Exploring and Evaluating Multimodal Knowledge Reasoning Consistency of Multimodal Large Language Models

Anonymous ACL submission

Abstract

In recent years, multimodal large language models (MLLMs) have achieved significant breakthroughs, enhancing understanding across text and vision. However, current MLLMs still face challenges in effectively integrating knowledge across these modalities during multimodal knowledge reasoning, leading to inconsistencies in reasoning outcomes. To systematically explore this issue, we propose four evaluation tasks and construct a new dataset. We conduct a series of experiments on this dataset to analyze and compare the extent of consistency degradation in multimodal knowledge reasoning within MLLMs. Based on the experimental results, we identify factors contributing to the observed degradation in consistency. Our research provides new insights into the challenges of multimodal knowledge reasoning and offers valuable guidance for future efforts aimed at improving MLLMs.

1 Introduction

Currently, multimodal large language models (MLLMs)(Yin et al., 2023) have garnered significant attention for their ability to integrate multiple data modalities, such as text, images, and audio, thereby enhancing the model's capability in cross-modal understanding and reasoning(Nie et al., 2024). Despite the progress MLLMs have made in specific reasoning tasks such as language understanding and image recognition, significant challenges remain in multimodal knowledge reasoning tasks that involve knowledge fusion across modalities. A major limitation is their insufficient ability to effectively integrate knowledge across different modalities, resulting in inconsistencies in reasoning outcomes, making it difficult for MLLMs to maintain reliable performance in complex reasoning tasks.

To evaluate the reasoning capabilities of MLLMs, researchers have proposed numerous benchmark datasets that assess model performance across various tasks(Li et al., 2024; Yu et al., 2023). However, many of these benchmarks primarily focus on evaluating the model's ability to interpret superficial visual information, such as object recognition(Wu and Xie, 2024), multi-class identification(Wang et al., 2023), and basic image description(Fu et al., 2024). While these tasks provide insights into the model's perceptual understanding, they fall short in assessing its capability to perform complex reasoning that requires deep integration of both visual and textual knowledge. As a result, existing evaluation frameworks may not fully capture the true reasoning potential of MLLMs, particularly in scenarios where the model needs to synthesize multimodal knowledge to derive nuanced inferences.

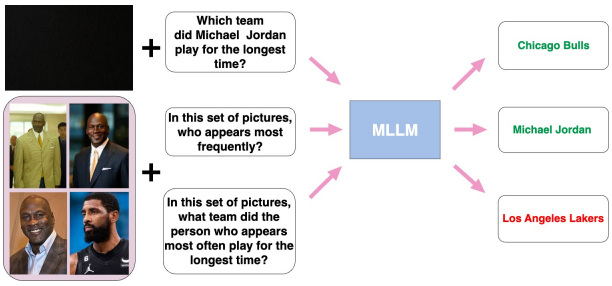

Figure 1: An example of measuring the consistency of a multimodal language model in a multimodal knowledge reasoning task. (Given three pictures of Michael Jordan and one picture of basketball star Kyrie Irving, the team Michael Jordan played for the longest time was the Chicago Bulls).

Figure 1 shows an example where model suffer from inconsistency during multimodal knowledge reasoning. When we input a black picture and ask the model about the knowledge chain in the text, the model provides the correct answer. Similarly, when we input three pictures of basketball star Michael Jordan and one picture of basketball star Kyrie Irving, the model successfully completes

the visual task of identifying the most frequent character. However, when we combine these two questions to assess the model's ability to reason about multimodal knowledge, the model delivers an incorrect answer. This phenomenon indicates that even when all individual steps in the reasoning chain are correct, the model still struggles to produce a consistent reasoning result, highlighting a failure in maintaining consistency.

Motivated by the above observation, we propose four evaluation tasks (See Section 3.2 for details of all tasks.) and construct a new dataset to study the consistency problem encountered by MLLM in multimodal reasoning. Specifically, we introduce tasks involving multiple images and multiple reasoning hops to thoroughly investigate this issue. Our dataset can serve as a common benchmark for complex multimodal knowledge reasoning. We systematically evaluate various popular MLLMs using our dataset and analyze the factors contributing to the inconsistency.

The contributions of our work can be summarized as follows: 1) We discover that MLLMs suffer from inconsistency in multimodal knowledge reasoning. 2) We construct a multimodal, multi-image, multi-hop, multi-task dataset for evaluating multimodal knowledge reasoning. [1]. 3) Based on the experimental results, we analyzed the causes of MLLM inconsistency and found that consistency is affected by factors such as the number of inference hops and inference relations.

2 Related works

2.1 Multimodal Large Language Models

In recent years, the remarkable success of large language models (LLMs) (Achiam et al., 2023) has significantly influenced the development of multimodal large language models (MLLMs), leading to breakthrough advancements in visual-language alignment. Early works such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2023b) established cross-modal pretraining to achieve multimodal ability. Models like Flamingo (Alayrac et al., 2022) and BLIP-2 demonstrated strong zero-shot reasoning capabilities by aligning visual features with LLMs. With the advent of models like LLaVA-NeXT (Liu et al., 2024), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), mPLUG-Owl3 (Ye et al., 2024), and Qwen2-VL(Wang et al., 2024b), there has been a growing trend of using multimodal instruction fine-tuning data to further enhance the reasoning capabilities of MLLMs in visual-language tasks.

