LT	LC	LO	PT	PL	PC	LT	LC	LO	PT	PL	PC	LT	LC	LO
Context-Memory Conflict																								
OAR	0.04	0.26	0.93	0.03	0.77	0.96	0.04	0.26	0.85	0.03	0.72	0.86	0.27	0.28	0.96	0.23	0.67	0.81	0.32	0.29	0.89	0.13	0.82	0.93
CAR	0.96	0.71	0.05	0.91	0.20	0.04	0.66	0.64	0.06	0.97	0.28	0.10	0.73	0.70	0.02	0.77	0.23	0.15	0.64	0.63	0.11	0.78	0.07	0.07
IAR	0.00	0.03	0.01	0.06	0.03	0.00	0.30	0.10	0.08	0.00	0.00	0.03	0.00	0.02	0.02	0.00	0.10	0.04	0.04	0.08	0.00	0.09	0.11	0.00
Inter-Context Conflict																								
OAR	0.03	0.35	0.97	0.02	0.74	0.93	0.01	0.22	0.87	0.04	0.80	0.86	0.24	0.13	0.83	0.22	0.66	0.74	0.18	0.23	0.78	0.11	0.83	0.84
CAR	0.94	0.62	0.02	0.97	0.24	0.07	0.94	0.77	0.11	0.94	0.20	0.12	0.76	0.85	0.10	0.78	0.29	0.24	0.76	0.64	0.17	0.79	0.08	0.13
IAR	0.00	0.02	0.01	0.00	0.02	0.00	0.05	0.02	0.02	0.01	0.00	0.01	0.00	0.02	0.08	0.00	0.05	0.01	0.06	0.13	0.05	0.10	0.09	0.03

Table 4: Results of context-memory and inter-context fine-grained Entity Knowledge conflicts on MMKC-Bench with multiple-choice question format. PT stands for character time knowledge, PL for character nationality knowledge, PC for character occupation knowledge. LT stands for brand time knowledge, LC for brand creator knowledge, and LO for brand product knowledge.

to 7B, 13B, and 70B, reflecting gains across entity recognition conflict, entity knowledge conflict, and visual semantic conflict. This trend suggests that larger models are more strongly influenced by their internal knowledge. This enhanced capability may stem from exposure to more extensive training data, enabling larger models to develop stronger mechanisms for resolving conflicts.

5) While performance differs between the two question formats, the overall trends remain consistent. Under the two question formats, the models exhibit different performance levels. For instance, in the open-ended question format, models tend to achieve higher IAR, suggesting that the open-ended nature of the task introduces greater variability

in the model outputs. Despite these differences in absolute performance, the overall trend across both formats remains consistent, demonstrating the robustness of the proposed benchmark across varying evaluation settings.

6) Multimodal large models are particularly sensitive to temporal knowledge conflicts and are easily misled by external knowledge. As shown in Table 4, they exhibit significant vulnerability in both person-time and location-time knowledge. On Qwen2.5-VL-7B, the CAR for person-time knowledge reaches 0.96 and 0.94 under the two settings, while for location-time knowledge, it even hits 0.91 and 0.97. This demonstrates that multimodal large models’ temporal knowledge is highly fragile and prone to being misled

	Coarse-Grained Detection									Fine-Grained Detection			
	ER			EK			VS			Avg.			
	Non-Conflict	Conflict	Avg.	Non-Conflict	Conflict	Avg.	Non-Conflict	Conflict	Avg.	Non-conflict	Conflict	Avg.	
Qwen2.5-VL-7B	0.92	0.87	0.89	0.89	0.51	0.70	0.67	0.89	0.78	0.79	0.76	0.65	0.71
InternVL3-8B	0.95	0.44	0.69	0.98	0.67	0.82	0.87	0.72	0.79	0.75	0.92	0.35	0.64
GPT-4o mini	0.73	0.88	0.80	0.66	0.76	0.71	0.63	0.82	0.73	0.76	0.82	0.61	0.72
Gemini-2.5	0.89	0.72	0.81	0.75	0.69	0.72	0.82	0.78	0.80	0.78	0.85	0.72	0.79

Table 5: Results of coarse-grained and fine-grained conflict detection on MMKC-Bench.

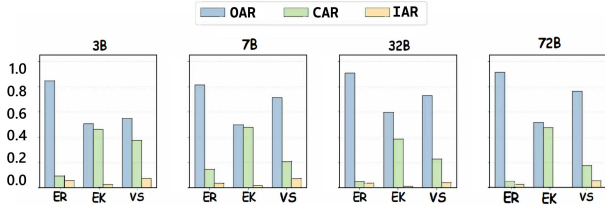


Figure 3: Results of Qwen2.5-VL with different model sizes under context-memory conflict with multi-choice question.

