GlFoMR: A Glance-then-Focus Multimodal Reasoning Framework for Diagram Question Answering

Yaxian Wang
School of Computer Science and Technology, Xi'an Jiaotong University
Xi'an, China
wyx1566@gmail.com

Bifan Wei ✉
School of Continuing Education, Xi'an Jiaotong University
Xi'an, China
weibifan@xjtu.edu.cn

Jun Liu
Lingling Zhang
School of Computer Science and Technology, Xi'an Jiaotong University
Xi'an, China
liukeen@xjtu.edu.cn
zhanglling@xjtu.edu.cn

Shuting He
Shanghai University of Finance and Economics
Shanghai, China
shuting.he@sufe.edu.cn

Jun Li
State Key Laboratory of Communication Content Cognition, People's Daily Online
Beijing, China
lijun@people.cn

Qika Lin
National University of Singapore
Singapore
qikalin@foxmail.com

Abstract

Diagram question answering (DQA) is a challenging task that requires models to combine with domain-specific knowledge and reason over the diagrams to answer questions. Multimodal Large Language Models (MLLMs) have recently made notable strides in combining textual and visual information, emerging as a promising solution for addressing the DQA task. However, they still encounter challenges in deliberate multimodal reasoning over the fine-grained visual details of content-rich and knowledge-grounded diagrams. The tight interweaving of visual and textual reasoning for MLLMs is also susceptible to hallucinations. To overcome these limitations, we propose a **Gl**ance-then-**Fo**cus **M**ultimodal **R**easoning framework named **GlFoMR** for DQA, which features a flexible architecture for comprehensive visual and text interaction. Firstly, the diagram is parsed into a hierarchical structure spanning different granularities including isolated single-object, object-group, and whole-diagram. Subsequently, the Glance-Plan and Focus-Reason stages collaborate to decouple the complex reasoning process. Glance-Plan first generates a preliminary plan by glancing at the multimodal context, specifying sub-goals related to knowledge extraction, visual perception, and visual reasoning. Based on these sub-goals, Focus-Reason further integrates domain-specific knowledge and visual details to enable more deliberate reasoning. The parsed multi-granularity diagram information is seamlessly incorporated into the corresponding sub-goal achievement process, enhancing the perception and reasoning capabilities of MLLMs for better DQA performance. Extensive experimental results on four DQA datasets demonstrate that GlFoMR achieves substantial improvements, showcasing its potential to advance the development of multimodal reasoning.

CCS Concepts

• **Information systems** → **Question answering**.

Keywords

Diagram question answering, multimodal reasoning, multimodal large language model.

ACM Reference Format:
Yaxian Wang, Bifan Wei ✉, Jun Liu, Lingling Zhang, Shuting He, Jun Li, and Qika Lin. 2025. GlFoMR: A Glance-then-Focus Multimodal Reasoning Framework for Diagram Question Answering. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3729990

1 Introduction

Diagram question answering (DQA) [13, 14, 19, 33, 45] requires models to answer questions based on the given diagrams. Diagrams usually use abstract visual language like symbols, various color blocks, and shapes to describe complex concepts and their relationships, further conveying abundant information and knowledge. DQA [19–21, 37–39] has recently emerged as a vital tool to evaluate the multimodal understanding and reasoning capability of AI machines. When answering a complex diagram question, the machines are required to perform multi-step reasoning to integrate multimodal information. As illustrated in Figure 1, the diagram question requires models to not only perceive visual objects within the diagram and reason over their relations but also retrieve relevant domain-specific knowledge to derive the answer.

Recent advancements in Multimodal Large Language Models (MLLMs) [1–3, 5] have demonstrated remarkable capabilities in combining textual and visual information. Chain-of-thought (CoT) promoting strategy imitates the step-by-step reasoning process of

✉ Corresponding authors.

SIGIR '25, July 13–18, 2025, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729990

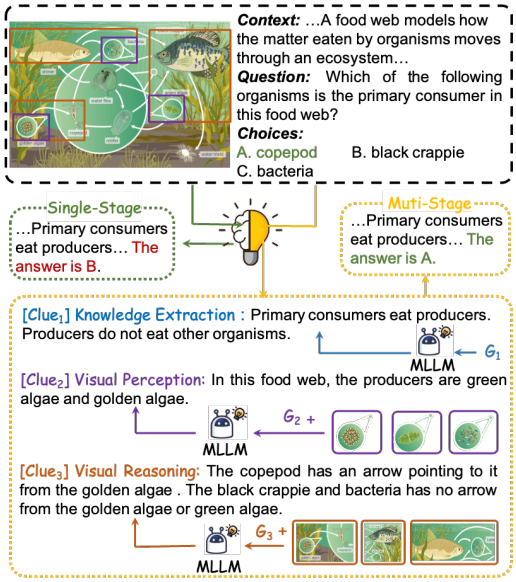

Figure 1: An example of DQA. Single-stage methods entangle all reasoning skills into a long-chain reasoning process, prone to hallucinations. Our multi-stage framework decouples the multimodal reasoning process into several sub-goals $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ about knowledge extraction, visual perception and visual reasoning. This enables MLLMs to interact with different-granularity visual information in multi-step reasoning for high-quality clues to support the final decision.

humans, significantly enhancing the problem-solving capabilities of Large Language Models (LLMs) in complex textual reasoning tasks. It has also been extended to the multimodal scenarios for MLLMs. MMCoT [10] develops a two-stage framework, which encodes the visual and text inputs into a joint space to first generate a rationale as intermediate reasoning steps and then infer the final answer. Subsequent work can be broadly divided into two categories based on their differing emphasis on textual and visual modalities. One type of method aims to enhance text comprehension, such as decomposing questions into several sub-questions [48], and introducing knowledge graphs [23]. The other type centers on improving visual understanding, such as generating composition scene graphs of images [22] or a multimodal knowledge graph [16], developing contrastive CoT prompting to compare the similarities and differences among multiple images [44], producing a multimodal latent space in a diffusion process [10], and leveraging external image processing tools for better diagram representation [34].

However, these existing methods still exhibit two primary limitations: (i) They still fall short of enabling MLLMs to fully capture the multi-granularity visual information in content-rich diagrams for deliberate multi-step reasoning. Eliciting MLLMs to effectively leverage visual information for step-by-step reasoning remains a pivotal challenge in DQA. (ii) They intertwine various skills, including visual perception, visual reasoning, and textual reasoning within a single stage to complete the problem-solving process. As a result, it is prone to introducing hallucinations in a long-chain, multi-step reasoning process, leading to unsatisfactory performance.

To overcome these limitations, we propose a **Gl**ance-then-**Fo**cus **M**ultimodal **R**easoning framework named **GlFoMR** for DQA, drawing inspiration from the human cognition process [27–29]. When facing a complex question about a diagram, humans typically begin by browsing the diagram, question, and candidate choices to formulate a solution plan based on their experience. Guided by this plan, they then proceed with deeper, step-by-step reasoning, focusing more on domain-specific knowledge and specific details of the diagram, particularly the question-related specific visual entities and their relationships. As illustrated in Figure 1, compared with the single-stage entangled reasoning methods, GlFoMR as a multi-stage framework decomposes the reasoning process, endowing MLLMs the chance to flexibly interact with the different-granularity visual information in step-by-step reasoning.