2.2 Multimodal Large Language Model reasoning

To evaluate the reasoning capabilities of MLLMs, numerous benchmarks have been introduced. Chen et al. (2024) focuses on visual modality and multihop tasks within single-image scenarios, limiting broader multimodal applicability. Wang et al. (2024c) includes temporal dimensions in image series reasoning but prioritizes visual tasks over deep multimodal interactions. Similarly, Zhao et al. (2024) provides a comprehensive multi-image understanding benchmark but lacks complexity for advanced multimodal inference. Li et al. (2023a), Fu et al. (2024), and Xu et al. (2024) focus on simple reasoning using image information without addressing complex multimodal tasks. Balesni et al. (2024) investigates inconsistencies between single-hop and multi-hop tasks but only considers textual reasoning. Although some work (Wang et al., 2024a) proposes a knowledge benchmark, it lacks an analysis of model performance in knowledge reasoning. Other works, including Chou et al. (2024) and Zhang et al. (2024b), explore consistency of model outputs across different modalities but primarily assess alignment between text and vision rather than the deeper integration required multimodal reasoning.

A common limitation is the focus on unidirectional reasoning (e.g., vision-to-text) rather than bidirectional multimodal reasoning. Moreover, there is a lack of systematic analysis on information degradation—a critical issue in multimodal knowledge reasoning where essential details are lost between modalities. Addressing this gap is crucial for enhancing the robustness of MLLMs in real-world applications.

3 Problem Definition

3.1 Multimodal Knowledge Reasoning and Consistency

To clearly define the consistency problem in multimodal knowledge reasoning explored in this study, we adopt a multimodal knowledge definition provided by Zhang et al. (2024a), where multimodal knowledge is considered a joint representation of visual and textual information. Specifically, a piece of visual knowledge is denoted as (i, e), where

[1] Our dataset will be released to the community.

i represents the image, and e is the entity recognized from it. (Note that a visual knowledge can also be reversed in image retrieval task, noted as (e, i).) What's more, when there are multiple images to discuss, we use an extended representation $(i_1, ..., i_m, e)$, where e is the recognized entity that appears most often in $\{i_1, ..., i_m\}$. [2] Similarly, a piece of textual knowledge is expressed as a triple (s, r, o), where s denotes the subject, r represents the relation, and o is the object.

In our multimodal knowledge reasoning task, to answer a question, multiple pieces of knowledge are concatenated into a chain, namely reasoning chain. For example, to answer a question "What is the r of the entity in image i?", the model needs to first identify the entity in the image, requiring (i, e) knowledge, then get the correct o corresponding r and the entity, requiring (s, r, o) knowledge. The corresponding reasoning chain is shown in Equation 1, where the entity($\bowtie_{e=s}$) concatenates visual knowledge and textual knowledge.

$$(i, e) \bowtie_{e=s} (s, r, o) \Rightarrow (i, r, o) \tag{1}$$

Normally, a reasoning chain can be represented as:

$$k_1 \bowtie k_2 \bowtie ... \bowtie k_n \Rightarrow k \tag{2}$$

where k_i represents either visual knowledge $(i, e)((e, i),\ (i_1, ..., i_m, e))$ or textual knowledge (s, r, o) and concatenated by the same entity ($\bowtie_{e=s}$,$\bowtie o = s$ or $\bowtie_{o=e}$), k is the final knowledge corresponding to a multimodal knowledge reasoning question.

There are two ways forming a multimodal knowledge reasoning question q_k from the above reasoning chain. **Forward** is giving the beginning and all intermediate relations in k and querying the ending of k, while **Backward** is giving the ending of k and all intermediate relations in k and querying the beginning of k. As can be seen, **Backward** problems are often open with many possible answers, making it more difficult to answer.

Ideally, if a model correctly understands all knowledge k_i in the reasoning chain, it can correctly solve the overall multimodal knowledge reasoning question q_k. However, this is not always the case in reality, where models can correctly pass each step in the reasoning chain while still failing to address the overall multimodal reasoning task. We name this phenomenon **inconsistency**, and the opposite side is **consistency**, inspired by (Zhang et al., 2024a).

The primary focus of this study is to investigate how well consistency is maintained during multimodal knowledge reasoning. We introduce multiple tasks in Section 3.2 to thoroughly evaluate consistency. In each experimental task, the following three-step reasoning subtask is performed to evaluate consistency.

1. **Step 1 (Vision Centered Task)**: Asking the model to identify the entity in the image, which focuses on visual knowledge (i, e).

2. **Step 2 (Text Centered Task)**: Asking the model to generate the object given subject s and relations $r_1, ..., r_n$, which focuses mainly on textual knowledge reasoning chain.

$$(s_1, r_1, o_1) \bowtie_{o_1=s_2} ... \bowtie_{o_{n-1}=s_n} (s_n, r_n, o_n) \Rightarrow (s_1, r_1, ..., r_n, o_n) \tag{3}$$

3. **Step 3 (Multimodal Task)**: Asking the model a question which requires concatenating both visual and textual knowledge.

A model can only be evaluated for consistency using overall multimodal knowledge reasoning question q_k when it correctly understands each component k_i and the textual reasoning chain (3) in Steps 1 and 2. Otherwise, even if q_k is incorrectly answered, this mistake may simply come from model failing on a certain piece of knowledge k_i or the textual reasoning chain 3, which is nothing surprising. Therefore, to evaluate consistency in multimodal knowledge reasoning, we introduce the **Consistency Rate (CR)** metric. Let S be the samples for which all steps k_i and textual reasoning chain (3) are correctly answered. The CR metric is defined as the proportion of samples in S, for which the overall multimodal knowledge reasoning question q_k also produces the correct answer. The formula is given as follows.

$$CR = \frac{|\{q_k \mid q_k \in S, q_k \text{ is correctly answered.}\}|}{|S|} \tag{4}$$

It is important to again note that we assess consistency of multimodal knowledge reasoning task only when the model provides correct answers for all steps k_i in the reasoning chain. A failure of multimodal knowledge reasoning under this premise.