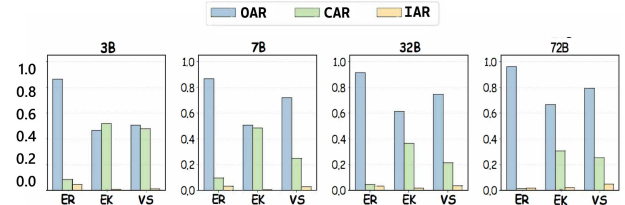


Figure 4: Results of Qwen2.5-VL with different model sizes under inter-context conflict with multi-choice question.

by conflicting external information, which undermines their safety and robustness.

Conflict Detection Analysis

The results of both coarse-grained and fine-grained conflict detection are shown in Table 5. Based on the results, we have the following findings:

1) **LMMs can effectively identify the presence of knowledge conflicts and generally perform better in recognizing conflicts under conflict scenarios than non-conflict scenarios.** As shown in Table 5, the average detection accuracy reaches 79%, 75%, 76%, and 78% for Qwen2.5-VL-7B, InternVL-8B, GPT-4o mini, and Gemini2.5-Pro, respectively, indicating that LMMs are capable of reliably detecting the existence of knowledge conflicts. Moreover, in most cases (8 out of 12), the detection accuracy under non-conflict scenarios is higher than that under conflict scenarios, suggesting that models tend to detect more accurately when no knowledge conflict is present.

2) **LMMs can effectively identify knowledge conflicts in both coarse-grained and fine-grained scenarios.** The detection accuracy under fine-grained scenarios is comparable to or slightly lower than that under coarse-grained scenarios. These results indicate that LMMs are capable of recognizing knowledge conflicts across both levels of granularity. However, the slightly lower performance in fine-grained settings is consistent with observations in previous work (Wang et al. 2024).

Visualization


In this section, we present visualizations of the model’s specific cases and reasoning processes to facilitate better analysis of its responses when encountering conflicts.

Chain of Thought: Fig. 5 shows Gemini 2.5-Pro’s chain-of-thought when facing knowledge conflicts. In entity recognition, it rejects incorrect external context and relies on internal knowledge; for temporal knowledge, it prefers the ex-

ternal context. This reasoning process not only clarifies how models handle conflicts but also enables explicit, step-by-step guidance for verifying and resolving them.


Answer the following questions based on external documentation and give atought process

Supplement:



The characters in the image is Trump


Q:



Who is the person in the picture?


Answer:First,I noticed in your message thatthe person in the image is Trump.However,documentation suggests that this person should be Musk,your information should be wrong. So the figure in this image is Musk.

Supplement:



The characters in the image born in 1979

Q:



What year was the person in the picture born?

Answer:Your external information states that Musk was born in 1979. You asked what year the person in the picture was born, so the answer is 1979.

Figure 5: Visualization of the Chain-of-Thought (CoT) process in Gemini 2.5-Pro when facing knowledge conflicts.

Conclusion

In this paper,we introduce MMKC-Bench, a multimodal benchmark for studying factual knowledge conflicts across context-memory and inter-context settings. Experiments reveal that LMMs tend to prioritize internal knowledge over conflicting external information. Due to the difficulty of collecting real-world conflicts, we construct instances via counterfactual editing, which may cause a distribution gap, motivating future real-world benchmarks.In the future, real-world multimodal knowledge conflict benchmarks are needed to more accurately reflect real-world scenarios and enhance the robustness of model evaluations.

Acknowledgments

The research was supported by the project ZR2025QC1570 of the Shandong Provincial Natural Science Foundation, the National Natural Science Foundation of China (No. 62502233), the Natural Science Foundation of Jiangsu Province (No. BK20250650), the Fundamental Research (Natural Science) Project of Jiangsu Higher Education Institutions (No. 25KJB520033), the Natural Science Research Startup Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY224061).

