ProReason: Multi-Modal Proactive Reasoning with Decoupled Eyesight and Wisdom

Jingqi Zhou[♡], **Sheng Wang**[♡†], **Jingwei Dong**[♡], **Kai Liu**[♡],
Lei Li[♡], **Jiahui Gao**[♡], **Jiyue Jiang**[♠], **Lingpeng Kong**[♡], **Chuan Wu**[♡†]
[♡] The University of Hong Kong, [♠] The Chinese University of Hong Kong,
{u3011211, u3009618, u3013005, u3638070}@connect.hku.hk,
nlp.lilei@gmail.com, sumiler@connect.hku.hk,
jiangjy@link.cuhk.edu.hk, {lpk, cwu}@cs.hku.hk

Abstract

Large vision-language models (LVLMs) have witnessed significant progress on visual understanding tasks. However, they often prioritize language knowledge over image information on visual reasoning tasks, incurring performance degradation. To tackle this issue, we first identify the drawbacks of existing solutions (*i.e.*, limited multi-modal reasoning capacities, and insufficient and irrelevant visual descriptions). We then decompose visual reasoning process into two stages: proactive visual perception (*i.e.*, eyesight) and textual reasoning (*i.e.*, wisdom), and introduce a novel visual reasoning framework named ProReason. This framework features decoupled vision-reasoning capabilities and multi-run proactive perception. Briefly, given a multi-modal question, ProReason iterates proactive information collection and reasoning until the answer can be concluded with necessary and sufficient visual descriptions. Notably, the disassociation of capabilities allows seamless integration of existing large language models (LLMs) to compensate for the reasoning deficits of LVLMs. Our extensive experiments demonstrate that ProReason outperforms existing multi-step reasoning frameworks on various benchmarks for both open-source and closed-source models, with the average performance gain reaching 13.2%. Besides, the integration of LLMs allows ProReason to produce high-quality visual reasoning data, which empowers ProReason-distilled models (*i.e.*, `ProReason-VL` and `ProReason-Q3`) to achieve superior performance in downstream tasks. Our insights into existing solutions and the decoupled perspective for feasible integration of LLMs illuminate future research on visual reasoning techniques, especially LLM-assisted ones. The code is available at `https://github.com/lian-tian-mo-zun/Pro_Reason`.

1 Introduction

In recent years, large language models (LLMs) (Yang et al., 2024; Dubey et al., 2024; Team et al., 2023; Jiang et al., 2023) have experienced explosive growth in their capabilities, driving significant advancements across various fields (Shao et al., 2023; Guo et al., 2024; Shao et al., 2024). This progress has also sparked interest in developing large vision-language models (LVLMs) (Xiaomi and Team, 2025; Bai et al., 2025; Chen et al., 2024b; Bai et al., 2023), which, like LLaVA (Li et al., 2024b), have achieved remarkable performance in multi-modal understanding tasks. However, state-of-the-art (SOTA) LVLMs still struggle to handle visual understanding with textual reasoning simultaneously due to inherent modality differences. For example, Ghosh et al. (2024) demonstrate that LVLMs often rely more on their prior language knowledge, neglecting visual information in multi-modal reasoning tasks, such as visual chart understanding and math reasoning, resulting in performance degradation. Figure 2.b illustrates a typical case of this issue, where the reasoning process remains irrelevant to the image.

To address the challenges, several visual reasoning frameworks have been proposed. Specifically, (Ghosh et al., 2024) and (Mitra et al., 2024) convert visual information in images into textual descriptions to aid LVLMs in reasoning. However, their visual extraction process is not targeted at a given question (*i.e.*, question-agnostic), termed as "passive", and omits reasoning mechanisms to infer extra information for better descriptions(*i.e.*, reasoning-free). These limitations result in irrelevant or inadequate information, ultimately degrading performance. Furthermore, these frameworks are powered by a single LVLM, leading to a reasoning process that conflates visual understanding with textual reasoning abilities, failing to mitigate the challenge faced by LVLMs in effectively managing both capabilities.

To resolve these problems, we propose ProReason, a multi-modal reasoning framework featur-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31662–31691
November 4-9, 2025 ©2025 Association for Computational Linguistics

ing decoupled vision-reasoning capabilities. As illustrated in Figure 1, we decouple multi-modal reasoning capacity into two sub-tasks: proactive visual perception (*i.e.*, eyesight) and textual reasoning (*i.e.*, wisdom). The former extracts visual information in a question-oriented and reasoning-involved manner, while the latter integrates all information to draw final conclusions. Specifically, during the visual perception stage, a Dispatcher first selectively engages a Vision Expert to capture additional visual information, or an Insight Expert to derive intermediate inferences. A Referee then determines whether sufficient information is gathered to proceed to the reasoning stage, where a Summarizer produces the final answer. Unlike passive methods, all sub-agents operate based on the given question and known information, effectively avoiding irrelevant information redundancy or insufficiency. Notably, decoupled vision-reasoning eliminates the need for LVLMs to handle vision-irrelevant roles (*i.e.*, Dispatcher, Insight Expert, Referee, and Summarizer), enabling seamless integration of existing LLMs with proven strong reasoning abilities (Chang et al., 2024), thereby alleviate the limitations of LVLMs. In addition, the high-quality reasoning data generated by LLM-assisted PROREASON can be effectively distilled into downstream models for inherent performance improvement.