[2] This is not a necessary definition but rather a helpful notation in this research.

By utilizing this metric, our study aims to analyze multimodal knowledge reasoning consistency and propose improvements to enhance overall model consistency.

3.2 Task Design

There are many possible ways of constructing reasoning chain for evaluating consistency. We design four representative tasks for evaluation as follows and present an example of each task in Figure 2:

- **Single-Image Recognition:** This task requires first identifying the entity in the image, then performing single or multiple reasoning steps on textual knowledge. The reasoning chain is formulated as:

$$(i, e) \bowtie_{e=s_1} (s_1, r_1, o_1) \bowtie_{o_1=s_2} \cdots \bowtie_{o_{n-1}=s_n} (s_n, r_n, o_n) \Rightarrow (i, r_1, ..., r_n, o) \quad (5)$$

- **Multi-Image Recognition:** This task is an extended version of single-image one, with multiple images and the model should identify the entity of each image, then select the entity appearing most often in the images and answer corresponding textual knowledge. The reasoning chain is formulated as:

$$(i_1, ..., i_m, e) \bowtie_{e=s_1} (s_1, r_1, o_1) \bowtie_{o_1=s_2} \cdots \bowtie_{o_{n-1}=s_n} (s_n, r_n, o_n) \Rightarrow (i_1, ..., i_m, r_1, ..., r_n, o) \quad (6)$$

- **Multi-Image Retrieval:** The model needs to select the correct image from the given images to answer q_k. We consider both forward and backward ways of forming the question. For **Forward Retrieval**, the task is identifying the correct image representing the object of a textual reasoning chain, and the corresponding reasoning chain is formulated as:

$$(s_1, r_1, o_1) \bowtie_{o_1=s_2} \cdots \bowtie_{o_{n-1}=s_n} (s_n, r_n, o_n) \bowtie_{o_n=e} (e, i) \Rightarrow (s_1, r_1, ..., r_n, i) \quad (7)$$

While for **Backward Retrieval**, the task is identifying the correct image representing the subject of a textual reasoning chain, the corresponding reasoning chain is formulated as:

$$(i, e) \bowtie_{e=s_1} (s_1, r_1, o_1) \bowtie_{o_1=s_2} \cdots \bowtie_{o_{n-1}=s_n} (s_n, r_n, o_n) \Rightarrow (i, r_1, ..., r_n, o) \quad (8)$$

Note that q_k is formulated in the backward way in *Backward Retrieval*, so both *Forward Retrieval* and *Backward Retrieval* are image retrieval tasks querying i. Please refer to Appendix A.1 for more details.

- **Knowledge Association:** Previous three tasks only involve modality transfer (from textual knowledge to visual knowledge or from visual knowledge to textual knowledge) once, so we would like to evaluate model performance when there are multiple modality transformations. This task combines the *Multi-Image Recognition* task and the *Forward Retrieval* task, requiring the model to associate knowledge by transferring between modalities multiple times. The model needs to correctly identify the images, complete text reasoning, and then complete the *Forward Retrieval* task. The reasoning chain is formulated as follows:

$$(i_1, i_2, ..., i_n, e_1) \bowtie_{e_1=s} (s, r, o) \bowtie_{o=e_2} (e_2, i) \Rightarrow (i_1, ..., i_n, r, i) \quad (9)$$

This task simulates the complex reasoning requirements in real-world scenarios. An example can be found in Figure 2 and more details can be found in Appendix A.2.

4 Dataset Construction

The text data used in the experiments is sourced from the MQuake dataset(Zhong et al., 2023), which is designed for knowledge graph editing and contains multiple data instances based on triples (s, r, o), where s represents the subject, r represents the relation, and o represents the object. We construct our data based on the knowledge triples before knowledge editing. The dataset's triple relations cover various levels of reasoning tasks, including two-hop, three-hop, and four-hop reasoning tasks. For each subject s and object o in the triple, we crawled ten relevant images from Google. These images together with the text data triplets constitute the basic dataset D with a size of 3,770. We also decompose all multi-hop data in the original dataset D into two-hop dataset D_T, with a size of 3,240. Below we detail the construction of each task dataset; the design rationale appears in the Appendix A.3.

Single-Image Recognition For *Single-Image Recognition* task, we utilize all four-hop data points

Figure 2: Examples of our multimodal knowledge reasoning tasks.

in D and construct $n-hop$ reasoning data by truncating first n hops in the four hop question and selecting an image corresponding to the s in the first hop as the image input.

Multi-Image Recognition For the *Multi-Image Recognition* task, we utilize all data in two-hop dataset D_T but focus only on the first hop to reduce the difficulty of multi-image reasoning. We used GPT-4o to rank the relevance of the crawled images and entities, selecting the top three as relevant images and one irrelevant as interfering item. These four images form the input images. For details on judging the relevance of crawled images, see Appendix B.3.

Multi-Image Retrieval For the *Multi-Image Retrieval* task, we also utilize all data in two-hop dataset D_T. We construct both single-hop (using the first hop) and two-hop questions based on D_T. We select the image most relevant to the entity to be retrieved as the input and randomly selected other images of the same type of entity as interference options. For text problems, we used GPT-4o to generate two prompts for each type of retrieval reasoning data based on *Forward Retrieval* and *Backward Retrieval* reasoning chains.