Specifically, GlFoMR mainly consists of a Hierarchical Diagram Parsing module and a Multimodal Interactive Reasoning module. For Hierarchical Diagram Parsing, the diagram is parsed into a hierarchical structure spanning different granularities including isolated single-object, object-group, and whole-diagram. Multimodal Interactive Reasoning consists of three stages: (i) Glance-Plan stage. The MLLM browses the diagram, context, question, and candidate choices to formulate a solution plan. Tailoring the different question complexities, each specific diagram-question pair is decomposed into different sub-goals. Given the nature of DQA, these sub-goals involve knowledge extraction, visual perception, and visual reasoning. Knowledge extraction focuses on acquiring question-related concepts and domain-specific background knowledge. Visual perception aims to delve deeply into the specific details of question-related visual entities. High-level object relations are also taken into consideration for visual reasoning based on the given question. A complex question requires achieving all three sub-goals. (ii) Focus-Reason stage. Based on the solution plan comprised of several sub-goals, the parsed multi-granularity diagram information is seamlessly integrated into the corresponding sub-goal achievement process, further enhancing MLLMs' perception and reasoning ability. (iii) Summarize-Answer stage. It gathers all question-related context clues to make a final decision. The experimental results on four DQA benchmarks including SQA-IMG [19], CSDQA [33], AI2D [13], and TQA-DMC [14], indicate the effectiveness of GlFoMR, with substantial average performance gains (up to 3.99%).

Our main contributions are summarized as follows:

- We propose a training-free Glance-then-Focus Mutimodal Reasoning framework GlFoMR, which features flexible visual-text interaction and disentangled multimodal reasoning capabilities.
- By decoupling the reasoning processes into distinct Glance-Plan and Focus-Reasoning stages, the MLLMs are endowed with chances to engage in full interactions with detailed visual information. It enables a holistic understanding of visual content, beginning with the global diagram and delving into individual objects and their relationships.
- The sub-goals decomposition facilitates the seamless integration of multi-granularity diagram information and domain-specific knowledge in step-wise reasoning, mitigating the hallucinations in the multi-step reasoning process.
- Extensive experiments on four DQA benchmarks showcase the significant advantages of our proposed GlFoMR, compared with its existing counterparts.

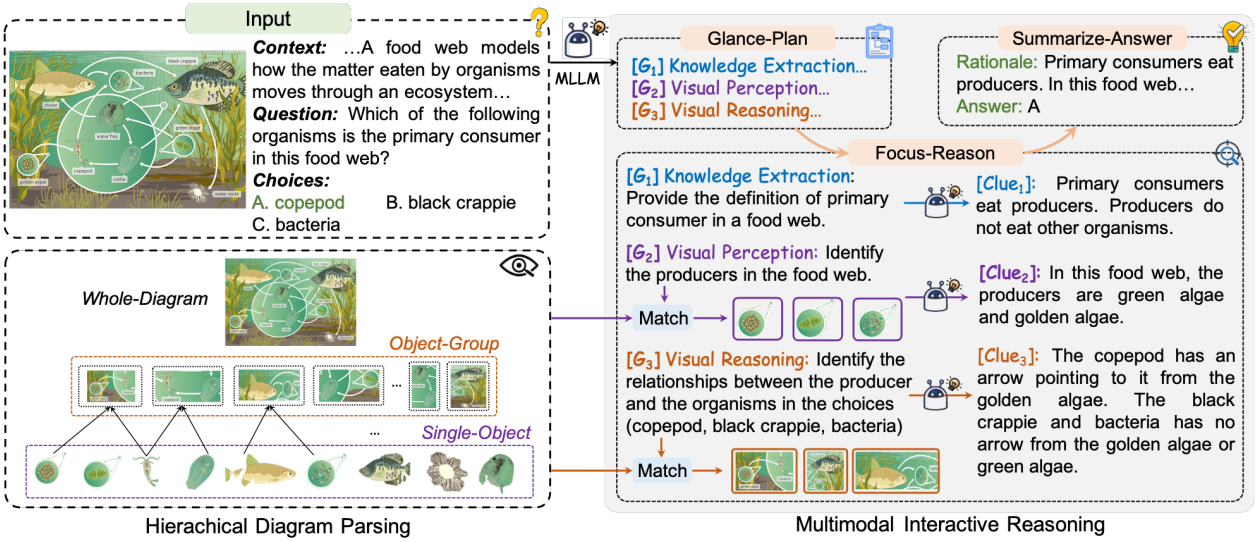

Figure 2: Overview of GlFoMR. (a) Hierarchical Diagram Parsing decomposes the diagram into a hierarchical structure spanning different granularities for subsequent visual perception and reasoning. (b) Multimodal interactive Reasoning includes three stages, Glance-Plan, Focus-Reason, and Summarize-Answer to complete the multimodal reasoning process.

2 Related Work

2.1 Multimodal Large Language Model

Recent research has increasingly concentrated on MLLMs [3, 5, 18, 30, 36], inspired by the remarkable progress of LLMs [4, 7, 32, 40]. Several MLLMs have achieved notable advancements in integrating textual and visual information to improve multimodal understanding. Popular open-source models such as the LLaVA series [18], InstructionBLIP [36], QwenVL [3], and InternVL [5] have seen substantial development. Similarly, closed-source models like GPT-4V [1], GPT-4o [11], Gemini [30], OpenAI o1 [12] and OpenAI o3 [6] have demonstrated exceptional performance in a range of multimodal tasks. However, these models still face challenges in leveraging detailed visual information for step-wise reasoning, particularly when employing a single-stage approach to tackle complex problem-solving tasks.

2.2 CoT Reasoning of MLLMs

The success of CoT prompting in textual tasks indicates its potential for effective extension to multimodal scenarios. MM-CoT [46] is the first to propose a two-stage framework, which fuses the text and image to first perform rationale generation and then answer inference. Based on the same framework, a series of works [10, 23, 31, 34] design strategies to enhance the multimodal understanding. He et al. [10] introduced a multimodal latent space to fuse visual and text features deeply via a diffusion process. Monda et al. [23] proposed to leverage knowledge graphs for better multimodal understanding. Lee et al. [16] utilized multimodal knowledge graphs to learn rich semantic knowledge, thereby enhancing the multimodal reasoning capabilities of LLMs. Wang et al. [31] utilized LLM-generated CoT for greater amounts of essential external knowledge, which can also mitigate the time-consuming human annotations. Wang et al. [34] proposed to introduce image processing tools to enhance image representation, further promoting rationale generation and answer inference. These methods rely on fine-tuning learning on ground truth rationale and answer annotation.

