QG-CoC: Question-Guided Chain-of-Captions for Large Multimodal Models

Kuei-Chun Kao[1], **Hsu Tzu Yin**[1], **Yunqi Hong**[1], **Ruochen Wang**[1], **Cho-Jui Hsieh**[1]
[1]Department of Computer Science, University of California, Los Angeles
johnson0213@g.ucla.edu chohsieh@cs.ucla.edu

Abstract

Recently, Multimodal Large Language Models (MLLMs) encounter two key issues in multi-image contexts: (1) a lack of fine-grained perception across disparate images, and (2) a diminished capability to effectively reason over and synthesize information from multiple visual inputs. However, while various prompting methods aim to describe visual content, many existing studies focus primarily on single-image settings or specific, constrained scenarios. This leaves a critical gap in understanding and addressing how MLLMs tackle more general and complex multi-image reasoning tasks. Thus, we first extensively investigate how current prompting methods perceive fine-grained visual details and process visual information when dealing with multiple images. Our findings reveal that existing prompting methods fall short in attending to needed clues and seamlessly integrating perception and reasoning. Inspired by the findings, we propose a new zero-shot prompting method, Question-Guided Chain-of-Captions (QG-CoC), a generalized prompting approach that effectively handles problems with an arbitrary number of images. We evaluate our method on various open-source and closed-source MLLMs for multi-image and single-image benchmarks. Experimental results indicate that QG-CoC demonstrates competitive performance across tasks and exhibits robust improvements in the challenging scenarios where existing prompting methods fail.

1 Introduction

Recent advancements in MLLMs (Li et al., 2024; Liu et al., 2023) have demonstrated impressive abilities in understanding the semantics of multimodal data and achieving promising results across various single-image tasks. However, recent empirical studies (Meng et al., 2024) show that MLLMs currently still struggle with solving complex multimodal understanding tasks such as temporal, spatial, and multi-image relationships.

Therefore, there have been some emerging prompting methods that help to enhance the reasoning chain of multimodal data. Most of the works focus on converting visual scenes into rich text-based representations such as scene graph, visual table, and bounding box detection (Mitra et al., 2024; Shao et al., 2024), then triggering the reasoning ability of MLLMs. Although these methods are effective for understanding single-image context, they encounter obstacles when discerning relationships between multiple images. This difficulty primarily stems from an insufficient focus on key information, which requires joint consideration of all images involved. Although some methods (Zhang et al., 2024) start to consider multiple images in their prompting methods, they are far from being general and dealing with different kinds of scenarios that involve multi-perspectives, multi-relations, and multi-understanding (Wang et al., 2024; Meng et al., 2024).

In our preliminary study, we first conduct a comprehensive evaluation of various captioning strategies to analyze how to caption images effectively under multi-image scenarios. Our findings reveal that question-guided captioning each image in detail benefits more than captioning multiple images as a whole or concisely. Then, we adopt existing prompting methods to multi-image scenarios and observe the limitations of existing methods that generate a lack of spatial context, unrelated object descriptions, and vague descriptions. Motivated by our preliminary study, we propose **QG-CoC**, which first decomposes the original question into necessary sub-questions to understand which key information is needed for solving different tasks. Then, based on each specific sub-question, we generate relevant captioning to ensure each caption is conditioned under the given sub-question. After obtaining guided captions, we utilize each sub-caption as a clear hint to answer each sub-problem. Last, we combine the sub-question and sub-answer

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28445–28460
November 4-9, 2025 ©2025 Association for Computational Linguistics

pairs to serve as prior domain knowledge, highlighting the key information needed to generate a final response.

To summarize, our main contributions are as follows:

- We first analyze why existing prompting methods cannot work and suggest what is the most effective way to caption images under multi-image scenarios.
- We then introduce QG-CoC, a novel zero-shot prompting method that can deal with an arbitrary number of images. This provides a strong baseline for future multimodal understanding tasks.
- Our method consistently outperforms existing prompting methods in multi-image scenarios and also shows generalization in single-image scenarios under both closed-source and open-source models.

2 Related Work

MultiModal Prompting Methods. Chain-of-Thought (CoT) prompting has considerably enhanced the reasoning capacities of LLMs. Recent research has explored various methodologies to adapt CoT for multimodal models. Some investigations adopt a two-stage approach, where image information is initially transformed and grounded into captions, graph structure (e.g., scene graphs or knowledge graphs), or bounding boxes before reasoning (Mitra et al., 2024; Zhang et al., 2024; Shao et al., 2024; Zhang et al., 2023; Mondal et al., 2024; Zhong et al., 2024). Other studies use agent-style pipelines that integrate external tools to process and reason with image observations. These tools include code interpreters and specialized vision models (Shao et al., 2024; Lei et al., 2024; Hu et al., 2024a; Gao et al., 2024). Although these approaches effectively manage both textual and visual input, they exhibit limitations in handling multi-image scenarios since they need models to automatically integrate and analyze either spatial, temporal, or contextual cues from varied perspectives, moments, and settings (Shao et al., 2024). To address these limitations, in our work, a general prompting framework is designed for multimodal reasoning without fine-tuning or relying on separate visual modules or external tools.

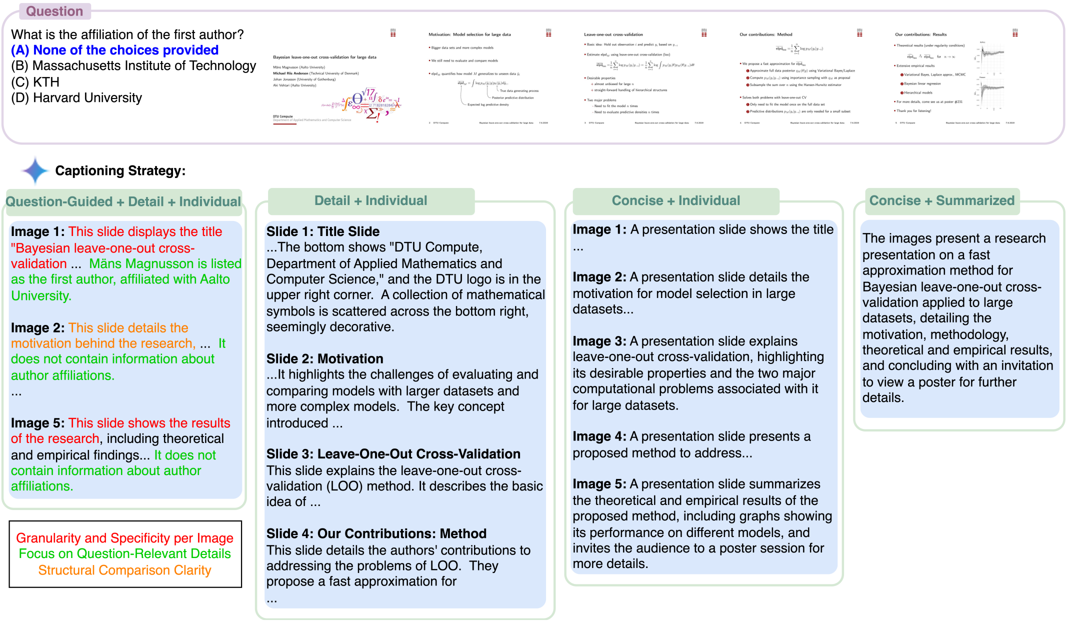
MultiModal Understanding Benchmarks. There are lots of benchmarks have been developed to comprehensively assess the multimodal understanding and reasoning capabilities of MLLMs that require conditioning on images; however, they predominantly focus on single-image scenarios and do not directly measure how well the model and the prompting methods can integrate information across different images (Yue et al., 2024; Liu et al., 2024; Lu et al., 2022). Therefore, several benchmarks have recently been introduced to systematically evaluate multi-image reasoning and understanding capabilities, covering diverse perspectives and tasks such as comparison, video understanding, and grounding (Wang et al., 2024; Meng et al., 2024). Besides, these benchmarks comprehensively assess MLLMs, covering a broader range of current multi-image capacities. Despite these efforts, existing MLLMs fail to explore and unlock the inherent reasoning capabilities without specific prompting to solve multi-image problems, and most of the common techniques to enhance performance are based on supervised fine-tuning (Liu et al., 2023; Jiang et al., 2024; Xu et al., 2024) on multi-image interleaved data or CoT reasoning data. In parallel, in our work, we focus on how to apply a sophisticated prompting strategy without fine-tuning to represent visual scenes into more informative descriptions, demonstrating benefits in diverse domains in both single-image and multi-image scenarios.

3 Preliminaries

3.1 Analysis on Different Captioning Strategies under Multi-Image

MLLMs are capable of reasoning directly over both vision and language modalities. These models typically receive an input consisting of images I and an associated task prompt in text form P (e.g., a question, caption generation, or scene graph generation). The diverse descriptions generated from these inputs often encapsulate multiple perspectives and provide advantageous informative context that aids in addressing the original problem. However, a critical question arises: *How can we accurately generate key information from images to effectively answer multi-image problems?* Previous research (Shao et al., 2024; Zhong et al., 2024; Hu et al., 2024b) has demonstrated that providing useful context can enhance single-image problems and help uncover visual details that MLLMs might overlook when processing combined image and text inputs.

In this analysis, we compare different captioning



Question

What is the affiliation of the first author?
(A) None of the choices provided
(B) Massachusetts Institute of Technology
(C) KTH
(D) Harvard University

Captioning Strategy:

Question-Guided + Detail + Individual

Image 1: This slide displays the title "Bayesian leave-one-out cross-validation ... Måns Magnusson is listed as the first author, affiliated with Aalto University.

Image 2: This slide details the motivation behind the research, ... It does not contain information about author affiliations.
...

Image 5: This slide shows the results of the research, including theoretical and empirical findings... It does not contain information about author affiliations.