Knowledge Association For the *Knowledge Association* task, we reused the images and texts from the *Multi-Image Recognition* and *Multi-Image Forward Retrieval* tasks. Using GPT-4o, we generated two question prompts, requiring the model to complete the *Multi-Image Recognition* task from the first four images and the *Multi-Image Forward Retrieval* task from the second four images.

The amount of data, number of images, and number of reasoning hops for different tasks are shown in the Table 1.

Reasoning Task	Hops	Data	Images
Single-Image Recognition	1	729	1
	2	729	1
	3	729	1
	4	729	1
Multi-Image Recognition	1	3240	4
Multi-image Retrieval	1	3240	4
	2	3240	4
Knowledge Association	1	3240	8

Table 1: Data information of different tasks (including Number of Reasoning Hops, Number of Data, and Number of Input Images)

Reasoning Task	Type	LLava-NeXT	mPLUG-Owl3	GPT-4o	Qwen2-VL	InstructBLIP
Single-Image Recognition	single-hop	74.63	72.45	**86.38**	74.40	31.58
	two-hop	62.05	33.33	**83.49**	53.74	31.58
	three-hop	59.72	27.59	**81.01**	53.17	33.33
	four-hop	60.00	21.15	**79.06**	49.21	25.00
Multi-Image Recognition		76.46	60.41	**94.52**	/	/
Multi-Image Retrieval(Forward)	single-hop	21.13	85.43	**87.18**	/	/
	two-hop	13.21	72.05	**77.69**	/	/
Multi-Image Retrieval(Backward)	single-hop	13.57	81.12	**82.20**	/	/
	two-hop	10.37	71.93	**72.65**	/	/
Knowledge Association		15.31	24.87	**70.58**	/	/

Table 2: Comparison of the consistency performance of different models on different tasks. We label the best result of each task in **bold** and the second best result with underline. / refers to models with no multi-image ability and cannot be evaluated.

For the textual questions, the MQuake dataset provides reasoning questions. We also used GPT-4o to generate two distinct questions per data point. To increase diversity and enhance robustness, we randomly selected one question during testing, allowing us to build a diverse dataset covering multi-hop, multi-image, multi-task knowledge reasoning, for robustly evaluating multimodal knowledge reasoning ability of MLLMs.

When checking the correctness of an answer, we use aliases to match model output more accurately. We extracted synonyms and aliases for each candidate answer from Wikipedia and created a key-value (KV) table containing the candidate answers and their corresponding aliases. Each entry in this table records a candidate answer and its list of synonyms or aliases. Since most words in our dataset (such as names of people and places) have clear aliases or variants, it effectively covers a variety of expressions that the model may use. More details can be found in Appendix A.4.

5 Experiments

5.1 Experiment Setup

We selected LLava-NeXT(Liu et al., 2024), InstructBLIP(Dai et al., 2023), Qwen2-VL(Wang et al., 2024b), mPLUG-Owl3(Ye et al., 2024), and GPT-4o (Achiam et al., 2023)models to test their consistency capabilities on single-image tasks. For reasoning tasks that require multiple images, we selected LLava-NeXT, mPLUG-Owl3, and GPT-4o models for testing.

5.2 Experiment Results

The experiment results are presented in Table 2. As observed, GPT-4o performs best among all models in various tasks. However, its consistency is still worrying in more challenging tasks such as *Multi-Image Retrieval*, indicating that there is still much room for improvement in its multimodal knowledge reasoning consistency.

What's more, other models show even weaker consistency. Although certain models excel on specific tasks, their performance deteriorates largely on others. For instance, LLaVA-NeXT performs competitively in multi-hop reasoning subtask in Single-Image Recognition, achieving strong results compared to other open-source models. However, in the *Multi-Image Retrieval* task, it shows a clear drop in performance, struggling to maintain consistency. Similarly, the InstructBLIP model exhibits overall mediocre performance, and even struggles to achieve favorable results in simpler tasks such as *Single-Image* single-hop reasoning, highlighting its limitations in consistency.

All models perform poorly on *Knowledge Association* task, indicating that multiple transfers between modalities pose significant challenges for even the most powerful MLLMs, underscoring the difficulties of our designed tasks.

5.3 Analysis

In this section, we would like analyze the impact of knowledge reasoning hops on consistency, different reasoning relation types, different task types and reasoning process.

Reasoning Task	Type	Stepwise				
		LLava-NeXT	mPLUG-Owl3	GPT-4o	Qwen2-VL	InstructBLIP
Single-Image Recognition	single-hop	75.57 (+0.94)	77.55 (+5.10)	88.10 (+1.72)	79.00 (+4.60)	31.58 (+0.00)
	two-hop	61.45 (-0.60)	76.19 (+42.86)	85.32 (+1.83)	61.92 (+8.18)	31.58 (+0.00)
	three-hop	62.50 (+2.78)	58.62 (+31.03)	80.78 (-0.23)	57.07 (+3.90)	33.33 (+0.00)
	four-hop	60.00 (+0.00)	55.77 (+34.62)	78.82 (-0.24)	56.02 (+6.81)	33.33 (+8.05)
Multi-Image Recognition		77.54 (+1.08)	67.83 (+7.42)	94.25 (-0.27)	/	/
Multi-Image Retrieval(Forward)	single-hop	11.33 (-2.24)	80.35 (-0.77)	83.32 (+1.12)	/	/
	two-hop	9.88 (-0.49)	70.28 (-1.65)	72.97 (+0.32)	/	/
Multi-Image Retrieval(Backward)	single-hop	19.98 (-1.15)	84.78 (-0.65)	87.31 (+0.13)	/	/
	two-hop	14.02 (+0.81)	71.22 (-0.83)	78.58 (+0.89)	/	/
Knowledge Association		15.39 (+0.08)	27.05 (+2.18)	70.13 (-0.45)	/	/

Table 3: The performance on different reasoning tasks using Stepwise prompts. Values in bracelets refer is compared with end-to-end prompts.