Another line of research utilizes the prompt-based mechanism to elicit the CoT capability, eliminating the need for time-consuming fine-tuning. DDCoT [48] uses language-only LLMs to decompose the original problem into sub-problems, and then utilizes VQA models to introduce visual information by answering the visual-related sub-problems. The question-decomposition methods [42, 48] overlook the visual information during the decomposition process, and rely solely on the image to answer sub-problems. CCoT [22] extracts the compositional information of images as scene graphs to enhance the CoT process. IoT [49] utilizes external image processing tools including segmentation, zoom-in, and color-space conversions to exploit visual information. CoCoT [44] develops a contrastive chain-of-thought prompting to compare the similarities and differences among multiple images, further promoting the visual perception ability of MLLMs. Our work follows this training-free prompt-based paradigm. In contrast, we disentangled the complex reasoning process into glance-plan and focus-reason stages to help alleviate hallucination issues aggravated by tightly entangled thinking and reasoning, while seamlessly introducing visual details of varying granularity for decomposed sub-goals.

3 Method

3.1 Overview

Our proposed GlFoMR is illustrated in Figure 2. It mainly consists of two modules including Hierarchical Diagram Parsing and Multimodal Interactive Reasoning. **Hierarchical Diagram Parsing** decomposes the diagram into multi-granularity visual information, encompassing the single-object, object-group, and whole-diagram

granularity. **Multimodal Interactive Reasoning** mainly comprises of three stages: (1) **Glance-Plan**, which browses the diagram, question, and candidate choices to formulate a solution plan. Tailored to the question difficulty, this solution plan decomposes the question-answering task into different sub-goals about knowledge extraction, visual perception, or visual reasoning. (2) **Focus-Reason**, with the solution plan composed of a chain of reasoning guidance about sub-goals, the parsed multi-granularity diagram information is seamlessly integrated into the corresponding subgoal-achieving process, further enhancing MLLMs' perception and reasoning ability. (3) **Summarize-Answer** summarizes the collected information from the previous stages to form a smooth and logical rationale, further supporting the final decision.

3.2 Hierarchical Diagram Parsing

A diagram usually employs visual elements like symbols, various color blocks, and shapes to describe complex concepts and their relationships, conveying rich information and knowledge in an intuitive visual form. To further enhance diagram understanding, we automatically parse the diagram into a hierarchical structure spanning different granularities as follows:

- **Single-object granularity.** We employ visual detection tools (e.g., SAM [15] or YOLOv3 [8]) to recognize the visual objects in the diagram, denoted as $\mathcal{R}^o = \{v_1, v_2, ...v_N\}$.
- **Object-group granularity.** It is the combination of paired objects by analyzing their relations. According to the visual perception principles, we aggregate the related objects into object groups. The visual perception principles are as follows: (i) Spatial Proximity Principle: The visual objects that are spatially proximate to each other are more likely to be related. For each visual object, the top-3 related objects are selected to form object groups. (ii) Connectivity Principle: The visual objects that are visually connected, such as through lines, arrows, or other linking elements are perceived as being associated in some way. We model relationships between objects following these two principles and then obtain the corresponding object groups. These object groups are denoted as $\mathcal{R}^g = \{g_1, g_2, ..., g_M\}$.
- **Whole-diagram granularity.** It denotes the global diagram D provided in each diagram-question pair.

In this way, we construct a candidate visual region list with different granularities $\mathcal{R} = \{\mathcal{R}^o, \mathcal{R}^g, D\}$. The whole diagram offers a holistic view of the scene, facilitating a deeper understanding of its overall context. Meanwhile, the regions representing visual objects and object groups highlight local areas, facilitating focused attention on crucial local details. By integrating insights from both the overall diagram and the fine-grained regions, a more comprehensive understanding of the visual content can be attained.

3.3 Glance-Plan Stage

The Glance-Plan stage involves quickly browsing the diagram, context (if it has), question, and candidate choices to formulate a solution plan. If the question is simple, it can be addressed directly with fast thinking using inherent commonsense knowledge, without the need to break it down to seek visual details. Otherwise, it requires slow and deliberate reasoning. Although current MLLMs have shown remarkable performance in various tasks, such as image captioning [9, 26] and object detection [35, 41, 43, 47], they still struggle with complex tasks that demand the integration of both visual understanding and reasoning capabilities. Moreover, they tend to hallucinate over questions that require deliberated multi-step reasoning on visual details. To mitigate hallucinations and compel more visual information into the reasoning process, we decompose the task into several sub-goals. Not all sub-goals are activated for each specific diagram-question pair, which varies with the difficulty of the questions. Emulating human thinking, the MLLM decouples the problem-solving task into a series of relevant sub-goals about different functions that serve as a chain of reasoning guidance.

Given the nature of DQA, a complex problem-solving process involves knowledge extraction, visual perception, and visual reasoning. The first focuses on domain-specific background knowledge, while the last two pay more attention to the visual details of objects and object relations in the diagram. Specifically, **Knowledge Extraction** focuses on acquiring question-related concepts and background knowledge required for perception and reasoning, for example, the concept of "*producer*" and "*primary consumer*". **Visual Perception** aims to delve deeply into the specific visual details of question-related visual objects, for example, "Recognize *green algae* and *golden algae*". **Visual Reasoning** requires reasoning over the visual objects and object relations, for example, "*the copepod has an arrow pointing to it from the golden algae*". We design a prompt to unveil their specific responsibilities and clarify the division of labor and their roles in the DQA task as follows:

Prompt for Glance-Plan (Part I)

You are a professional and helpful science question-answering expert. Your task is to guide the users to select the right answer to the question. You are guided by three kind of crucial sub-goals to help analyze and answer questions about diagrams:
1. Knowledge Extraction: This sub-goal introduces additional knowledge that is necessary for the questions. If this sub-goal is required, please specify your request as: " Knowledge Extraction : <specify the concepts>, providing the concepts whose related knowledge needs to be obtained."
2. Visual Perception: This sub-goal interprets and makes sense of images, enabling you to deal with questions related to the diagram's content. If this sub-goal is required, specify your request as: "Visual Perception: <specify the objects or text need to recognize from the diagram >," providing the objects or text in the diagram you believe need to be recognized for further analysis.
3. Visual Reasoning: This sub-goal analyzes the relationships between the objects within the diagrams, such as semantic or spatial relationships. When this sub-goal is required, specify your request as: "Visual Reasoning: <the objects need to compare or reason further>," providing the objects that you believe need to compare or reason further.

Specifically, given a diagram D, the context C (if has), a question Q, and candidate choices $\mathcal{A} = \{a_1, a_2, ..., a_n\}$, we adopt an MLLM $\mathcal{F}(\cdot)$ to generate a preliminary solution plan including a chain of reasoning guidance $\mathcal{G} = \{\mathcal{G}_1, ..., \mathcal{G}_k\}$ about different sub-goals:

$$\mathcal{G} = \mathcal{F}[D, C, Q, \mathcal{A}]. \quad (1)$$

The decomposition-related prompt is designed as shown in the Prompt for Glance-Plan (Part II).