Granularity and Specificity per Image
Focus on Question-Relevant Details
Structural Comparison Clarity

Detail + Individual

Slide 1: Title Slide
...The bottom shows "DTU Compute, Department of Applied Mathematics and Computer Science," and the DTU logo is in the upper right corner. A collection of mathematical symbols is scattered across the bottom right, seemingly decorative.

Slide 2: Motivation
...It highlights the challenges of evaluating and comparing models with larger datasets and more complex models. The key concept introduced ...

Slide 3: Leave-One-Out Cross-Validation
This slide explains the leave-one-out cross-validation (LOO) method. It describes the basic idea of ...

Slide 4: Our Contributions: Method
This slide details the authors' contributions to addressing the problems of LOO. They propose a fast approximation for
...

Concise + Individual

Image 1: A presentation slide shows the title
...

Image 2: A presentation slide details the motivation for model selection in large datasets...

Image 3: A presentation slide explains leave-one-out cross-validation, highlighting its desirable properties and the two major computational problems associated with it for large datasets.

Image 4: A presentation slide presents a proposed method to address...

Image 5: A presentation slide summarizes the theoretical and empirical results of the proposed method, including graphs showing its performance on different models, and invites the audience to a poster session for more details.

Concise + Summarized

The images present a research presentation on a fast approximation method for Bayesian leave-one-out cross-validation applied to large datasets, detailing the motivation, methodology, theoretical and empirical results, and concluding with an invitation to view a poster for further details.

Figure 1: An example multi-image question with different captioning settings. Text in red, green, and orange highlights our advantages. Text in blue is the correct answer. The actual prompt used for each captioning setting can be found in Appendix B.

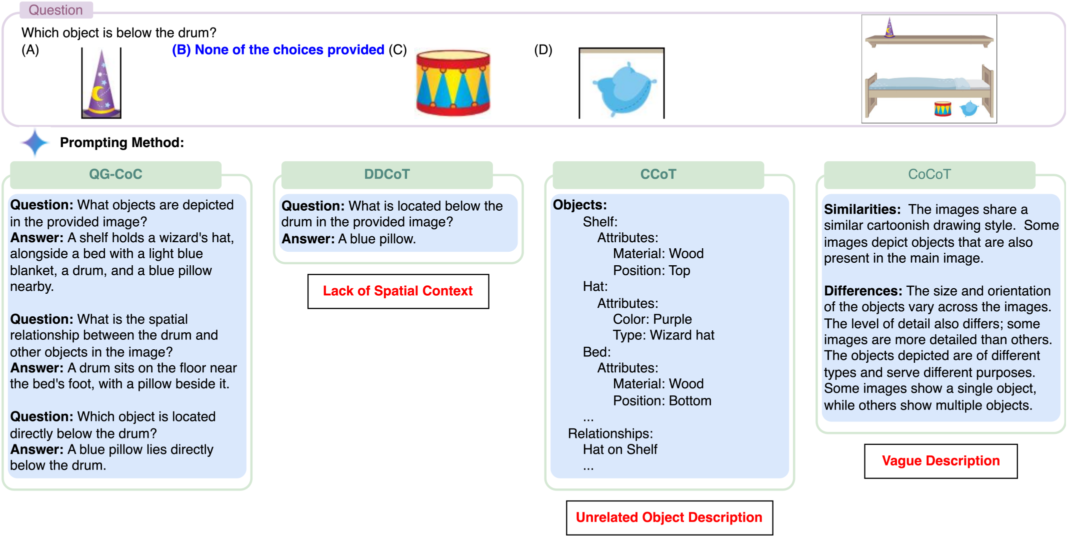
Model	**Gemini-Flash**		**LLaVA-OV**		**Mantis**	
Dataset	MMIU	MUIR	MMIU	MUIR	MMIU	MUIR
Concise vs. Detailed	54.1 → **54.9**	65.2 → **66.3**	47.3 → **48.0**	43.7 → **44.0**	45.3 → **46.4**	42.3 → **44.5**
Summarize vs. Individual	54.1 → **54.5**	66.0 → **66.5**	46.5 → **48.6**	**44.1** → 43.9	45.3 → **46.4**	43.1 → **43.5**
Question-Guided (N/Y)	53.3 → **55.3**	65.4 → **66.2**	47.4 → **47.8**	43.1 → **44.7**	45.5 → **46.0**	42.4 → **44.1**

Table 1: Comparison of captioning settings across models and multi-image datasets. Metrics represent answer accuracy (%).

strategies and derive insights into their effectiveness, focusing on four key settings: (1) concise versus detailed captions, (2) individual captions for each image versus a summarized caption across multiple images, and (3) the inclusion of questions when doing captioning. To comprehensively assess performance, we evaluate both closed-source and open-source models across all possible combinations of these factors, resulting in 8 experimental settings. For each control factor, results are averaged over the 4 relevant variations, enabling a fair and robust comparison of the different strategies.

1. **Caption Length (Concise vs. Detailed)**: To examine whether the level of detail in image captions affects multi-image understanding, we compare two captioning length settings: Concise (describe the image in a sentence) vs. Detailed (describe the image in detail). Table 1 indicates that detailed captions improve multi-image accuracy due to enhanced modality matching and comprehensive image descriptions. In Figure 1, we can observe that detailed captioning will contain the information such as author and school list needed for answering the question.
Insight: Detailed captions are superior to concise ones, as they mitigate information loss and better support complex reasoning tasks.

2. **Caption Scope (Summarized vs. Individual)**: When dealing with multiple images related to the question, a key decision is whether to summarize image set as a whole or describe each image independently. We evaluated two settings: Summarized (generate a summarized caption that describes the content across the whole set) vs. Individual (generate a separate caption for each image). Table 1 indicates that when handling multiple images, generating individual captions for each image outperforms producing a single summarized caption across all images. In Figure 1, we can observe that individual captioning provides more information than summarized captioning.
Insight: Individual captions are more effective than summarized captions, particularly in multi-image scenarios requiring precise,


Question
Which object is below the drum?
(A)
(B) None of the choices provided
(C)
(D)
Prompting Method:
QG-CoC
DDCoT
CCoT
CoCoT

Question: What objects are depicted in the provided image?
Answer: A shelf holds a wizard's hat, alongside a bed with a light blue blanket, a drum, and a blue pillow nearby.

Question: What is the spatial relationship between the drum and other objects in the image?
Answer: A drum sits on the floor near the bed's foot, with a pillow beside it.

Question: Which object is located directly below the drum?
Answer: A blue pillow lies directly below the drum.

Question: What is located below the drum in the provided image?
Answer: A blue pillow.

Objects:
Shelf:
Attributes:
Material: Wood
Position: Top
Hat:
Attributes:
Color: Purple
Type: Wizard hat
Bed:
Attributes:
Material: Wood
Position: Bottom
...
Relationships:
Hat on Shelf
...

Similarities: The images share a similar cartoonish drawing style. Some images depict objects that are also present in the main image.

Differences: The size and orientation of the objects vary across the images. The level of detail also differs; some images are more detailed than others. The objects depicted are of different types and serve different purposes. Some images show a single object, while others show multiple objects.

Figure 2: An example multi-image question with different prompting methods. Text in red highlights the disadvantages. Text in blue is the correct answer. The actual prompt used for each method can be found in Appendix C.

image-specific information.

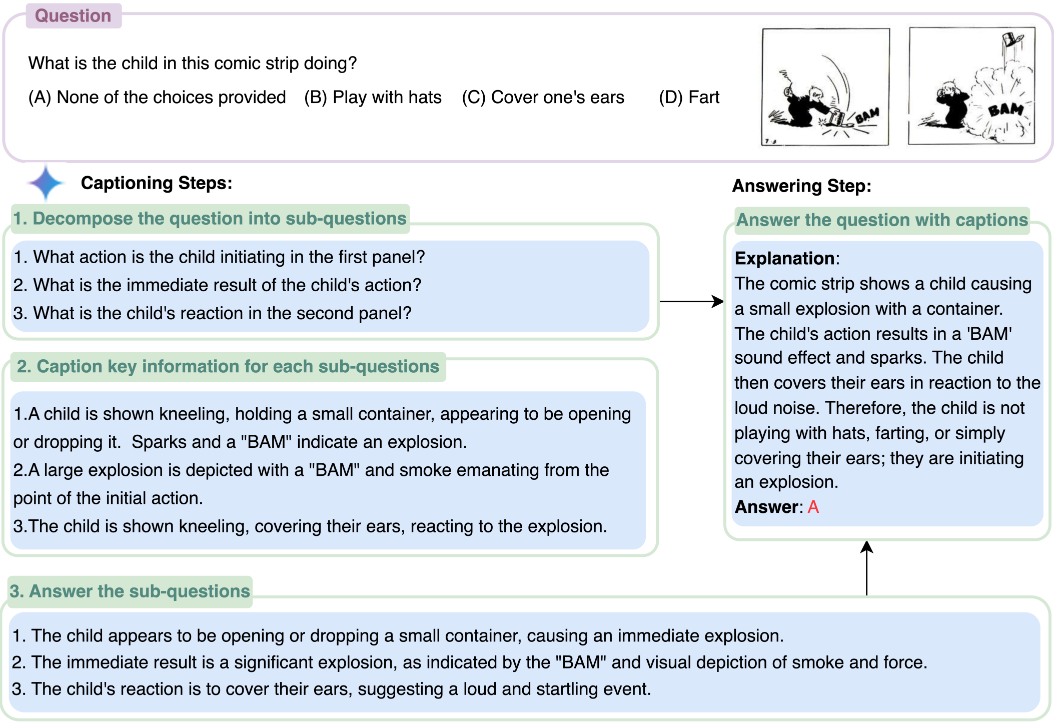
3. **Question-Guided (No vs. Yes)**: To understand whether integrating the question during the caption generation influences the performance, we compare two captioning settings: No Question-Guided (captions are generated based on images solely) vs. Question-Guided (captions are generated based on images and the question). Table 1 and Figure 1 show that question-guided captions improve overall multi-image task accuracy, focusing on task-relevant visual elements.
Insight: Question-guided captioning outperforms unguided captioning by aligning generated context more closely with the question.