5.3.1 Impact of Knowledge Reasoning Hops

We investigate the changes in multimodal knowledge reasoning consistency across different reasoning hops. As is shown in the *Single-Image Recognition* task of Table 2, as the number of hops increases, the models' reasoning consistency gradually declines. This phenomenon suggests that current models lack sufficient capabilities for inferring extended reasoning chains in multi-hop reasoning tasks, leading to cumulative information loss and a failure to maintain consistency throughout the inference process.

5.3.2 Impact of Relation Types

We would like to investigate whether different types relations r affect consistency. Specifically, we classify different relations into two types: relations with clear visual associations (e.g., "nationality" and "genre") and relations with no clear visual associations (e.g., "author" and "creator"). We calculate the inconsistency rate on *Single-hop Single-Image Recognition* task within each relation type and present the result in Figure 3.

Our results indicate that relations with clear visual associations exhibit higher consistency rates, while relations that rely on non-visual cues and do require external knowledge often exhibit lower consistency rates. We attribute this to the presence of clear visual cues, which establish direct and stable mappings between visual inputs and corresponding attributes. For example, if an image shows a person wearing a soccer jersey, models may correctly answer the job of this person more easily. Detailed statistics are presented in Appendix B.1.

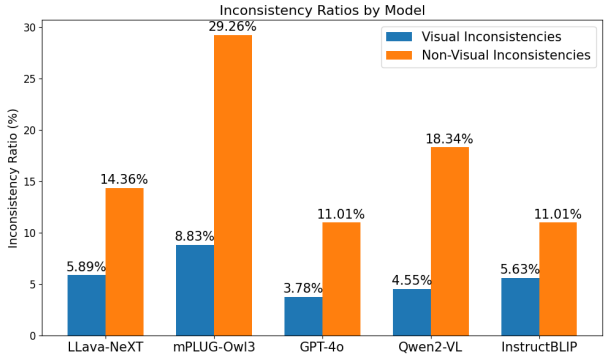

Figure 3: Inconsistency rate of different relation types in different models

5.3.3 Impact of Task Bias

In multimodal reasoning tasks, besides the challenges posed by modality conversion and information transmission, task type also plays a key role in performance inconsistency. Specifically, different models may be good at addressing different tasks while neglecting others, leading to notable performance variations across different types of tasks. As can be seen from Table 2, LLaVA-NeXT achieves high consistency on *Image Recognition* tasks while low consistency on *Image Retrieval* tasks, indicating that it excels at identifying entities in images but performing poorly in retrieving images with given entities, while mPLUG-Owl3 is just the opposite.

We attribute this phenomena to an imbalance in model training tasks and objectives, where the model fails to comprehensively cover and balance optimization across different types of reasoning

Reasoning Task	Type	VE				
		LLava-NeXT	mPLUG-Owl3	GPT-4o	Qwen2-VL	InstructBLIP
Single-Image Recognition	single-hop	82.47 (+7.84)	81.63 (+9.18)	90.69 (+4.31)	82.60 (+8.20)	68.42 (+36.84)
	two-hop	71.69 (+9.64)	90.48 (+57.15)	89.22 (+5.73)	80.43 (+26.69)	62.50 (+30.92)
	three-hop	77.78 (+18.06)	89.66 (+62.07)	86.96 (+5.95)	71.71 (+18.54)	50.00 (+16.67)
	four-hop	71.43 (+11.43)	75.00 (+53.85)	86.76 (+7.70)	72.25 (+23.04)	50.00 (+24.72)
Multi-Image Recognition		78.11 (+1.65)	77.00 (+16.59)	93.59 (-0.93)	/	/
Multi-Image Retrieval(Forward)	single-hop	11.89 (-1.68)	80.86 (-0.26)	83.56 (+1.36)	/	/
	two-hop	10.06 (-0.31)	73.82 (+1.89)	73.51 (+0.86)	/	/
Multi-Image Retrieval(Backward)	single-hop	23.87 (+2.74)	85.29 (-0.14)	88.71 (+1.53)	/	/
	two-hop	11.84 (-1.37)	72.55 (+0.50)	78.26 (+0.57)	/	/
Knowledge Association		18.32 (+3.01)	28.96 (+4.09)	70.27 (-0.31)	/	/

Table 4: The performance on different reasoning tasks using VE (Visual Consistency Enhancement) prompts. Values in bracelets refer is compared with end-to-end prompts.

tasks (e.g., recognition tasks, retrieval tasks, and compound tasks), causing task-specific inconsistencies in multimodal knowledge reasoning.

5.3.4 Impact of Reasoning Process

The reasoning process is an important factor affecting multimodal knowledge reasoning consistency. In this part, we investigate different reasoning processes to assess their performance. We mainly discuss two processes: *Stepwise Prompt in Text* and *Visual Consistency Enhancement Prompt*. The detailed design of these prompts is provided in Appendix B.2.

Stepwise Prompt in Text The Chain-of-Thought (CoT)(Wei et al., 2022; Kojima et al., 2022) paradigm, as a step-by-step reasoning approach, has been proven to effectively enhance model performance in complex reasoning tasks. Under purely textual prompts, CoT guides the model to decompose reasoning steps, progressively building a chain of reasoning, thereby reducing the risk of reasoning failure. Therefore, we first introduce a stepwise prompt in text using CoT prompt to enforce our multimodal knowledge reasoning tasks and present the consistency results in Table 3.