Each reasoning guidance is a concise summary of the corresponding sub-goal. As shown in Figure 2, the MLLM generates a deliberated solution plan $\mathcal{G}$, acting as the reasoning guidance in the next stage. By separating visual perception, reasoning, and knowledge extraction, the framework allows the flexible integration of different-granularity visual details and knowledge, effectively reducing hallucinations arising from long-chain reasoning. Compared with the question-decomposition methods [42, 48], the goals-based

Prompt for Glance-Plan (Part II)

Given a diagram, the context (if has) as the hint, a question, choices, please think step-by-step to provide a rationale about how to answer the question. Note that, do not reveal the answer in the process. You need to specify which sub-goal should be used to obtain the essential information for question answering.

The expected response format is as follows:
Plan:
["This question is a common sense question, which can be answered directly without the information from the diagram."] or [Rationale: Your plan to solve the question based on the question and diagram provided. Explain how each sub-goal contributes to the question answering.]

Duties of Sub-goals:
1. Knowledge Extraction: [Provide the relevant domain knowledge about the question, if required.]
1. Visual Perception: [List the objects or text to be recognized or extracted in the diagram, if required.]
3. Visual Reasoning: [Point out the objects whose relationships need deeper analysis, if necessary.]

Please ensure that your response follows this format. Note that, if both Visual Perception and Visual Reasoning are needed, Visual Perception should come before Visual Reasoning. If Knowledge Extraction is required, it should be the first step of the plan.

Here are some examples:
....

Please refer to the prompt and in-context examples above to help me solve the following problem: Context: {context}, Question: {question}, Choices: {choices}

decomposition at a higher level offers greater flexibility to handle various problems in a more structural manner, mitigating the issues caused by over- or under-decomposition in sub-problems. Moreover, the functionality of each sub-goal and the reasoning logic are clear, allowing the system's behavior to be traced for a specific diagram-question pair. This detachment also simplifies error diagnosis, making it easier to identify issues such as insufficient knowledge, inaccurate perception, or flawed reasoning.

3.4 Focus-Reason Stage

For reasoning guidance $\mathcal{G}_i$ about sub-goals of visual perception and visual reasoning, the parsed single objects and object groups are seamlessly integrated into the reasoning process. Specifically, each visual guidance is linked to the diagram content with the corresponding granularity. We retrieve relevant single objects or object groups corresponding to the guidance $\mathcal{G}_i$ from the candidate region list $\mathcal{R}^o$ and $\mathcal{R}^g$, respectively. This is achieved by first computing their matching scores $\{S_{ij}\}_{j=1}^{\{N,M\}}$, and then selecting the visual regions with top k highest matching scores. Taking visual perception as an example, the computation is as follows:

$$\{S_{ij}\}_{j=1}^{N} = \text{softmax}(\mathcal{F}_m(\mathcal{G}_i, \mathcal{R}^o)), \tag{2}$$

$$\{j_1, j_2, ..., j_k\} = \text{Topk}(\text{argsort}(\{S_{ij}\}_{j=1}^{N})), \tag{3}$$

$$\mathcal{R}_k^o(\mathcal{G}_i) = [v_{j_1}, v_{j_2}, ..., v_{j_k}], \tag{4}$$

where $\mathcal{G}_i \in \mathcal{G}$, $\mathcal{F}_m$ denotes a image-text matching function, argsort denotes a sorting operation for the matching scores, Topk denotes a function to extract the indices with the top k highest scores, and softmax is used to normalize the scores.

With the retrieved related visual regions, the MLLM reasons over multi-granularity visual contents to provide more detailed information. For knowledge extraction, the MLLM can directly generate responses based on its inherent domain-specific knowledge.

Algorithm 1 GlFoMR

Input: Diagram D, Context C (if it exists), Question Q, and Candidate choices $\mathcal{A}$.
Output: Rationale Rat and Final Answer Ans.
$\mathcal{R} = \{\mathcal{R}^o, \mathcal{R}^g, D\} \leftarrow$ Hierarchical Diagram Parsing.
$\mathcal{G} = \{\mathcal{G}_1, ..., \mathcal{G}_n\} \leftarrow$ MLLM.Glance_Plan(D, C, Q, $\mathcal{A}$).
for each reasoning guidance $\mathcal{G}_i$ **do**
  $Clue_i \leftarrow$ MLLM.Focus_Reason($\mathcal{G}_i$, $\mathcal{R}$).
end for
$SR = \{[\mathcal{G}_1, Clue_1]; ...; [\mathcal{G}_n, Clue_n]\} \leftarrow$ Get the chain-of-clues.
$Rat, Ans \leftarrow$ MLLM.Summarize_Answer(D, C, Q, $\mathcal{A}$, SR).
Return Rat, Ans.

Table 1: The statistics of the test set of four DQA datasets.

Dataset	SQA-IMG	CSDQA	TQA-DMC	AI2D
Diagram	2,017	238	660	308
Question	2,017	618	3,285	978

For visual perception and visual reasoning, the corresponding guidance G_i, the retrieved regions $\mathcal{R}_k^{\{o,g\}}(G_i)$ along with the whole diagram are provided as inputs to the MLLM, enabling it to produce the necessary clue $Clue_i$. Note that the clue generated from the visual perception is also integrated into the visual reasoning subgoal-achievement process, allowing low-level visual perception to enhance high-level visual reasoning.

3.5 Summarize-Answer Stage

In the Focus-Reason stage, we have collected the question-related context clues $\{Clue_1, ...Clue_n\}$ for question answering. Then, in the Summarize-Answer stage, an MLLM is introduced to summarize the supplementary information as chain-of-clues, denoted as $SR = \{[\mathcal{G}_1, Clue_1]; ..., [\mathcal{G}_n, Clue_n]\}$ to support the decision-making process. As demonstrated in Algorithm 1, these clues together with the corresponding diagram, context (if it has), question, and candidate choices are fed into the MLLM to generate a coherent and logical rationale, culminating in the final answer, which can be formalized as follows:

$$Rat, Ans = \mathcal{F}(D, C, Q, \mathcal{A}, SR). \tag{5}$$

The specific prompt for the Summarize-Answer stage is introduced as follows:

Prompt for Summarize-Answer

You are a knowledgeable and skilled science expert. Please gradually think and answer the questions based on the given diagram, context (if has), question, choices, and supplementary information. Please note that we need not only the answer, but more importantly, the rationales of getting the answer.
Note that the supplementary information given may not always be valid, please maintain critical thinking and select effective information to form the rationale and select the most correct choice as the answer.

The expected response format is as follows:
Rationale: <rationale>,
Answer: <one of the choices, such as A, B >.
Please note that the answer must be one of the choices. Do not write a full sentence, just provide a letter choice.
Please answer the following case: Context:{context}, Question:{question}, Choices:{choices}, Supplementary information:{supplementary_information}.

Table 2: Performance comparisons on the test set of four DQA datasets. Question classes: NAT = natural science, SOC = social science, and LAN = language science. TF denotes True-or-False questions, and MC denotes Multiple-Choice questions.