Based on the above findings regarding effective image captioning in multi-image scenarios, the next subsection examines if adjusting the previous single-image prompting methods to multi-image scenarios can provide the necessary context for multi-image problems.

3.2 Adjusting Existing Prompting Methods to Multi-Image Scenarios

We conduct the following study to verify whether existing prompting methods can be effectively extended to address the complexities of multi-image scenarios. Our study focused on prominent methods such as DDCoT (Duty-Distinct Chain-of-Thought) (Zheng et al., 2023), which we adapted to decompose a central question into sub-questions applicable across multiple images; CCoT (Compositional Chain-of-Thought) (Mitra et al., 2024), explored for its potential to generate a composite scene graph from each given image; and CoCoT (Contrastive Chain-of-Thought) (Zhang et al., 2024), which, while originally designed for discerning similarities and differences between just two images, we considered for its conceptual applicability to broader multi-image comparisons. As illustrated in Figure 2 using Gemini-1.5-Flash (Team et al., 2024), we present a case study and reveal a consistent pattern. While these adapted existing methods demonstrate some capability in identifying individual entities, their characteristics, and straightforward, explicit relationships between images, they exhibit significant limitations. Specifically, they struggle to extract deeper, implicit context or perform complex reasoning that requires synthesizing information from an arbitrary number of images. For example, DDCoT lacks present spatial context from images, CCoT presents unrelated object descriptions since it does not understand what information is needed to answer the question, and CoCoT only vaguely describes the similarity and difference between images. To further validate these observations, Section 5 provides quantitative support that demonstrates these limitations.

Thus, since the above study highlights the need for more specialized prompting methods tailored to multi-image context, we propose a new zero-shot prompting method **Question-Guided Chain-of-Captions** that involves balancing detail, specificity, and relevance.



Question

What is the child in this comic strip doing?

(A) None of the choices provided (B) Play with hats (C) Cover one's ears (D) Fart

Captioning Steps:

1. Decompose the question into sub-questions

1. What action is the child initiating in the first panel?
2. What is the immediate result of the child's action?
3. What is the child's reaction in the second panel?

2. Caption key information for each sub-questions

1.A child is shown kneeling, holding a small container, appearing to be opening or dropping it. Sparks and a "BAM" indicate an explosion.
2.A large explosion is depicted with a "BAM" and smoke emanating from the point of the initial action.
3.The child is shown kneeling, covering their ears, reacting to the explosion.

3. Answer the sub-questions

1. The child appears to be opening or dropping a small container, causing an immediate explosion.
2. The immediate result is a significant explosion, as indicated by the "BAM" and visual depiction of smoke and force.
3. The child's reaction is to cover their ears, suggesting a loud and startling event.

Answering Step:

Answer the question with captions

Explanation: The comic strip shows a child causing a small explosion with a container. The child's action results in a 'BAM' sound effect and sparks. The child then covers their ears in reaction to the loud noise. Therefore, the child is not playing with hats, farting, or simply covering their ears; they are initiating an explosion.

Answer: A

Figure 3: An example multi-image question and its corresponding reasoning steps using QG-CoC. The prompts used for each step can be found Appendix D.

4 Question-Guided Chain-of-Captions

As shown in Figure 3, Question-Guided Chain-of-Captions (QG-CoC) is a structured reasoning approach designed to enhance multi-image understanding. The method involves three key steps:

Step 1: Decompose the question into sub-questions. First, given a complex question, the method breaks it down into a series of simpler, interpretable sub-questions. Each sub-question targets a specific aspect of the image(s), such as the subject's action, outcome, or reaction. This decomposition ensures that the reasoning is detailed and aligned with the intent of the question.

Step 2: Caption key information for each sub-question. The MLLM then generates targeted captions for each sub-question. These captions extract and describe the most relevant visual evidence (e.g., objects, actions, effects, or scene changes), providing intermediate interpretations. This step directly connects each piece of reasoning to the image content.

Step 3: Answer the sub-questions and integrate reasoning. Finally, the model answers each sub-question based on the captions, forming a coherent reasoning chain. These individual answers are then combined to produce the final answer to the original question, supported by visual evidence from the images. This step-by-step process improves both the accuracy and the explainability of the model predictions.

5 Experimental Results

5.1 Experimental Setting

Implementation. We conduct experiments using different zero-shot prompting methods on both closed-source and open-source MLLMs. For experiments in this section, we utilize GPT-4o (Hurst et al., 2024) and Gemini-1.5-Flash (Team et al., 2024) as representatives of general-purpose MLLMs. We also utilize two open-sourced MLLMs: Mantis-idefics2-8B (Jiang et al., 2024), LLaVA-OneVision-7B (Li et al., 2024), and Qwen-2.5-VL-7B (Bai et al., 2025), which support multiple image inputs. However, they have limited capacity to process and follow long prompts to generate additional context in the first stage. From open-source model evaluation, we use Gemini-1.5-Flash as oracle captioning in the first stage. The versions of these models we used for the experiments are listed in Appendix A.

Baselines. First, to evaluate the added benefit of our method to pretrained MLLMs, our default baseline is to apply the model to the benchmark without any prompt engineering. Then, we compare **QG-CoC** prompting to five state-of-the-art

Model	Method	Dataset				
		MUIR	MMIU	ScienceQA	MMMU	MMBench
			Open-Source			
LLaVA-One-Vision	w/o prompt	41.2	44.6	**94.5**	45.4	85.1
	Caption	42.0 (+0.8)	48.1 (+3.5)	91.7 (-2.8)	**49.7** (+4.3)	85.1 (+0.0)
	QG-Caption	44.7 (+3.5)	49.4 (+4.8)	93.1 (-1.4)	45.4 (+0.0)	85.6 (+0.5)
	DDCoT	**53.4** (+12.2)	50.5 (+5.9)	92.9 (-1.6)	**49.7** (+4.3)	84.3 (-0.8)
	CCoT	44.6 (+3.4)	46.9 (+2.3)	93.0 (-1.5)	46.8 (+1.4)	86.0 (+0.9)
	CoCoT	44.2 (+3.0)	46.4 (+1.8)	–	–	–
	QG-CoC	53.3 (+12.1)	**50.9** (+6.3)	**94.5** (+0.0)	48.9 (+3.5)	**87.6** (+2.5)
Mantis-idefics2	w/o prompt	43.4	45.0	80.3	41.8	79.0
	Caption	43.9 (+0.5)	46.7 (+1.7)	79.7 (-0.6)	44.7 (+2.9)	80.4 (+1.4)
	QG-Caption	44.5 (+1.1)	47.7 (+2.7)	79.1 (-1.2)	44.0 (+2.2)	79.7 (+0.7)
	DDCoT	47.9 (+4.5)	**50.1** (+5.1)	83.0 (+2.7)	**49.7** (+7.9)	78.3 (-0.7)
	CCoT	44.4 (+1.0)	44.9 (-0.1)	80.7 (+0.4)	46.1 (+4.3)	82.1 (+3.1)
	CoCoT	42.6 (-0.8)	45.4 (+0.4)	–	–	–
	QG-CoC	**48.9** (+5.5)	49.8 (+4.8)	**83.8** (+3.5)	48.9 (+7.1)	**83.4** (+4.4)
Qwen-2.5-VL	w/o prompt	62.1	50.3	90.2	58.2	88.2
	Caption	62.8 (+0.7)	50.9 (+0.6)	88.0 (-2.2)	59.4 (+1.2)	88.3 (+0.1)
	QG-Caption	62.4 (+0.3)	50.1 (-0.2)	88.9 (-1.3)	60.0 (+1.8)	88.5 (+0.3)
	DDCoT	63.7 (+1.6)	54.1 (+3.8)	90.5 (+0.3)	61.5 (+3.3)	87.9 (-0.3)
	CCoT	62.3 (+0.2)	51.6 (+1.3)	89.5 (-0.7)	59.5 (+1.3)	88.5 (+0.3)
	CoCoT	62.6 (+0.5)	52.3 (+2.0)	–	–	–
	QG-CoC	**65.3** (+3.2)	**56.9** (+6.6)	**91.9** (+1.7)	**64.8** (+6.6)	**89.4** (+1.2)
			Closed-Source			
GPT-4o	w/o prompt	70.8	63.3	89.5	63.1	86.0
	Caption	71.8 (+1.0)	63.6 (+0.3)	86.8 (-2.7)	66.0 (+2.9)	88.1 (+2.1)
	QG-Caption	70.0 (-0.8)	65.1 (+1.8)	89.6 (+0.1)	61.7 (-1.4)	**89.5** (+3.5)
	DDCoT	73.1 (+2.3)	62.9 (-0.4)	89.3 (-0.2)	64.5 (+1.4)	86.6 (+0.6)
	CCoT	70.4 (-0.4)	60.9 (-2.4)	87.8 (-1.7)	61.0 (-2.1)	88.1 (+2.1)
	CoCoT	74.0 (+3.2)	64.5 (+1.2)	–	–	–
	QG-CoC	**74.9** (+4.1)	**65.8** (+2.5)	**90.3** (+0.8)	**66.7** (+3.6)	88.9 (+2.9)
Gemini-1.5-Flash	w/o prompt	66.0	55.0	87.0	64.5	**86.0**
	Caption	66.8 (+0.8)	53.7 (-1.3)	86.9 (-0.1)	61.0 (-3.5)	84.5 (-1.5)
	QG-Caption	66.0 (+0.0)	54.9 (-0.1)	86.8 (-0.2)	**66.7** (+2.2)	84.9 (-1.1)
	DDCoT	67.6 (+1.6)	51.5 (-3.5)	86.9 (-0.1)	53.9 (-10.6)	84.5 (-1.5)
	CCoT	66.3 (+0.3)	51.9 (-3.1)	85.5 (-1.5)	53.2 (-11.3)	85.6 (-0.4)
	CoCoT	65.4 (-0.6)	**55.5** (+0.5)	–	–	–
	QG-CoC	**68.2** (+2.2)	55.4 (+0.4)	**87.2** (+0.2)	63.7 (-0.8)	85.2 (-0.8)

Table 2: Multi-Image and Single-Image benchmark performance of different models with various prompting methods. Numbers in (+/-) indicate delta compared to the w/o prompt baseline of the same model. Metrics represent answer accuracy (%).

methods including: (1) **Detailed Captioning**: In the previous section, we find that captioning image individually in detail enhance the performance the most, (2) **Question-Guided Detailed Captioning**: In the previous section, we find that adding question in the prompt enhances the performance, (3) **DDCoT** (Zheng et al., 2023): First, decompose the question, then utilizes MLLMs to answer the sub-questions and uses it as rationale, (4) **CCoT** (Mitra et al., 2024): Utilize MLLMs to generate a scene graph based on each image, and (5) **CoCoT** (Zhang et al., 2024): Utilize MLLMs to describe the similarity and difference between multiple images. All these methods work in a two-step pipeline. The first step generates an additional textual representation from the instructions of different methods. The second step involves passing the images, question, and output from the first step to answer the question.