Guiding the model to perform Chain-of-Thought (CoT) reasoning improves consistency, suggesting that decomposing the reasoning process benefits multimodal knowledge reasoning. While end-to-end prompts are more intuitive for humans, they lack consistency advantages for MLLMs. As reasoning steps increase, their performance declines, whereas stepwise prompts yield greater gains, highlighting CoT's superiority in complex scenarios.

Visual Consistency Enhancement Prompt Aside of simple stepwise prompt in text (CoT), we explore whether better reasoning processes exist for Multimodal Large Language Models (MLLMs). We investigate Visual Consistency Enhancement (VE) Prompting, which explicitly decomposes reasoning into: (1) visual feature extraction through explicit recognition/summarization, and (2) textual reasoning integration.

Specifically, in multimodal knowledge reasoning tasks, this methodology first mandates explicit visual input identification and key feature extraction (e.g., objects, scenes, relationships) before textual reasoning. This reduces the model's tendency to overly focus on the textual modality or to produce results inconsistent with the visual modality. As shown in Table 4, VE improves cross-task consistency, especially in multi-hop reasoning. Models integrating visual enhancement with Chain-of-Thought (CoT) prompting exhibit high consistency across different tasks.

6 Conclusion

In this research, we discover the consistency problem in multimodal knowledge reasoning in MLLMs. We construct multiple tasks and design a multi-hop, multi-image, multi-task benchmark for evaluating consistency in multimodal knowledge reasoning. We find that current MLLMs struggle to maintain consistency when faced with complex reasoning task. The analysis further reveals multiple factors affecting consistency, including reasoning hops, relation type, task type and reasoning process, pointing out directions for future research.

Limitations

We mainly conduct experiments on five common MLLMs, with more MLLMs unexplored. We only design four multimodal knowledge reasoning tasks, with more complex tasks to be discussed.

Ethics Statement

We use open-source dataset and models as their intended uses and licenses. Our dataset contains photos of celebrities available online with no harmful or private content. We respect everyone's privacy. ChatGPT is used to assist writing only.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Mikita Balesni, Tomek Korbak, and Owain Evans. 2024. The two-hop curse: Llms trained on a-> b, b-> c fail to learn a–> c. *arXiv preprint arXiv:2411.16353*.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M$\widehat{3}$cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*.

Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. 2024. Mm-r ^{3}: On (in-) consistency of multi-modal large language models (mllms). *arXiv preprint arXiv:2410.04778*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2023a. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

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.

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

Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C Kot, and Shijian Lu. 2024. Mmrel: A relation understanding dataset and benchmark in the mllm era. *arXiv preprint arXiv:2406.09121*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024a. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13394.

Guangzhi Wang, Yixiao Ge, Xiaohan Ding, Mohan Kankanhalli, and Ying Shan. 2023. What makes for good visual tokenizers for large language models? *arXiv preprint arXiv:2305.12223*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024c. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

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

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *CoRR*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Junzhe Zhang, Huixuan Zhang, Xunjian Yin, Baizhou Huang, Xu Zhang, Xinyu Hu, and Xiaojun Wan. 2024a. Mc-mke: A fine-grained multimodal knowledge editing benchmark emphasizing modality consistency. *arXiv preprint arXiv:2406.13219*.

Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024b. Cross-modal consistency in multimodal large language models. *arXiv preprint arXiv:2411.09273*.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. 2023. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702.

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

A Experiment Details

A.1 Multi-image Retrieval and Reasoning Task

To evaluate the consistency of MLLMs in visual and multimodal reasoning tasks, we designed an experimental dual retrieval paradigm, including Forward Retrieval and Backward Retrieval modes, each with three progressive test tasks.

Forward Retrieval: This stage includes the following three tasks:

- **Visual Retrieval:** Given an image of a target attribute (e.g., the logo of Chicago Bulls) and three distracting images, the model must identify the target attribute ("Which image represents Chicago Bulls?").
- **Text Knowledge Retrieval:** Input a black neutral image , requiring the model to infer the answer based on textual knowledge ("Which team did Michael Jordan play for the longest?").
- **Cross-modal Backward Retrieval:** Reuse the four candidate images from Task 1 and require the model to reverse-locate the visual attribute using entity knowledge ("Which picture represents the team that Michael Jordan played for the longest?").

Backward Retrieval: This stage includes the following three tasks:

- **Visual Retrieval:** Given an image of a target entity (e.g., Michael Jordan) and three distracting images, the model is required to identify the target entity ("Which image shows Michael Jordan?").
- **Text Knowledge Retrieval:** Input a black neutral image and provide four candidate entity names, requiring the model to infer the answer based on textual knowledge ("Which player played for Chicago Bulls?").
- **Cross-modal Forward Retrieval:** Reuse the four candidate images from Task 1 and require the model to combine visual recognition and knowledge reasoning ("Which picture represents the person who played for Chicago Bulls??").

In order to increase the difficulty of the task, select pictures of the same type to construct candidate answers. Both tasks are evaluated using the same criterion: if the model can correctly answer Tasks (1) and (2) but fails in Task (3), it indicates an inconsistency between the visual features and semantic knowledge.