Backbone	Model	SQA-IMG				CSDQA			TQA-DMC	AI2D	Avg.
		NAT	SOC	LAN	All	TF	MC	All			
Open-Source Models	LLaVa-1.5-7B [17]	-	-	-	65.08	62.65	28.64	47.31	35.38	65.76	49.44
	LLaVa-1.5-13B [17]	-	-	-	68.45	64.43	29.36	48.65	37.64	67.77	51.91
	CCoT (LLaVa-1.5-13B) [22]	68.65	72.77	75.00	70.35	61.81	46.93	54.37	55.98	67.80	61.71
	MM-CoT-large [46]	-	-	-	73.53	63.11	51.46	57.28	52.51	75.76	62.38
	CoG-DQA [34]	76.10	79.45	65.91	78.85	72.82	63.75	68.28	54.60	79.14	66.40
GPT-3.5/4	CoT (GPT-3.5) [1]	70.22	62.30	70.45	67.23	65.02	27.04	46.83	36.20	53.39	48.72
	CoT (GPT-4) [1]	71.71	70.16	81.82	71.34	66.85	27.45	48.03	38.47	58.68	51.85
	DDCoT (GPT-3.5) [48]	-	-	-	72.53	-	-	-	-	-	-
	GlFoMR (GPT-3.5)	**82.71**	**82.85**	**90.91**	**82.95**	**68.61**	**67.31**	**67.96**	**72.09**	**81.90**	**76.29**
Gemini-1.5-flash	Direct	66.09	65.97	70.45	66.14	71.52	61.17	66.34	66.92	80.98	68.63
	CoT [1]	80.23	82.72	93.18	81.46	72.17	69.58	70.87	70.71	82.82	75.58
	CCoT [22]	75.60	67.41	**97.73**	72.98	69.90	69.26	69.58	72.24	82.21	73.63
	GlFoMR	**85.61**	**86.26**	**97.73**	**86.12**	**73.46**	**72.17**	**72.82**	**74.09**	**84.97**	**79.04**
GPT-4o-mini	Direct	68.65	67.41	79.55	68.42	70.87	63.78	66.83	67.88	81.70	69.90
	CoT [1]	80.89	85.86	**97.73**	83.14	68.28	61.17	64.72	70.32	84.05	75.51
	CCoT [22]	77.67	80.37	95.45	79.08	74.76	67.64	71.52	73.00	83.23	76.10
	GlFoMR	**87.26**	**87.17**	95.45	**87.40**	**76.70**	**68.93**	**72.82**	**75.13**	**86.30**	**80.09**

4 Experiments

We conduct extensive experiments across four DQA benchmarks to validate the effectiveness of our method. In Section 4.1, we first introduce the details of four DQA datasets, baseline models, and the implementation details of GlFoMR. Subsequently, the experimental results and analyses are presented in Section 4.2. Next, we further evaluate the effectiveness of GlFoMR through several ablation studies in Section 4.3. The sub-goal statistics are introduced in Section 4.4. Finally, the qualitative results and error analysis are shown in Section 4.5 and 4.6 for more insights.

4.1 Experimental Settings

4.1.1 Datasets. In this section, we evaluate our proposed GlFoMR on the test set of four DQA datasets: SQA-IMG, CSDQA, TQA, and AI2D. The specific dataset statistics are shown in Table 1.

- **SQA-IMG** [19] is a subset of the multimodal science question benchmark, containing only questions with accompanying diagrams. It encompasses three major scientific disciplines: natural sciences (NAT), social sciences (SOC), and language sciences (LAN). The test set is comprised of 2,017 samples.
- **CSDQA** [33] is collected from computer science in university courses, containing 618 questions about 238 diagrams.
- **TQA-DMC** [14] is derived from Textbook Question Answering (TQA) dataset, involving life science, earth science, and physical science textbooks of middle school science curricula. The test set contains 3,285 questions about 660 diagrams.
- **AI2D** [13] is collected from grade school science, with a test set comprising of 308 diagrams and 978 questions.

4.1.2 Implementation Details. In the experiment, we use different prevalent MLLMs as backbones, including Gemini-flash-1.5 and GPT-4o-mini, by calling their API to evaluate the model performance. For GlFoMR (GPT-3.5), the Glance-Plan and Summarize-Answer stages utilize GPT-3.5, with the diagram-related information for these stages provided by Gemini-flash-1.5. Meanwhile, the Focus-Reason stage relies on the corresponding MLLM to meet the demands of multimodality. In the Hierarchical Diagram Parsing module, we employed the Segment Anything Model (SAM) [15] to obtain isolated visual objects. In the Glance-Plan stage, we provide a four-shot in-context example to guide the solution guidance generation. Tailored to the specific characteristics of different datasets, we design distinct in-context examples in the Glance-Plan stage, while keeping the remaining parts of the prompts consistent. In the Focus-Reason stage, we use CLIP [25] to match each reasoning guidance with the top k most related visual content at the corresponding granularity by calculating their similarities.

4.1.3 Baseline Models. We mainly selected several classic prompt-based models as our baselines. We divide them into two categories:

- Open-Source MLLMs: LLaVa-1.5-7B [17], LLaVa-1.5-13B [17], MM-CoT-large [46], and CoG-DQA [34].
- Closed-Source MLLMs with large-scale parameters: GPT-3.5 [24], GPT-4 [1], Gemini-1.5-flash [2], GPT-4o-mini [11]. There are several prompt-based methods to further improve their performance, including CoT [1], CCoT [22], and DDCOT [48].

4.2 Main Results

We comprehensively evaluate the performance of GlFoMR on top of Gemini-flash-1.5 and GPT-4o-mini through SQA-IMG [19], CSDQA [33], AI2D [13], and TQA-DMC [14]. The results on four different DQA benchmarks are displayed in Table 2. We primarily use accuracy as the evaluation metric to quantify the models' performance. We can make the following five observations: (1) GlFoMR with

Table 3: Ablation studies of GlFoMR based on Gemini-1.5-flash and GPT-4o-mini on SQA-IMG, CSDQA and AI2D.

Model	SQA-IMG				CSDQA			AI2D
	NAT	SOC	LAN	Avg.	TF	MC	Avg.	
Gemini-1.5-flash								
Base	**85.61**	**86.26**	**97.73**	**86.12**	**73.46**	**72.17**	**72.82**	**84.97**
w/o HDP	82.21 (↓3.40)	84.69 (↓1.57)	95.45 (↓2.28)	83.44 (↓2.68)	72.49 (↓0.97)	71.52 (↓0.65)	72.00 (↓0.82)	82.21 (↓2.76)
w/o GP	80.23 (↓5.38)	83.12 (↓3.14)	93.18 (↓4.55)	81.46 (↓4.66)	72.17 (↓1.29)	69.58 (↓2.59)	70.71 (↓2.11)	82.82 (↓2.15)
w/o R	83.20 (↓2.41)	84.55 (↓1.71)	97.73 (↓0.00)	84.04 (↓2.08)	73.14 (↓0.32)	70.55 (↓1.62)	71.84 (↓0.98)	82.71 (↓2.26)
GPT-4o-mini								
Base	**87.26**	**87.17**	**95.45**	**87.40**	**76.70**	**68.93**	**72.82**	**86.30**
w/o HDP	84.37 (↓2.89)	86.00 (↓1.17)	93.18 (↓2.27)	85.18 (↓2.22)	74.43 (↓2.27)	66.34 (↓2.59)	70.39 (↓2.43)	83.23 (↓3.07)
w/o GP	80.89 (↓6.37)	85.86 (↓1.31)	97.73 (↑2.28)	83.14 (↓4.26)	68.28 (↓8.42)	61.17 (↓7.76)	64.72 (↓8.10)	84.05 (↓2.25)
w/o R	82.63 (↓4.63)	86.12 (↓1.05)	95.45 (↓0.00)	84.23 (↓3.17)	72.50 (↓4.20)	64.10 (↓4.83)	68.28 (↓4.54)	83.54 (↓2.76)