Evaluation Dataset. We select two representative and multi-faceted benchmarks: Muir-

Bench (Wang et al., 2024) and MMIU (Meng et al., 2024). MuirBench is a comprehensive benchmark consisting of 12 diverse multi-image tasks, such as scene understanding, ordering, etc. It contains 2,600 multiple-choice questions with 11,264 images in total. We report the overall average performance across the 12 tasks. MMIU is a multi-image benchmark encompassing 7 types of multi-image relationships, 52 tasks, 77K images, and 11K multiple-choice questions. We report the overall average performance across all the tasks. However, during the evaluation, we observe some tasks in MMIU exhibit low quality, so we filter out some tasks in the spatial and semantic relationships. We also compare our method on various single-image tasks, including MMMU (Yue et al., 2024), MMBench (Liu et al., 2024), and ScienceQA (Lu et al., 2022), to validate the generalizability of our method. However, since CoCoT is constructed under image comparison, we cannot evaluate CoCoT on single-image benchmarks.

5.2 Main Results

To investigate which prompting methods and models better solve multi-image problems, we summarize the answer accuracy performance in Table 2.

Comparison with various prompting baselines. QG-CoC demonstrates strong performance across both multi-image and single-image benchmarks, as shown in Table 2:

1. **Comparison over Caption:** While providing detailed captions for individual images ("Caption" method) is beneficial, QG-CoC not only provides image captions but also ensures these captions are directly relevant to specific parts of the sub-question. This relevance is achieved by first decomposing the main question into sub-questions (*Step 1*) and captioning key information for sub-questions (*Step 2*). As a result, the generated captions are targeted, leading to more focused and effective reasoning compared to general detailed captions.

2. **Comparison over QG-Caption:** QG-Caption incorporates the question into the prompt to improve caption relevance. Instead of guiding captions with a single, potentially complex main question, QG-CoC decomposes the question into simpler sub-questions (*Step 1*) and then generates targeted captions for each sub-question (*Step 2*). This question-guided captioning at each sub-question typically yields better results than a single pass of QG-Caption.

3. **Comparison over DDCoT:** DDCoT also involves question decomposition. However, QG-CoC introduces a crucial intermediate step: generating explicit, targeted captions for each sub-question (*Step 2*) before proceeding to answer them and integrate reasoning (*Step 3*). This step of grounding each sub-problem in visual evidence through dedicated captions often leads to more robust reasoning. While DDCoT shows competitive performance, QG-CoC frequently outperforms it.

4. **Comparison over CCoT:** While scene graphs can be informative, they might produce overly detailed or less relevant information for a specific question. Our method of generating captions related to sub-questions (*Step 2*), guided by the initial question decomposition (*Step 1*), ensures that the visual information extracted is directly relevant to the task. Thus, QG-CoC consistently demonstrates higher accuracy than CCoT.

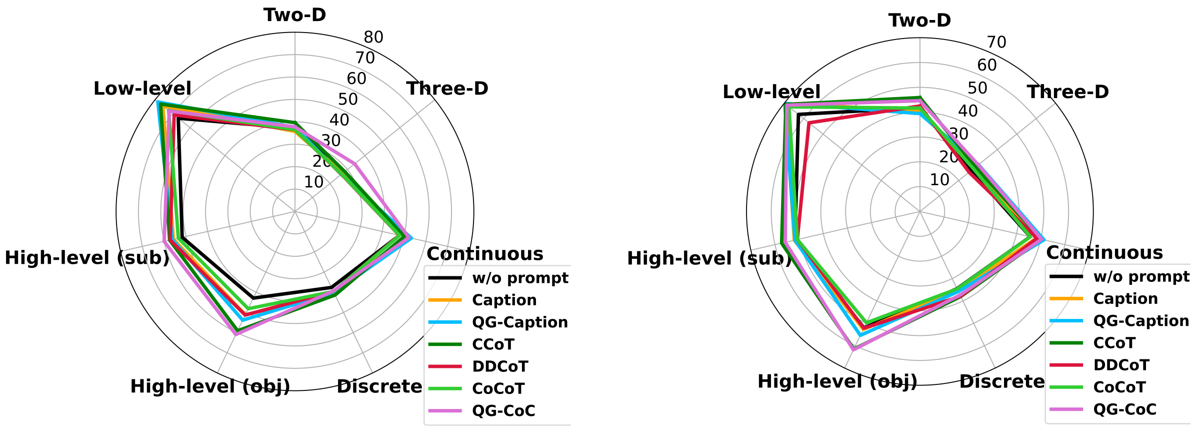
5. **Comparison over CoCoT:** CoCoT utilizes MLLMs to describe the similarity and difference between multiple images. This can be effective for comparative tasks but may not be optimal for all types of multi-image tasks. QG-CoC, through its sub-question decomposition (*Step 1*) and subsequent targeted captioning (*Step 2*), offers a more general framework that can adapt to various reasoning needs beyond simple comparison. As a result, QG-CoC generally achieves higher accuracy than CoCoT.

Overall, the results show the effectiveness of QG-CoC in leveraging both detailed image understanding and question-aware reasoning.

6 Discussion

We conduct an analysis of QG-CoC through multiple perspectives, including detailed breakdowns of different visual domains on MMIU and MUIR benchmarks, the impact of incorporating each component of QG-CoC, and common error analysis.

Different Prompting Methods Performance Across Various Image Relationships. As shown



(a) LLaVA-OV

(b) Mantis

Figure 4: Prompting methods performance by image relationships on different models (MMIU dataset).

in Figure 4, models exhibit different capabilities across various image relationships in MMIU. We also record all model performance on all tasks in MMIU (Table 8) and MUIR (Table 9).

1) In semantic relationships, direct prompting generally performs better on multi-image semantic tasks involving low-level relationships than adding more context. Since low-level relationships usually involve intuitive understanding, providing more details will not help with reasoning. Inversely, in high-level tasks, for subjective tasks such as Causality Reasoning and Emotion Recognition, which require the identification and reasoning of implicit visual information, and objective tasks, such as retrieval tasks, QG-CoC outperforms existing methods significantly since our method provides more key information to tackle them. 2) In temporal relationships, all prompting methods can handle discrete and continuous temporal relationships relatively well, but perform poorly on reasoning-intensive tasks such as Visual Ordering and Temporal Ordering. 3) In spatial relationships, we find that all prompting method struggles with understanding both 2D and 3D positional relations. Since these prompting methods cannot provide spatial information in multiple images and reason correctly, QG-CoC overall provides more spatial-related information compared to other methods.

Importance of each component on QG-CoC. We analyze the contribution of each component in QG-CoC through an ablation study on the MUIR and MMIU benchmarks. In Table 3, starting from the zero-shot baseline, each successive module leads to consistent performance gains. Introducing Question Decomposition improves MUIR accuracy from 66.0 to 66.5, showing the benefit of simplifying complex queries. Adding the Question-Guided Captioning module further raises the score to 67.2, highlighting the importance of context-aware visual grounding. Finally, incorporating the full QG-CoC model achieves the highest accuracy of 68.2 on MUIR and 55.4 on MMIU, confirming that the combined reasoning and generation steps effectively enhance overall understanding. These results underscore the complementary roles of each module and validate the design of our compositional reasoning pipeline.

Method	MUIR	MMIU
Zero-shot	66.0	55.0
+ Question-Decompose	66.5	54.8
+ Question-Guided Caption	67.2	55.1
+ QG-CoC	**68.2**	**55.4**

Table 3: Ablation experiment results across MMIU and MUIR benchmarks using Gemini-1.5-Flash. Our method achieves the highest accuracy among all.

Error Reason	Percentage (%)
(E1) Wrong question understanding	33.3% (40/120)
(E2) Inaccurate perception	31.7% (38/120)
(E3) Wrong reasoning	35.0% (42/120)

Table 4: Statistics of error analysis under Gemini-1.5-Flash using QG-CoC.

Error Analysis. We delve deeper into the primary challenges that MLLMs encounter when solving multi-image problems using QG-CoC. To gain a quantitative understanding of model failures, we

Error Type	Geographic	Diagram	Matching	Difference	Retrieval	Counting	Attribute	Scene	Action	Grounding	Cartoon	Ordering
E1	50	30	30	30	30	40	30	30	30	30	30	40
E2	30	40	40	20	40	40	40	20	20	30	30	30
E3	20	30	30	50	30	20	30	50	50	40	40	30

Table 5: Distribution of error types (%) across MUIR tasks for Gemini-1.5-Flash under QG-CoC prompting.

randomly sample 10 error instances for every task and a total of 120 error instances made by Gemini-1.5-flash on MuirBench, and annotate the main reasons for these mispredictions. We categorize into the three error types, including: **(E1) Wrong question understanding**, which means MLLMs do not understand the question accurately, leading to the incorrect question decomposition. **(E2) Wrong perception**, which means the failure to capture details in or between images. **(E3) Wrong reasoning**, which means even if we get accurate decomposition and captioning, MLLMs still infer the wrong reasoning path to answer the question.