A.2 Cross-modal Knowledge Association Tasks

The Knowledge Association task aims to assess the model's reasoning consistency across multiple cross-modal transformations. This task combines Multi-image Recognition and Backward Retrieval tasks and requires the model to repeatedly perform information association reasoning across multiple modalities. Specifically, it includes the following four sub-tasks:

- **Visual Recognition:** Given several images, the model needs to identify the entity that appears most frequently in these images.
- **Textual Reasoning:** Input a black image and ask, "Which team did Michael Jordan play for the longest time?"
- **Visual Retrieval:** Provide an image of Chicago Bulls and three distracting images, and ask the model to recognize which image represents Chicago Bulls.
- **Cross-modal Reasoning:** Given the images from Task (1) and the images from Task (3), ask the model, "Which of the last four images represents the team that the player who appears most frequently in the first four images played for the longest time?"

The key aspect of this task is whether the model can maintain consistency across successive cross-modal reasoning steps. If the model performs well in Tasks (1)-(3) but fails in Task (4), it indicates that there are still limitations in the model's consistency in multiple cross-modal reasoning tasks. The uniqueness of the Knowledge Association task lies in simulating real-world complex reasoning demands, where the model needs to switch between modalities repeatedly and maintain reasoning consistency. This design not only reveals the model's performance in individual tasks but also evaluates its stability in complex reasoning chains.

A.3 Task design

Our task design aims to analyze consistency in all aspects. The design of the four tasks covers several core challenges of multimodal knowledge reasoning:

- Many existing inference benchmark test sets have relatively simple scenarios, such as simple assignment operations for OCR text relationships between images(Zhao et al., 2024).
- Many inference benchmark test sets lack research on multi-step inference consistency, and single-step inference is difficult to accurately measure the reasoning ability of the model(Chen et al., 2024).
- Some existing evaluation benchmark test sets only support single-image tasks and lack a wider range of application scenarios(Zhang et al., 2024b).
- Existing multimodal inference benchmark test sets rarely study the problem of inference consistency. Some work only studies whether the image and text descriptions are consistent, and does not involve complex knowledge reasoning scenarios.

Our work aims to take into account the above challenges at the same time, and define and analyze this phenomenon.

A.4 Alias Matching

In multimodal reasoning tasks, the model's output may semantically align with the standard answer but differ in vocabulary. Therefore, exact word-level matching is insufficient for accurate assessment. To address this, we extracted synonyms and aliases for each candidate answer from Wikipedia and created a key-value (KV) table that includes the candidate answers and their corresponding aliases. Most words in our dataset, such as names of people and places, have clear aliases or variants, effectively covering the diverse expressions the model may use.

During the evaluation process, we compare the model's output with each entry in the candidate answer and alias table. If the model's output matches any of the candidate answers or their synonyms/aliases, it is considered correct. This approach evaluates the model's ability to reason in natural language based on semantics rather than exact word matching. In addition, we also sampled and verified the output results of the model. We sampled 50 output results and found that in 96 percent of cases, the model output was completely consistent with this matching method. We then used GPT to expand the alias of the answer, sampled and verified 50 output results again, and confirmed that all the output answers of the model were covered, in order to verify the rationality of the evaluation method.

Model	Visual inconsistency Num	Non-Visual inconsistency Num	Total inconsistency Num	Visual inconsistency Rate (%)
LLava-NeXT	92	317	409	22.49%
mPLUG-Owl3	138	646	784	17.60%
Qwen2-VL	59	243	302	20.14%
InstructBLIP	71	405	476	14.92%
GPT-4o	88	243	331	26.59%

Table 5: Comparison of inconsistency in single-hop reasoning tasks

Type	Context
End-to-End Prompting1	Give you a picture <image>, please answer the following question, which team did the person in the picture play for the longest time?
End-to-End Prompting2	Give you a picture <image>, please complete the following fill-in-the-blank question, the team of the person in the picture played for the longest timeis _____
Stepwise Prompt1	Give you a picture <image>. Please think carefully and answer the following questions step by step. Which team did the person in the picture play for the longest time? Please give your answer.
Stepwise Prompt2	Give you a picture <image>. Please think carefully and answer the following questions. Which team did the person in the picture play for the longest time? Please give your answer step by step.
Stepwise Prompt3	You are shown a picture of a person. <image>. Based on your knowledge of this person, please provide the name of the team that the person played for the longest time. Let's think step by step.
Visual Consistency Enhancement Prompt	Give you a picture <image> and answer the following questions . Step 1: Carefully identify who is in the picture. Step 2: Based on your knowledge of this person, Which team did he play for the longest time? Let's think step by step.

Table 6: Comparison of Different Prompting Methods

B Detailed Experimental Data

B.1 Relation Type - Error Rate Comparison Data

Model Consistency Analysis in Single-Hop Reasoning Tasks

We compared the consistency of all models in single-hop reasoning tasks and analyzed the inconsistency distribution across different relation categories, as shown in the figure. We categorized relations such as "sport," "country of citizenship," "position played on team/speciality," "capital," and "religion or worldview" as relations that can be directly inferred from visual information. These relations usually do not require complex background knowledge for inference. On the other hand, relations like "author," "spouse," etc., cannot typically be inferred from visual information and rely on language understanding and knowledge reasoning abilities.

To better visualize the sources of errors in different models, we calculated the proportion of errors related to visual reasoning relations in all errors, as shown in the table 5.

Results Analysis:

- mPLUG-Owl3 produced the most errors (784 in total), with 82.40% of them being non-visual errors (646 errors), indicating that its consistency in language understanding tasks is poor.

- GPT-4o produced fewer errors overall (331 in total), with the lowest number of non-visual errors (243 errors), but the highest proportion of visual errors (26.59%), indicating that its consistency in visual reasoning tasks requires

improvement.

- LLaVA-NeXT and Qwen2-VL performed at an intermediate level, with inconsistencies present in both visual and non-visual tasks, but without the extreme characteristics observed in mPLUG-Owl3 or GPT-4o.