different MLLM backbones exhibits significant and consistent performance improvement over all other baselines on four benchmarks. It shows an average performance improvement of 3.46% over CoT based on Gemini-1.5-flash and 3.99% over CCoT based on GPT-4o-mini, demonstrating its effectiveness and generalization ability. Moreover, GlFoMR utilizing both Gemini-1.5-flash and GPT-4o-mini surpasses GlFoMR with GPT-3.5, suggesting that the incorporation of holistic visual information enables MLLMs to formulate more rational reasoning plans. (2) GlFoMR with Gemini-1.5-flash and GPT-4o-mini achieves a peak of 4.66% and 4.26% improvement respectively on the SQA-IMG dataset compared with the best baseline CoT. We analyze that the integration of multi-granularity diagram information can effectively promote problem-solving performance. (3) GlFoMR receives a relatively small performance boost 1.95% and 1.3% on the CSDQA dataset compared to the best baseline with Gemini-1.5-flash and GPT-4o-mini, respectively. We analyze that the abstract visual representation and their complex object interactions in the computer science domain still pose a challenge for MLLMs. (4) Compared to CoT with Gemini-1.5-flash or GPT-4o-mini as backbones, our model achieves 3.46% and 4.66% performance improvement, respectively, which demonstrates the advantage of our two decoupled designs in GlFoMR over the tight interweaving of visual and textual reasoning in CoT. (5) Compared to other multimodal prompt-based methods, the performance gains of GlFoMR further highlight that the seamless integration of multi-granularity visual details into the reasoning process yields greater effectiveness and advantages.

4.3 Ablation Studies

4.3.1 The impact of main modules. To further evaluate the effectiveness of each design in our GlFoMR framework, we conduct a series of ablation studies across SQA-IMG, CSDQA, and AI2D in Table 3. When Hierarchical Diagram Parsing (HDP) is removed, we replace the multi-granularity visual information in the Focus-Reasoning stage with the whole diagram, leading to noticeable performance degradation for either Gemini-1.5-flash or GPT-4o-mini. This proves the importance of HDP in providing crucial different-granularity visual details compared with the entire content-rich diagram. Since the Focus-Reason stage depends on the Glance-Plan stage, w/o GP

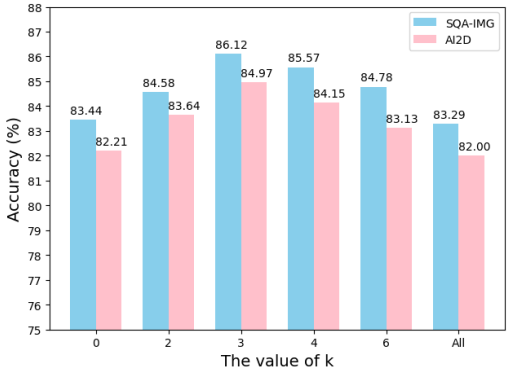

Figure 3: The impact of different numbers of visual regions k relevant to the sub-goals from the Glance-Plan stage. $k = 0$ denotes no parsing visual region is used, and $k = All$ denotes all parsing visual regions are used.

represents the removal of both Glance-Plan (GP) and Focus-Reason (FR). In this case, GlFoMR degrades to CoT. The results highlight the effectiveness of our decomposed design in mitigating the hallucinations that arise from the tightly entangled reasoning process. In the summarize-answer stage, we remove the generation of rationale R. We find that rationale promotes a smooth and logical flow of thought, helping generate a coherent reasoning process, and ultimately helping guide the model to the correct answer. To a certain extent, it helps to improve the question-answering accuracy, especially in complex questions that require multi-step reasoning. Overall, our GlFoMR framework is a cohesive system wherein each design is essential to performance.

4.3.2 The impact of the number of matching visual regions. To explore the impact of the number of visual regions k on sub-goals, we evaluate different values of k and compare the results on SQA-IMG and AI2D. The results are as shown in Figure 3. The model's performance initially improves as k increases, peaking at 86.12% and 72.82% for SQA-IMG and AI2D, respectively, when $k = 3$. These experimental results demonstrate that parsing diagrams into fine-grained visual regions enables the model to focus on relevant local information and capture detailed visual features, thereby enhancing overall performance. However, when $k = 4$, the accuracy begins

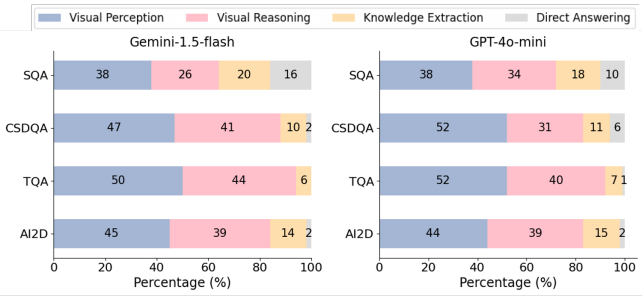

Figure 4: Statistics of GlFoMR's utilization of three sub-goals across the four DQA datasets.

to decline slightly. When $k = All$, a significant drop occurs. The performance is even worse than that of $k = 0$. We attribute the performance drops to the introduction of excessive noise or redundant information when the number of visual regions reaches a threshold.

4.4 Sub-goal Statistics

The statistics of different sub-goals generated by GlFoMR for four DQA datasets are shown in Figure 4. Through the analysis, we can gain deeper insights into the distribution of question types for different datasets. We find that Gemini-1.5-flash and GPT-4o-mini exhibit relatively similar glance-plan patterns. The utilization distribution of visual perception, visual reasoning, knowledge extraction, and direct answering shows a comparable trend in both models, with a higher reliance on visual perception and visual reasoning. For all four datasets, visual perception is the most commonly employed sub-goal, followed by visual reasoning and knowledge extraction, as depicted in Figure 4. It aligns with the intuition that visual perception is the cornerstone of diagram question answering tasks. The proportion of visual reasoning is slightly lower than that of visual perception, but much higher than that of knowledge extraction. Compared to other datasets, the SQA dataset contains the highest proportion of questions that can be directly answered using commonsense knowledge. It also has the highest proportion of questions requiring domain-specific knowledge for reasoning. In contrast, the TQA-DMC has the lowest proportion of questions that can be answered directly, and most questions require visual perception and reasoning to answer.