In Table 4, we observe that the most common error category (35.0% of error cases) is failure of reasoning. We conclude that even if the given context is accurate, MLLMs still infer incorrectly. The other error category (33.3% of error cases) is due to inaccurate question understanding and influences the generation of incorrect captions and reasoning. The rest 31.7% of errors are due to the failure to capture details in images. The detailed qualitative examples are provided in Figure 10.

We further analyze errors by task category in MUIR (Table 5). We observe that tasks requiring holistic multi-image understanding (e.g., Difference, Scene, Action) are dominated by E3. In contrast, tasks relying on fine-grained perception (e.g., Matching, Attribute, Counting) are more prone to E2. Meanwhile, E1 is consistently present, with higher prevalence in abstract tasks like Ordering and Geographic. Overall, the breakdown confirms that reasoning across multiple images remains the most significant challenge.

Method	#Tokens	Runtime
w/o prompt	0	3.5s
Caption	349	8.5s
QG-Caption	169	6.6s
DDCoT	108	5.8s
CCoT	372	8.7s
CoCoT	111	5.9s
QG-CoC	**127**	**6.1s**

Table 6: Computational Overhead Analysis on MMIU Benchmark. Runtime means the average runtime(seconds) per sample. #Tokens means the average additional tokens per sample.

Inference Time Comparison Analysis. We analyze the computational overhead of our method, QG-CoC. The method involves a two-stage pipeline, which inherently introduces additional costs compared to direct prompting. To quantify this, we measured the extra token usage for closed-source models, using Gemini-1.5-Flash as an example, and the inference runtime for open-source models, exemplified by LLaVA-OneVision-7B. The results, averaged on 100 data samples randomly selected from the MMIU benchmark and run on 4 NVIDIA A6000 GPUs for open-source models, are detailed in Table 6. For Gemini-1.5-Flash, token estimation was based on the Google-provided API. As the table indicates, QG-CoC does increase token usage and runtime. However, we contend that this is a justifiable trade-off for the consistent performance improvements documented in our paper. This is particularly evident for open-source models, where QG-CoC leads to more significant gains, with a +12% improvement for LLaVA-OV and +5% for Mantis. The overhead is comparable to other two-stage methods while achieving superior accuracy. We believe this represents an efficient utilization of resources to unlock more advanced reasoning capabilities.

7 Conclusion

In this work, we introduce a novel prompting method called Question-Guided Chain-of-Captions (QG-CoC), which first incorporates problem decomposition and then generates each sub-question-guided image captioning to provide a clue to answer the sub-question, then combines the sub-question and sub-answer pair as prior knowledge to answer the original problem. Our extensive experiments demonstrate the advantages of our method for different MLLMs on various benchmarks.

Limitations

This work only provides a strong baseline for the single-image and multi-image reasoning of MLLMs. Although we experiment with many representative models and reasoning methods in this paper, we acknowledge that this does not cover all models and frameworks. Our proposed method relies on the captioning ability of advanced MLLMs. Therefore, it might cause performance deterioration in less advanced language models or more challenging tasks. To strengthen QG-CoC, a more diverse and complicated scenario should be explored in the future, such as complex geometric shapes and even 2D, 3D-spatial information.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024a. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024b. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, et al. 2024. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *arXiv preprint arXiv:2408.02718*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kamcot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

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

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

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Yiwu Zhong, Zi-Yuan Hu, Michael R Lyu, and Liwei Wang. 2024. Beyond embeddings: The promise of visual table in visual reasoning. *arXiv preprint arXiv:2403.18252*.

A Model Hyperparameters

The hyperparameters for the experiments for studying QG-CoC and other prompting methods are set to their default values to ensure consistency in our experiment. Table 7 details the specific generation parameters for the various MLLMs we evaluate.

B Detail Studies of Different Captioning Strategies under Multi-Image

B.1 Full Model Prompt

In Figure 5, we show the full model prompt of different captioning settings.

C Detail Studies of Adjusting Existing Prompting Methods to Multi-Image Scenarios

C.1 Full Model Prompt

In Figure 6, we show the full model prompt of different methods.

D Detail Studies of Question-Guided Chain-of-Captions

D.1 Full Model Prompt

In Figure 7, we show the full model prompt of QG-CoC.

D.2 Full Quantitative Results Across Various Image Relationships

We further show the overall performance of QG-CoC across various image relationships and compare it with different prompting methods and models. The results of MMIU and MUIR datasets are shown in Table 8 and Table 9, and we also illustrate the task performance of different prompting methods under MUIR benchmark in Figure 8. The findings remain the same as MMIU, and our method outperforms other methods. Additionally, we observe that the performance of each task under open-source models generally has a larger difference compared to closed-source models across various datasets and prompting methods.

D.3 More Qualitative Examples

In Figure 9, we show more examples for each multi-image task using QG-CoC in Gemini-1.5-Flash.

D.4 Qualitative Analysis of Error Cases

We present every type of error case that Gemini-1.5-Flash cannot answer correctly in Figure 10a, 10b,10c. From E1, the model understands the wrong meaning of the question that "tortoise" is not "duck", and decomposes the question into wrong sub-questions (sub-goals). From E2, in step 2, the model incorrectly captions that "L shape has 4 squares", when the correct caption is "3 squares". From E3, since the generated sub-questions and captions are accurate, we can observe that the model correctly points out the difference between the two images, "a person walking". However, the model does incorrect reasoning in the final response.

Model	Version	Generation Setup
Close-source		
GPT-4o	gpt-4o-2024-05-13	temperature = 0, max tokens = 2048
Gemini-Flash	gemini-1.5-flash	temperature = 0, max tokens = 2048
Open-source		
LLaVA-OneVision-7B	lmms-lab/llava-onevision-qwen2-7b-ov	do_sample=False, temperature=0, max tokens = 2048
Mantis-Idefics2-8B	TIGER-Lab/Mantis-8B-Idefics2	do_sample=False, temperature=0, max tokens = 2048

Table 7: Model names, versions, and generating setups for various MLLMs.

Model	Method	Overall	Discrete	Continuous	Low-level	High-sub	High-obj	Two-D	Three-D
LLaVA-OV	w/o prompt	44.6	37.6	47.9	66.8	51.8	42.9	37.1	27.8
	Caption	48.1	40.5	50.6	75.6	55.8	51.2	35.8	27.5
	QG-Caption	49.4	40.1	53.4	78.4	56.3	53.8	37.6	26.5
	CCoT	50.5	41.4	50.2	76.9	57.5	59.1	39.6	28.5
	DDCoT	46.9	39.6	47.8	69.1	57.3	51.3	36.4	26.6
	CoCoT	46.4	39.6	48.0	72.3	53.5	48.2	36.5	26.8
	QG-CoC	50.9	39.4	52.3	71.9	60.0	61.0	37.8	34.1
Mantis	w/o prompt	45.0	34.5	45.7	62.7	51.8	52.0	41.8	26.4
	Caption	46.7	35.4	45.7	69.5	52.0	52.7	40.7	28.6
	QG-Caption	47.7	35.8	51.4	69.8	51.8	55.4	39.4	30.3
	CCoT	50.1	38.0	50.3	69.2	57.3	61.5	45.9	28.8
	DDCoT	44.9	37.9	48.5	57.3	50.8	52.2	42.5	25.4
	CoCoT	45.4	34.6	45.7	67.6	50.8	49.8	41.6	27.6
	QG-CoC	49.8	37.4	50.4	68.7	55.8	61.9	44.6	30.1
GPT-4o	w/o prompt	63.3	60.6	60.7	94.8	60.0	67.3	53.3	46.4
	Caption	63.6	59.0	57.5	95.1	65.8	65.9	53.3	48.6
	QG-Caption	65.1	58.1	61.4	93.1	66.0	67.7	55.8	53.5
	CCoT	60.9	53.4	60.0	91.7	60.8	63.7	53.4	43.0
	DDCoT	62.9	57.3	58.3	94.1	64.0	65.1	54.4	47.0
	CoCoT	64.5	60.3	60.9	95.4	65.8	65.0	56.3	48.0
	QG-CoC	65.8	59.3	61.4	93.3	66.0	68.5	56.2	55.9
Gemini-Flash	w/o prompt	55.0	49.4	53.0	82.1	62.0	61.3	46.4	30.9
	Caption	53.7	51.4	52.1	83.1	60.3	63.3	47.2	18.4
	QG-Caption	54.9	52.8	55.1	78.3	59.5	63.0	47.5	28.1
	CCoT	51.9	48.1	52.3	72.2	59.8	60.9	45.6	24.5
	DDCoT	51.5	47.8	51.6	80.4	58.8	61.4	42.4	18.4
	CoCoT	55.5	50.8	52.3	79.6	59.8	63.2	49.1	33.8
	QG-CoC	55.4	51.1	54.6	76.8	60.3	63.4	48.1	33.6

Table 8: MMIU performance across dimensions with different prompting methods and models.