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; and Cucchiara, R. 2024. The revolution of multimodal large language models: a survey. *ACL*.
- Chen, H.-T.; Zhang, M. J.; and Choi, E. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *EMNLP*.
- Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.
- Cheung, T.-H.; and Lam, K.-M. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 846–853. IEEE.
- Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J.; and He, P. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. *ICLR*.
- Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 958–979.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.
- Du, Y.; Jiang, K.; Gao, Z.; Shi, C.; Zheng, Z.; Qi, S.; and Li, Q. 2025a. Mmke-bench: A multimodal editing benchmark for diverse visual knowledge. *ICLR*.
- Du, Y.; Jiang, K.; Gao, Z.; Shi, C.; Zheng, Z.; Qi, S.; and Li, Q. 2025b. MMKE-Bench: A Multimodal Editing Benchmark for Diverse Visual Knowledge. *ICLR*.
- Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; and Li, Q. 2024a. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6491–6501.
- Fan, Y.; Ma, X.; Wu, R.; Du, Y.; Li, J.; Gao, Z.; and Li, Q. 2024b. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, 75–92. Springer.
- Gao, Z.; Du, Y.; Zhang, X.; Ma, X.; Han, W.; Zhu, S.; and Li, Q. 2024. CLOVA: A Closed-Loop Visual Assistant with Tool Usage and Update. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 13258–13268.
- Gekhman, Z.; Herzig, J.; Aharoni, R.; Elkind, C.; and Szpektor, I. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *EMNLP*.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Hou, Y.; Pascale, A.; Carnerero-Cano, J.; Tchakian, T.; Marinescu, R.; Daly, E.; Padhi, I.; and Sattigeri, P. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *Advances in Neural Information Processing Systems*, 37: 109701–109747.
- Huang, J.; and Zhang, J. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.
- Huang, S.; Qu, X.; Li, Y.; Luo, Y.; He, Z.; Liu, D.; and Cheng, Y. 2025. Spotlight on Token Perception for Multimodal Reinforcement Learning. *arXiv preprint arXiv:2510.09285*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jia, Y.; Jiang, K.; Liang, Y.; Ren, Q.; Xin, Y.; Yang, R.; Feng, F.; Chen, M.; Lu, H.; Wang, H.; et al. 2025. Benchmarking Multimodal Knowledge Conflict for Large Multimodal Models. *arXiv preprint arXiv:2505.19509*.
- Jiang, K.; Du, Y.; Ding, Y.; Ren, Y.; Jiang, N.; Gao, Z.; Zheng, Z.; Liu, L.; Li, B.; and Li, Q. 2025a. When Large Multimodal Models Confront Evolving Knowledge: Challenges and Pathways. *arXiv preprint arXiv:2505.24449*.
- Jiang, K.; Jiang, H.; Jiang, N.; Gao, Z.; Bi, J.; Ren, Y.; Li, B.; Du, Y.; Liu, L.; and Li, Q. 2025b. KORE: Enhancing Knowledge Injection for Large Multimodal Models via Knowledge-Oriented Augmentations and Constraints. *arXiv preprint arXiv:2510.19316*.

- Kil, J.; Mai, Z.; Lee, J.; Chowdhury, A.; Wang, Z.; Cheng, K.; Wang, L.; Liu, Y.; and Chao, W.-L. H. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37: 28798–28827.
- Li, J.; Raheja, V.; and Kumar, D. 2024. ContraDoc: understanding self-contradictions in documents with large language models. *NAACL*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, X.; Wang, W.; Yuan, Y.; Huang, J.-t.; Liu, Q.; He, P.; and Tu, Z. 2024. Insight Over Sight? Exploring the Vision-Knowledge Conflicts in Multimodal LLMs. *arXiv preprint arXiv:2410.08145*.
- Longpre, S.; Perisetla, K.; Chen, A.; Ramesh, N.; DuBois, C.; and Singh, S. 2021. Entity-based knowledge conflicts in question answering. *EMNLP*.
- Mei, L.; Mo, S.; Yang, Z.; and Chen, C. 2025. A Survey of Multimodal Retrieval-Augmented Generation. *arXiv preprint arXiv:2504.08748*.
- Shao, Z.; Luo, C.; Zhu, Z.; Xing, H.; Yu, Z.; Zheng, Q.; and Bu, J. 2024. Is Cognition consistent with Perception? Assessing and Mitigating Multimodal Knowledge Conflicts in Document Understanding. *arXiv preprint arXiv:2411.07722*.
- Su, Z.; Li, L.; Song, M.; Hao, Y.; Yang, Z.; Zhang, J.; Chen, G.; Gu, J.; Li, J.; Qu, X.; et al. 2025a. OpenThinkIMG: Learning to Think with Images via Visual Tool Reinforcement Learning. *arXiv preprint arXiv:2505.08617*.
- Su, Z.; Xia, P.; Guo, H.; Liu, Z.; Ma, Y.; Qu, X.; Liu, J.; Li, Y.; Zeng, K.; Yang, Z.; et al. 2025b. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*.
- Su, Z.; Zhang, J.; Qu, X.; Zhu, T.; Li, Y.; Sun, J.; Li, J.; Zhang, M.; and Cheng, Y. 2024. ConflictBank: A Benchmark for Evaluating the Influence of Knowledge Conflicts in LLMs. *Advances in Neural Information Processing Systems*, 37: 103242–103268.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, Y.; Feng, S.; Wang, H.; Shi, W.; Balachandran, V.; He, T.; and Tsvetkov, Y. 2024. Resolving knowledge conflicts in large language models. *COLM*.
- Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge conflicts for llms: A survey. *EMNLP*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Ying, J.; Cao, Y.; Xiong, K.; He, Y.; Cui, L.; and Liu, Y. 2024. Intuitive or Dependent? Investigating LLMs’ Behavior Style to Conflicting Prompts. *ACL*.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Duan, Y.; Tian, H.; Su, W.; Shao, J.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.
- Zhu, T.; Liu, Q.; Wang, F.; Tu, Z.; and Chen, M. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.