4.5 Case Studies

To further intuitively demonstrate the effectiveness of our proposed framework GlFoMR, we conduct case studies on three success cases and two failure cases, as illustrated in Figure 6. These cases are selected from SQA, CSDQA and AI2D from different subject areas. Specifically, GlFoMR first generates a preliminary plan via the Glance-Plan stage. The first and third cases involve all three sub-goals about knowledge extraction, visual perception, and visual reasoning. It guides MLLMs to output the domain-specific knowledge about *magnetic force comparison* and *First-In-Last-Out principles*. For sub-goals about visual perception and visual reasoning, GlFoMR seamlessly integrates the most related visual details of different granularities from the parsed diagram to elicit the MLLMs to conclude the high-quality clues. The decomposition of the reasoning process enables the flexible integration of different-granularity

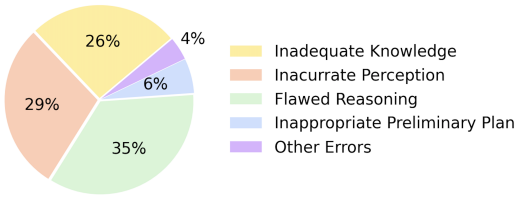

Figure 5: Error distribution over 80 problems across the four DQA datasets using Gemini-1.5-flash.

visual details and domain-specific knowledge, effectively reducing hallucinations arising from long-chain reasoning. It also facilitates the identification of errors, such as insufficient knowledge, inaccurate perception, or flawed reasoning. For example, in the first failure case, GlFoMR makes mistakes in reasoning about whether node C can have another node. The second failure case provides insufficient knowledge about the kinetic energy. It is calculated using the formula $KE = \frac{1}{2}mv^2$, where m is the mass and v is the velocity of particles. This knowledge point is crucial to addressing the problem.

4.6 Error Analysis

To thoroughly examine the limitations of GlFoMR and help point towards areas needing further research, we conduct a detailed error analysis when using Gemini-1.5-flash as the backbone. We randomly sampled 80 error samples from four datasets, with 20 samples from each dataset. The errors were categorized as follows. **(1) Inadequate Knowledge (26%).** Irrelevant or insufficient knowledge provision can result in deviations in subsequent reasoning. For example, when guiding the comparison of magnetic forces between magnets, the impact of magnet size is overlooked, leading to reasoning errors. **(2) Inaccurate Perception (29%).** Despite being capable of acquiring fine-grained visual information, the model may still struggle to accurately interpret some key visual cues. This limitation often arises from the difficulty in understanding small-scale visual details or domain-specific nuances in the diagram. These errors manifest as misidentification or misclassification of objects, particularly when the task demands precise visual perception. For instance, the model might incorrectly perceive solution particles or misidentify empty stacks. **(3) Flawed Reasoning (35%).** This emerges as a significant cause of errors. While the model successfully interprets text and images and retrieves relevant knowledge, it often struggles with complex reasoning effectively, resulting in inaccurate answers. For instance, the model might make errors when reasoning about species changes in a food web. **(4) Inappropriate Preliminary Plan (6%).** In the Glance-Plan stage, the proposed solution plan is not closely aligned with the actual problem-solving process, introducing irrelevant or distracting information. **(5) Other Errors (4%).** For instance, the reasoning process may be correct, but the final incorrect answer is chosen.

5 Conclusion

In this paper, we propose a novel training-free Glance-then-Focus Multimodal Reasoning framework GlFoMR for diagram question answering, which features flexible visual-text interaction and disentangled multimodal reasoning capabilities. The diagram is first parsed into a hierarchical structure with visual contents of different

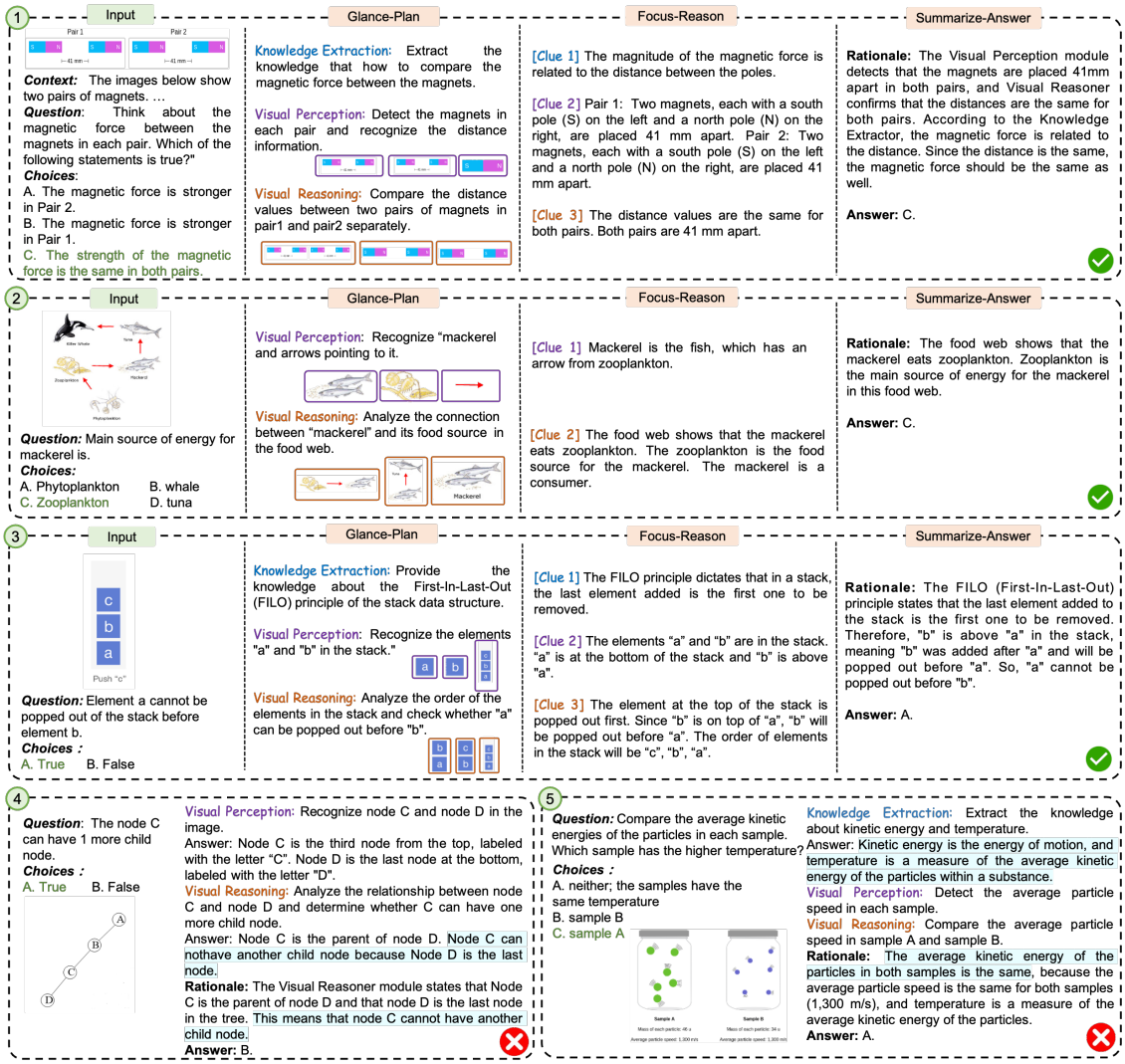

Figure 6: Case studies of our proposed GlFoMR using Gemini-1.5-flash. These cases are selected from SQA, AI2D, and CSDQA, respectively. We display the intermediate output from Glance-Plan, Focus-Reason, and Summarize-Answer stages to delve into the effectiveness of GlFoMR. The blue background highlights the errors for the two failure cases in the last row.