Model	Method	Overall	Geographic.	Diagram.	Matching.	Difference.	Retrieval.	Counting.	Attribute.	Scene.	Action.	Grounding.	Cartoon.	Ordering
LLaVA-OV	w/o prompt	41.2	37.0	54.0	44.0	30.0	45.9	26.5	34.2	63.4	40.2	29.8	38.5	15.6
	Caption	42.0	46.0	56.0	44.0	32.4	38.4	34.2	28.6	66.7	42.1	32.1	37.2	20.3
	QG-Caption	44.7	40.0	60.1	49.6	33.2	41.4	36.3	37.2	66.1	43.3	29.8	38.5	20.3
	CCoT	44.6	44.0	58.8	47.8	32.7	43.5	35.9	36.7	69.9	40.2	32.1	38.5	18.8
	DDCoT	53.4	41.0	69.6	61.0	46.2	54.5	34.2	56.1	74.2	42.1	32.1	41.0	21.9
	CoCoT	44.2	42.0	56.8	46.3	34.4	50.3	31.6	35.7	67.2	42.1	31.0	35.9	17.2
	QG-CoC	53.3	42.0	70.1	60.1	38.8	54.1	41.9	56.6	76.9	43.9	29.8	42.3	20.3
Mantis	w/o prompt	43.4	25.0	62.1	53.7	28.8	35.3	38.0	46.9	56.5	34.2	28.6	38.5	17.2
	Caption	43.9	29.0	61.3	53.0	32.7	31.9	39.3	33.7	62.9	44.5	28.6	43.6	17.2
	QG-Caption	44.5	32.0	63.6	53.5	28.5	37.0	41.0	38.8	62.4	41.5	28.6	38.5	15.6
	CCoT	44.4	30.0	63.3	56.5	28.2	34.6	41.5	35.7	66.1	37.8	27.4	38.5	10.9
	DDCoT	47.9	35.0	59.8	57.8	35.9	42.1	39.3	52.0	71.0	38.4	34.5	41.0	15.6
	CoCoT	42.6	26.0	59.6	52.6	33.8	31.5	39.3	35.2	55.9	38.4	29.8	38.5	17.2
	QG-CoC	48.9	37.0	64.3	59.1	34.5	41.4	44.0	48.0	70.4	39.0	32.1	46.2	15.6
GPT-4o	w/o prompt	70.8	50.0	90.2	84.1	58.5	63.0	78.6	63.3	86.6	50.6	54.8	53.9	28.1
	Caption	71.8	62.0	91.0	85.6	65.3	59.9	79.1	56.1	83.3	54.9	53.6	52.6	34.4
	QG-Caption	67.0	44.0	90.2	84.9	63.8	58.2	75.2	60.7	85.0	51.2	52.4	50.0	23.4
	CCoT	70.4	51.0	90.2	83.9	66.2	61.6	75.6	60.2	83.3	46.3	54.8	44.9	31.3
	DDCoT	73.1	50.0	89.7	85.8	66.5	64.4	79.9	61.7	87.6	57.3	56.0	56.4	40.6
	CoCoT	74.0	57.0	90.5	87.3	70.6	70.9	76.5	59.2	88.2	50.0	54.8	57.7	37.5
	QG-CoC	74.9	61.0	91.0	87.9	68.5	68.5	79.1	62.2	87.0	57.9	57.1	56.4	43.8
Gemini-Flash	w/o prompt	66.0	53.0	84.7	82.5	53.5	75.3	51.3	54.1	82.8	43.3	51.2	46.2	18.8
	Caption	66.9	58.0	84.2	83.2	56.2	69.2	50.9	58.2	80.7	47.6	50.0	50.0	32.8
	QG-Caption	66.0	47.0	83.4	83.4	55.0	64.4	52.1	61.2	83.3	53.1	48.8	42.3	25.0
	CCoT	66.3	54.0	85.7	82.3	52.4	69.9	50.0	60.7	81.2	49.4	47.6	43.6	34.4
	DDCoT	67.6	44.0	87.7	84.3	56.5	74.7	46.6	62.2	75.8	49.4	56.0	53.9	32.8
	CoCoT	65.4	44.0	84.4	81.7	50.9	73.3	48.7	57.1	80.7	47.0	51.2	52.6	25.0
	QG-CoC	68.2	46.0	88.7	84.3	57.4	76.0	50.4	59.2	79.0	50.6	52.4	51.3	28.1

Table 9: MUIR performance across tasks with different prompting methods and models.

Caption Prompt Template

Caption Length (Concise vs. Detailed)

1. Describe each given image individually **in one sentence**. {Image Set}
2. Describe each given image as a whole **in one sentence**. {Image Set}
3. Given the multi-image question, generate only a caption highlighting the key information related to the question **in one sentence**. {Question} {Image Set}
4. Given the multi-image question, generate a question-relevant image caption for each image individually **in one sentence**. {Question} {Image Set}

1. Describe each given image individually **in detail**. {Image Set}
2. Describe each given image as a whole **in detail**. {Image Set}
3. Given the multi-image question, generate a question-relevant image caption for each image individually **in detail**. {Question} {Image Set}
4. Given the multi-image question, generate only a caption highlighting the key information related to the question **in detail**. {Question} {Image Set}

Caption Scope (Individual vs. Summarized)

1. Describe each given image **individually** in one sentence. {Image Set}
2. Describe each given image **individually** in detail. {Image Set}
3. Given the multi-image question, generate a question-relevant image caption for each image **individually** in one sentence. {Question} {Image Set}
4. Given the multi-image question, generate a question-relevant image caption for each image **individually** in detail. {Question} {Image Set}

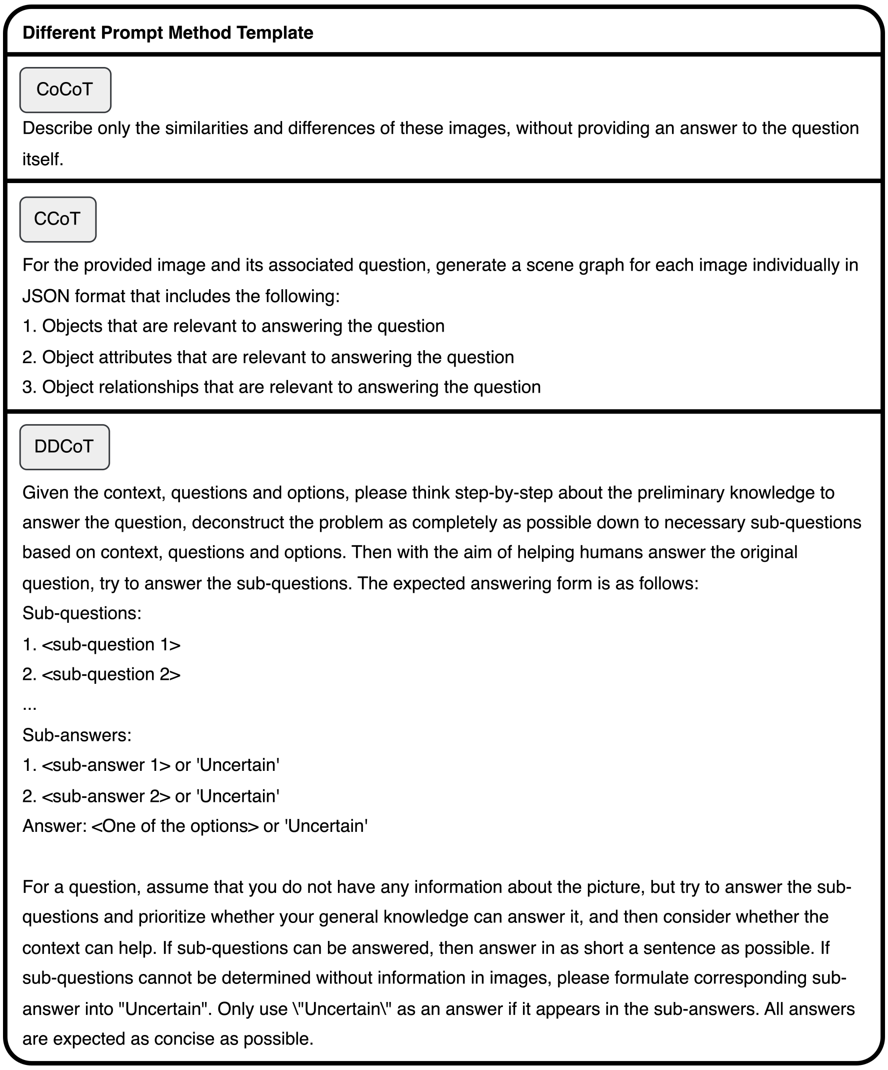
1. Describe the given images as **a summarized caption** in one sentence. {Image Set}
2. Describe the given images as **a summarized caption** in detail. {Image Set}
3. Given the multi-image question, generate only **a summarized caption** highlighting the key information related to the question in one sentence. {Question} {Image Set}
4. Given the multi-image question, generate only **a summarized caption** highlighting the key information related to the question in detail. {Question} {Image Set}

Question-Guided (No vs. Yes)

1. Describe each given image individually in one sentence. {Image Set}
2. Describe each given image individually in detail. {Image Set}
3. Describe the given images as a whole in one sentence. {Image Set}
4. Describe the given images as a whole in detail. {Image Set}

1. **Given the multi-image question**, generate a question-relevant image caption for each image individually in one sentence {Question} {Image Set}
2. **Given the multi-image question**, generate a question-relevant image caption for each image individually in detail. {Question} {Image Set}
3. **Given the multi-image question**, generate only a caption highlighting the key information related to the question in detail. {Question} {Image Set}
4. **Given the multi-image question**, generate only a caption highlighting the key information related to the question in one sentence. {Question} {Image Set}

Figure 5: Actual prompts with different captioning settings.



Different Prompt Method Template

CoCoT

Describe only the similarities and differences of these images, without providing an answer to the question itself.