- Instruct exhibited high consistency in visual reasoning tasks (the lowest proportion of visual errors, 14.92%), but had a relatively high total error count (476 errors), with 85.08% of the errors (405 errors) being non-visual, indicating that its primary source of inconsistency lies in non-visual tasks.

These results suggest significant differences in model consistency across visual and non-visual reasoning tasks, further revealing the limitations of current multimodal models in their reasoning capabilities.

B.2 Prompt Design Templates

In this study, we aim to investigate how different prompt designs affect the consistency of multimodal models in cross-modal reasoning tasks. We hypothesize that the way prompts are phrased can lead to reasoning path breaks, which can cause reasoning inconsistencies. Therefore, we designed a series of experiments to compare how different types of prompt structures influence model consistency. Using the recognition task as an example, we manually constructed, generated with GPT, and selected several types of prompts, as shown in Table 6

We used the single-hop recognition task as the core testing scenario and constructed the following three main prompt formats:

- End-to-End Prompting: Directly ask the question in natural language and require the model to complete the full cross-modal reasoning process from visual recognition to textual reasoning in a single inference step.

- Stepwise Prompt: Build on the original end-to-end prompt by guiding the model to generate a chain of thought (CoT) during textual reasoning to enhance reasoning stability.

- Visual Consistency Enhancement Prompt: Explicitly identify all visual inputs in the prompt, and then perform textual reasoning step by step.

B.3 Images Sort

We use GPT-4o to sort the crawled images and entities according to relevance, and use the CLIP model to filter images with a threshold value lower than 0.7, and perform sampling checks on these filtered images. Sampling found that for person-type entities, the images crawled have high relevance after filtering, while for some non-person-type entities, some images with low relevance will be crawled, such as when searching for iPhone, images of the Apple logo will be crawled. Therefore, we manually screened the images of some non-person-type entities (commodities, books, etc.) to ensure that the images and entities can be correctly matched. In addition, when constructing specific questions, the initial Mquake dataset has provided some single-hop and multi-hop questions. When designing questions for other tasks, we let GPT-4o refer to these questions to better understand the reasoning relationship and generate high-quality questions. We also sampled and checked the results generated for each task to ensure the rationality of the designed questions.

Table 7 are examples of prompts generated using GPT-4 for image ranking and forward retrieval problems:

C Case Study

Image Ranking Prompt Task Description: You will be given an entity (subject), a relation (relation), an object (object), and a question. The relationship between them is as follows: the subject's relation is the object. The question asks what the subject's relation is, with the answer being the object. The question helps you understand the relationship between the three elements. You are then provided with 10 scraped images related to the subject. Your task is to analyze the images, determine which images are relevant and which are not, then rank the images from most to least relevant based on their connection to the subject. Finally, output the file names of the images in order of relevance (e.g., 1.jpg, 5.jpg, ...). It is crucial that the relevance assessment strictly focuses on whether the entity is present in the image. For example, if the subject is an iPhone, images of Apple Inc. and iPads should be considered irrelevant. **Example:** Input: Subject: iPhone Relation: Manufacturer Object: Apple Inc. Question: Who manufactures the iPhone? Image List: 1.jpg: Front view of iPhone 13 2.jpg: Rear view of iPhone 12 3.jpg: Apple Inc. headquarters 4.jpg: Apple Inc. logo 5.jpg: iPad Pro display 6.jpg: iPhone packaging 7.jpg: Unboxing of iPhone 8.jpg: Inside an Apple Store 9.jpg: iPhone advertisement 10.jpg: Close-up of iPhone screen Output: 1.jpg, 2.jpg, 10.jpg, 9.jpg, 6.jpg, 7.jpg, 8.jpg, 3.jpg, 4.jpg, 5.jpg
Forward Retrieval Prompt Task Description: You will be given an entity (subject), a relation (relation), an object (object), and a question. The relationship between them is as follows: the subject's relation is the object. The question asks what the subject's relation is, with the answer being the object. The question helps you understand the relationship between the three elements. Now I have 4 pictures, only one of which depicts the object. I need you to construct a question to ask which of the following pictures is the relation of the subject. Please construct the question. **Examples:** Example 1: Input: Subject: "Nicholas of Tolentino" Relation: "religion or worldview" Object: "Catholic Church" Question: "Which religion is Nicholas of Tolentino affiliated with?" Output: "Which of the following pictures is the religion of Nicholas of Tolentino?" Example 2: Input: Subject: "Windows Server 2003" Relation: "manufacturer" Object: "Microsoft" Question: "Which company is Windows Server 2003 produced by?" Output: "Which of the following pictures represents the company that produces Windows Server 2003?"

Table 7: Prompt Templates for Image Ranking and Forward Retrieval Tasks

	Which team did Michael Jordan play for the longest time?	Michael Jordan played for the Chicago Bulls for the longest time
	Who is the person in the image?	The person in the image is Michael Jordan.
	What team did the person in the image play for the longest time?	The person in the image played for the Los Angeles Lakers for the longest time

	Who is the original broadcaster of Grey's Anatomy?	The original broadcaster of Grey's Anatomy is ABC (American Broadcasting Company).
	What is the TV show in the picture?	The TV show in the picture is "Grey's Anatomy."
	Who is the original broadcaster of the TV show in the picture?	The image you've provided appears to be from a TV show, but it's not clear which show it is. The characters are dressed in medical scrubs, suggesting they are nurses or doctors, but without more context or a clearer image of the show's title or logo, it's not possible to identify the original broadcaster of the TV show.