granularities for multimodal interactive reasoning. The reasoning process is mainly decoupled into distinct Glance-Plan and Focus-Reason stages, drawing inspiration from the human cognition habit. The two stages collaborate to first generate a preliminary plan about different sub-goals by glancing at the multimodal context, and then focus on the visual details based on corresponding sub-goals for deliberated reasoning. The sub-goal decomposition facilitates the seamless integration of the multi-granularity diagram content and domain-specific background knowledge in step-wise reasoning, mitigating the hallucinations in the multi-step reasoning process to further improve the DQA performance. Our comprehensive experiments across a diverse range of DQA benchmarks demonstrate that GlFoMR achieves SOTA performance, outperforming previous baselines. The ablation study also highlights the significant contribution of GlFoMR in empowering MLLMs to integrate detailed information for effective reasoning. Furthermore, our error analysis highlights several inherent limitations in our framework, offering valuable insight for future improvements.

Acknowledgments

This work was supported by National Key Research and Development Program of China (2022YFC3303600), National Natural Science Foundation of China (62137002, 62293550, 62293553, 62293554, 62450005, 62437002, 62477036, 62477037, 62176209, 62192781, and 62306229), "LENOVO-XJTU" Intelligent Industry Joint Laboratory Project, the Shaanxi Provincial Social Science Foundation Project (2024P041), the Natural Science Basic Research Program of Shaanxi (2023-JC-YB-593), and the Youth Innovation Team of Shaanxi Universities "Multi-modal Data Mining and Fusion".

References

[1] 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* (2023).

[2] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* 1 (2023).

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* 1, 2 (2023), 3.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24185–24198.

[6] F Chollet. 2024. Openai o3 breakthrough high score on arc-agi-pub. ARC Prize Foundation. *https://arcprize.org/blog/oai-o3-pubbreakthrough* (2024).

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[8] Ali Farhadi and Joseph Redmon. 2018. Yolov3: An incremental improvement. In *Computer vision and pattern recognition*, Vol. 1804. Springer Berlin/Heidelberg, Germany, 1–6.

[9] Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024. Visual Fact Checker: Enabling High-Fidelity Detailed Caption Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14033–14042.

[10] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. 2024. Multi-modal latent space learning for chain-of-thought reasoning in language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18180–18187.

[11] 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* (2024).

[12] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

[13] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 235–251.

[14] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*. 4999–5007.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[16] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal Reasoning with Multimodal Knowledge Graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 10767–10782.

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

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

[19] Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[20] Jie Ma, Qi Chai, Jingyue Huang, Jun Liu, Yang You, and Qinghua Zheng. 2022. Weakly supervised learning for textbook question answering. *IEEE Transactions on Image Processing* 31 (2022), 7378–7388.

[21] Jie Ma, Jun Liu, Qi Chai, Pinghui Wang, and Jing Tao. 2024. Diagram perception networks for textbook question answering via joint optimization. *International Journal of Computer Vision* 132, 5 (2024), 1578–1591.

[22] 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*. 14420–14431.

[23] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. KAM-CoT: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18798–18806.

[24] OpenAI. 2023. ChatGPT. *https://chat.openai.com*. (2023).

[25] 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*. PMLR, 8748–8763.

[26] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5689–5700.

[27] Robert L Solso, M Kimberly MacLin, and Otto H MacLin. 2005. *Cognitive psychology*. Pearson Education New Zealand.

[28] Keith Stenning and Michiel Van Lambalgen. 2012. *Human reasoning and cognitive science*. MIT Press.

[29] Christopher Summerfield and Tobias Egner. 2009. Expectation (and attention) in visual cognition. *Trends in cognitive sciences* 13, 9 (2009), 403–409.

[30] 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* (2024).

[31] Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2024. T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 19162–19170.

[32] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2609–2634.

[33] Shaowei Wang, Lingling Zhang, Xuan Luo, Yi Yang, Xin Hu, Tao Qin, and Jun Liu. 2022. Computer science diagram understanding with topology parsing. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16, 6 (2022), 1–20.

[34] Shaowei Wang, Lingling Zhang, Longji Zhu, Tao Qin, Kim-Hui Yap, Xinyu Zhang, and Jun Liu. 2024. CoG-DQA: Chain-of-Guiding Learning with Large Language Models for Diagram Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13969–13979.

[35] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* 36 (2024).

[36] Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. 2023. InstructTA: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886* (2023).

[37] Yaxian Wang, Jun Liu, Jie Ma, Hongwei Zeng, Lingling Zhang, and Junjun Li. 2023. Dynamic dual graph networks for textbook question answering. *Pattern Recognition* 139 (2023), 109441.

[38] Yaxian Wang, Bifan Wei, Jun Liu, Qika Lin, Lingling Zhang, and Yaqiang Wu. 2023. Spatial-Semantic Collaborative Graph Network for Textbook Question Answering. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 7 (2023), 3214–3228.

[39] Yaxian Wang, Bifan Wei, Jun Liu, Lingling Zhang, Jiaxin Wang, and Qianying Wang. 2023. DisAVR: Disentangled Adaptive Visual Reasoning Network for Diagram Question Answering. *IEEE Transactions on Image Processing* 32 (2023), 4812–4827.

[40] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research* (2022).

[41] Yukun Xiao, Yisheng An, Ting Li, Naiqi Wu, Wei He, and Peng Li. 2024. Autonomous driving policy learning from demonstration using regression loss function. *Knowledge-Based Systems* 295 (2024), 111766.

[42] Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11289–11303.

[43] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. 2024. Contextual object detection with multimodal large language models. *International Journal of Computer Vision* (2024), 1–19.

[44] 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* (2024).

[45] Lingling Zhang, Wenjun Wu, Jun Liu, Xiaojun Chang, Xin Hu, Yuhui Zheng, Yaqiang Wu, and Qinghua Zheng. 2025. LFSRM: Few-Shot Diagram-Sentence Matching via Local-Feedback Self-Regulating Memory. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence (2025).

[46] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024. Multimodal Chain-of-Thought Reasoning in Language Models. *Trans. Mach. Learn. Res.* 2024 (2024).

[47] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361* (2024).

[48] 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 (2023), 5168–5191.

[49] Qiji Zhou, Ruochen Zhou, Zike Hu, Panzhong Lu, Siyang Gao, and Yue Zhang. 2024. Image-of-Thought Prompting for Visual Reasoning Refinement in Multimodal Large Language Models. *arXiv preprint arXiv:2405.13872* (2024).