CCoT

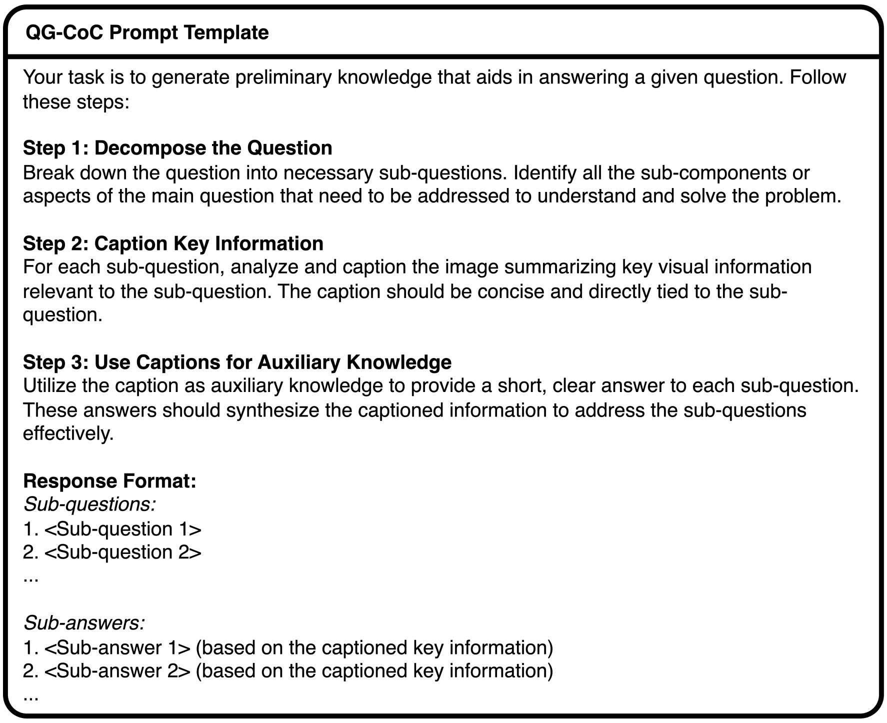
For the provided image and its associated question, generate a scene graph for each image individually in JSON format that includes the following:
1. Objects that are relevant to answering the question
2. Object attributes that are relevant to answering the question
3. Object relationships that are relevant to answering the question

DDCoT

Given the context, questions and options, please think step-by-step about the preliminary knowledge to answer the question, deconstruct the problem as completely as possible down to necessary sub-questions based on context, questions and options. Then with the aim of helping humans answer the original question, try to answer the sub-questions. The expected answering form is as follows:
Sub-questions:
1. <sub-question 1>
2. <sub-question 2>
...
Sub-answers:
1. <sub-answer 1> or 'Uncertain'
2. <sub-answer 2> or 'Uncertain'
Answer: <One of the options> or 'Uncertain'

For a question, assume that you do not have any information about the picture, but try to answer the sub-questions and prioritize whether your general knowledge can answer it, and then consider whether the context can help. If sub-questions can be answered, then answer in as short a sentence as possible. If sub-questions cannot be determined without information in images, please formulate corresponding sub-answer into "Uncertain". Only use \"Uncertain\" as an answer if it appears in the sub-answers. All answers are expected as concise as possible.

Figure 6: Different actual prompts of existing prompting methods adapted to multi-image scenarios.



QG-CoC Prompt Template

Your task is to generate preliminary knowledge that aids in answering a given question. Follow these steps:

Step 1: Decompose the Question
Break down the question into necessary sub-questions. Identify all the sub-components or aspects of the main question that need to be addressed to understand and solve the problem.

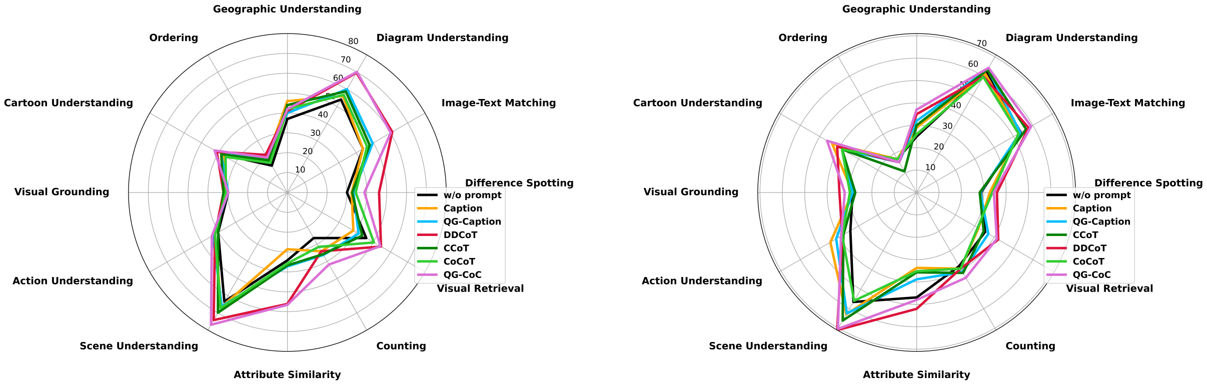
Step 2: Caption Key Information
For each sub-question, analyze and caption the image summarizing key visual information relevant to the sub-question. The caption should be concise and directly tied to the sub-question.

Step 3: Use Captions for Auxiliary Knowledge
Utilize the caption as auxiliary knowledge to provide a short, clear answer to each sub-question. These answers should synthesize the captioned information to address the sub-questions effectively.

Response Format:
Sub-questions:
1. <Sub-question 1>
2. <Sub-question 2>
...

Sub-answers:
1. <Sub-answer 1> (based on the captioned key information)
2. <Sub-answer 2> (based on the captioned key information)
...

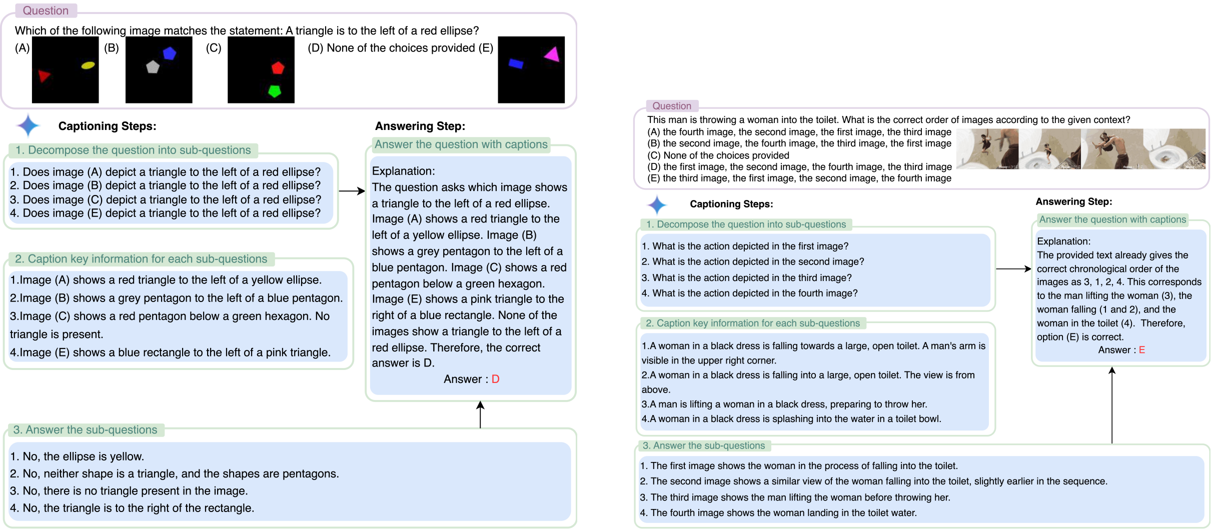
Figure 7: An actual prompt of QG-CoC.



(a) LLaVA-OV

(b) Mantis

Figure 8: Prompting methods performance by tasks on different models. (MUIR)



Question

Which of the following image matches the statement: A triangle is to the left of a red ellipse?

(A) (B) (C) (D) None of the choices provided (E)

Captioning Steps:

1. Decompose the question into sub-questions

1. Does image (A) depict a triangle to the left of a red ellipse?
2. Does image (B) depict a triangle to the left of a red ellipse?
3. Does image (C) depict a triangle to the left of a red ellipse?
4. Does image (E) depict a triangle to the left of a red ellipse?

2. Caption key information for each sub-questions

1.Image (A) shows a red triangle to the left of a yellow ellipse.
2.Image (B) shows a grey pentagon to the left of a blue pentagon.
3.Image (C) shows a red pentagon below a green hexagon. No triangle is present.
4.Image (E) shows a blue rectangle to the left of a pink triangle.

3. Answer the sub-questions

1. No, the ellipse is yellow.
2. No, neither shape is a triangle, and the shapes are pentagons.
3. No, there is no triangle present in the image.
4. No, the triangle is to the right of the rectangle.

Answering Step:

Answer the question with captions

Explanation:
The question asks which image shows a triangle to the left of a red ellipse. Image (A) shows a red triangle to the left of a yellow ellipse. Image (B) shows a grey pentagon to the left of a blue pentagon. Image (C) shows a red pentagon below a green hexagon. Image (E) shows a pink triangle to the right of a blue rectangle. None of the images show a triangle to the left of a red ellipse. Therefore, the correct answer is D.
Answer : D

(a) Task: Image Text Matching

Question

This man is throwing a woman into the toilet. What is the correct order of images according to the given context?
(A) the fourth image, the second image, the first image, the third image
(B) the second image, the fourth image, the third image, the first image
(C) None of the choices provided
(D) the first image, the second image, the fourth image, the third image
(E) the third image, the first image, the second image, the fourth image

Captioning Steps:

1. Decompose the question into sub-questions

1. What is the action depicted in the first image?
2. What is the action depicted in the second image?
3. What is the action depicted in the third image?
4. What is the action depicted in the fourth image?

2. Caption key information for each sub-questions

1.A woman in a black dress is falling towards a large, open toilet. A man's arm is visible in the upper right corner.
2.A woman in a black dress is falling into a large, open toilet. The view is from above.
3.A man is lifting a woman in a black dress, preparing to throw her.
4.A woman in a black dress is splashing into the water in a toilet bowl.

3. Answer the sub-questions

1. The first image shows the woman in the process of falling into the toilet.
2. The second image shows a similar view of the woman falling into the toilet, slightly earlier in the sequence.
3. The third image shows the man lifting the woman before throwing her.
4. The fourth image shows the woman landing in the toilet water.

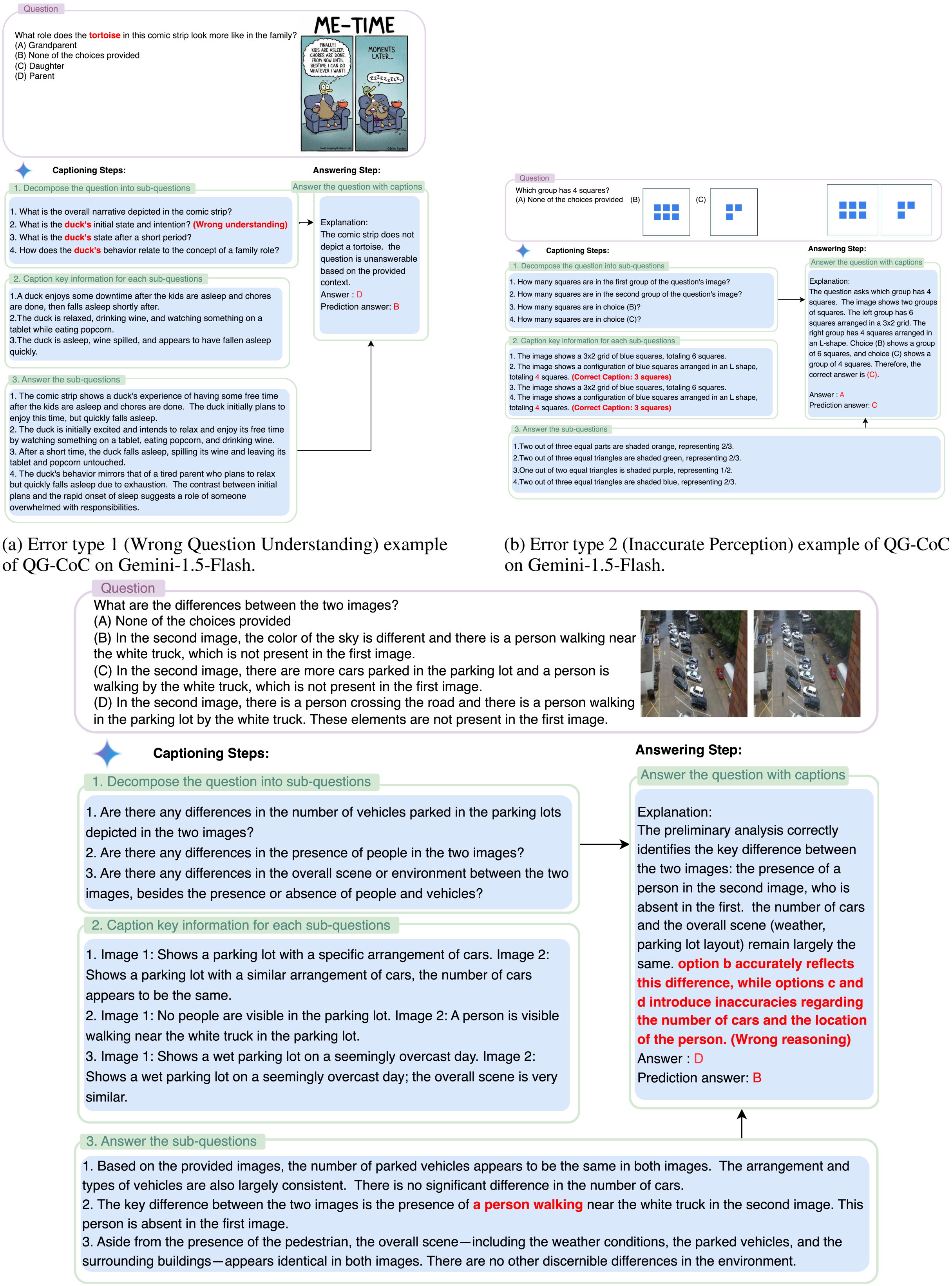
Answering Step:

Answer the question with captions

Explanation:
The provided text already gives the correct chronological order of the images as 3, 1, 2, 4. This corresponds to the man lifting the woman (3), the woman falling (1 and 2), and the woman in the toilet (4). Therefore, option (E) is correct.
Answer : E

(b) Task: Ordering

Figure 9: Examples of different tasks using QG-CoC on Gemini-1.5-Flash.


ME-TIME
FINALLY! KIDS ARE ASLEEP. CHORES ARE DONE. FROM NOW UNTIL BEDTIME I CAN DO WHATEVER I WANT!
MOMENTS LATER...
zzzzzzzzz

Question

What role does the **tortoise** in this comic strip look more like in the family?
(A) Grandparent
(B) None of the choices provided
(C) Daughter
(D) Parent

Captioning Steps:

1. Decompose the question into sub-questions

1. What is the overall narrative depicted in the comic strip?
2. What is the **duck's** initial state and intention? **(Wrong understanding)**
3. What is the **duck's** state after a short period?
4. How does the **duck's** behavior relate to the concept of a family role?

2. Caption key information for each sub-questions

1.A duck enjoys some downtime after the kids are asleep and chores are done, then falls asleep shortly after.
2.The duck is relaxed, drinking wine, and watching something on a tablet while eating popcorn.
3.The duck is asleep, wine spilled, and appears to have fallen asleep quickly.

3. Answer the sub-questions

1. The comic strip shows a duck's experience of having some free time after the kids are asleep and chores are done. The duck initially plans to enjoy this time, but quickly falls asleep.
2. The duck is initially excited and intends to relax and enjoy its free time by watching something on a tablet, eating popcorn, and drinking wine.
3. After a short time, the duck falls asleep, spilling its wine and leaving its tablet and popcorn untouched.
4. The duck's behavior mirrors that of a tired parent who plans to relax but quickly falls asleep due to exhaustion. The contrast between initial plans and the rapid onset of sleep suggests a role of someone overwhelmed with responsibilities.

Answering Step:

Answer the question with captions

Explanation:
The comic strip does not depict a tortoise. the question is unanswerable based on the provided context.
Answer : D
Prediction answer: B

(a) Error type 1 (Wrong Question Understanding) example of QG-CoC on Gemini-1.5-Flash.

Question

Which group has 4 squares?
(A) None of the choices provided (B) (C)

Captioning Steps:

1. Decompose the question into sub-questions

1. How many squares are in the first group of the question's image?
2. How many squares are in the second group of the question's image?
3. How many squares are in choice (B)?
4. How many squares are in choice (C)?

2. Caption key information for each sub-questions

1. The image shows a 3x2 grid of blue squares, totaling 6 squares.
2. The image shows a configuration of blue squares arranged in an L shape, totaling 4 squares. **(Correct Caption: 3 squares)**
3. The image shows a 3x2 grid of blue squares, totaling 6 squares.
4. The image shows a configuration of blue squares arranged in an L shape, totaling 4 squares. **(Correct Caption: 3 squares)**

3. Answer the sub-questions

1.Two out of three equal parts are shaded orange, representing 2/3.
2.Two out of three equal triangles are shaded green, representing 2/3.
3.One out of two equal triangles is shaded purple, representing 1/2.
4.Two out of three equal triangles are shaded blue, representing 2/3.

Answering Step:

Answer the question with captions

Explanation:
The question asks which group has 4 squares. The image shows two groups of squares. The left group has 6 squares arranged in a 3x2 grid. The right group has 4 squares arranged in an L-shape. Choice (B) shows a group of 6 squares, and choice (C) shows a group of 4 squares. Therefore, the correct answer is (C).

Answer : A
Prediction answer: C

(b) Error type 2 (Inaccurate Perception) example of QG-CoC on Gemini-1.5-Flash.

Question

What are the differences between the two images?
(A) None of the choices provided
(B) In the second image, the color of the sky is different and there is a person walking near the white truck, which is not present in the first image.
(C) In the second image, there are more cars parked in the parking lot and a person is walking by the white truck, which is not present in the first image.
(D) In the second image, there is a person crossing the road and there is a person walking in the parking lot by the white truck. These elements are not present in the first image.

Captioning Steps:

1. Decompose the question into sub-questions

1. Are there any differences in the number of vehicles parked in the parking lots depicted in the two images?
2. Are there any differences in the presence of people in the two images?
3. Are there any differences in the overall scene or environment between the two images, besides the presence or absence of people and vehicles?

2. Caption key information for each sub-questions

1. Image 1: Shows a parking lot with a specific arrangement of cars. Image 2: Shows a parking lot with a similar arrangement of cars, the number of cars appears to be the same.
2. Image 1: No people are visible in the parking lot. Image 2: A person is visible walking near the white truck in the parking lot.
3. Image 1: Shows a wet parking lot on a seemingly overcast day. Image 2: Shows a wet parking lot on a seemingly overcast day; the overall scene is very similar.

Answering Step:

Answer the question with captions

Explanation:
The preliminary analysis correctly identifies the key difference between the two images: the presence of a person in the second image, who is absent in the first. the number of cars and the overall scene (weather, parking lot layout) remain largely the same. **option b accurately reflects this difference, while options c and d introduce inaccuracies regarding the number of cars and the location of the person. (Wrong reasoning)**
Answer : D
Prediction answer: B

3. Answer the sub-questions

1. Based on the provided images, the number of parked vehicles appears to be the same in both images. The arrangement and types of vehicles are also largely consistent. There is no significant difference in the number of cars.
2. The key difference between the two images is the presence of **a person walking** near the white truck in the second image. This person is absent in the first image.
3. Aside from the presence of the pedestrian, the overall scene—including the weather conditions, the parked vehicles, and the surrounding buildings—appears identical in both images. There are no other discernible differences in the environment.

(c) Error type 3 (Wrong Reasoning) example of QG-CoC on Gemini-1.5-Flash.

Figure 10: Examples of three common error types made by QG-CoC on Gemini-1.5-Flash.