Empirically, we evaluate PROREASON across multiple challenging visual reasoning benchmarks with both open-source and closed-source models. Extensive experiments demonstrate that PROREASON exhibits significant advantages in two key aspects: (1) As a visual reasoning framework, PROREASON achieves consistent and substantial performance improvements across multiple benchmarks, with the average performance gain reaching 13.2%, validating the effectiveness of its decoupled vision-reasoning architecture and proactive visual feature extraction mechanism; (2) PROREASON effectively integrates existing LLMs to generate high-quality visual reasoning process, empowering PROREASON-distilled models (*i.e.*, `ProReason-VL` and `ProReason-Q3`) with superior visual reasoning capabilities. The above results, coupled with the ablation study in Section 4.5, demonstrate the substantial advantages of decoupled vision-reasoning, while highlighting the potential of LLM-assisted LVLM reasoning and distillation strategies.

The main contributions of this work are three-fold:

- We propose a novel multi-modal reasoning framework named PROREASON, featuring decoupled vision-reasoning and iterative proactive perception capabilities, effectively mitigating the drawbacks of previous methods.
- Extensive experiments consistently highlight the significant superiority of PROREASON and necessity of each component across multiple visual reasoning tasks and model series, illuminating the great potential of LLM-assisted LVLM reasoning.
- PROREASON-distilled models also exhibits remarkable enhancements over vanilla counterparts, showcasing the feasibility of LLM-assisted LVLM improvement in the future.

2 Preliminary Observations

Ghosh et al. (2024) demonstrate that the limited multimodal reasoning abilities of LVLMs lead to an overreliance on linguistic priors, thus neglecting visual inputs and ultimately degrading their performance. Their Visual Description Grounded Decoding (VDGD) mitigates visual oversight by converting images into comprehensive textual descriptions to inform reasoning processes. However, **such passive visual reasoning techniques suffer insufficient and irrelevant visual information**. To support this claim, we generate fine-grained image captions using `GPT-4o-mini`[1] with the prompt shown in Figure 9. We then incorporate these captions into the prompts for LVLMs to facilitate the reasoning process. We analyze the performance of this approach on the challenging multi-modal MMMU dataset, which requires college-level knowledge and fine-grained reasoning, using recent open-source LVLMs listed in Section 4.1 As shown in Table 4, while these image descriptions improve the performance of LVLMs, the gains are marginal, consistently amounting to less than 1%. This underscores the limited utility of captions generated by passive methods.

For further demonstration, inspired by Liu et al. (2023c), we use the prompt instructions in Figure 8 to instruct `GPT-4` to analyze the generated captions along three dimensions: Detail Level, Question Relevance, and Reasoning Effective Info Inclusion, measuring the richness of detail, relevance to the given question, and the inclusion of

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence

information that is essential for reasoning, respectively. Meanwhile, since the reasoning process of `GPT-4o-mini` on MMMU contains key information necessary for solving the problems, we use it as a reference answer to aid evaluation. As shown in Table 5, the captions for correct responses of `Llama3-LLaVA-NeXT-8B` receive higher scores across all three criteria, highlighting the importance of better captions for multi-modal reasoning. Additionally, all captions score significantly lower in the Question Relevance and Reasoning Effective Info Inclusion dimensions than the Detail Level dimension, indicating that **while the captions are detailed, they often lack relevance to the questions.** Figure 2 shows a case where `Llama3-LLaVA-NeXT-8B` utilizes fine-grained captions to solve a question from the MMMU benchmark. As illustrated, although the caption exhaustively describes the image content, it incorrectly describes the wires in the image as octagons, and misses information about the locations of these wires. This information is irrelevant to the target question, thus offering minimal assistance to LVLMs. In summary, our analysis highlights the drawbacks of passive visual reasoning enhancement techniques in terms of information insufficiency and redundancy, due to their question-agnostic property.

3 Method

As depicted in Figure 1 and exemplified in Figure 3, ProReason presents an innovative decoupling of the visual reasoning process into two distinct phases framed through the lens of LVLM capabilities: Proactive Visual Perception (*i.e.*, eyesight) and Textual Reasoning (*i.e.*, wisdom). The entire workflow consists of five functionally distinct yet inter-cooperative sub-agents, along with a Memory component, facilitating multi-modal reasoning performance.

3.1 Proactive Visual Perception

Proactive Visual Perception is the core of question-oriented visual information extraction, driven by four sub-agents: Dispatcher, Vision Expert, Insight Expert and Referee. The Dispatcher breaks down the original question, selectively directing the Vision Expert to capture specific visual information, or instructing the Insight Expert to analyze known information to derive more. The responses from both experts are stored in a textual Memory component. The Referee then evaluates whether the information stored in Memory is sufficient to answer the original question.

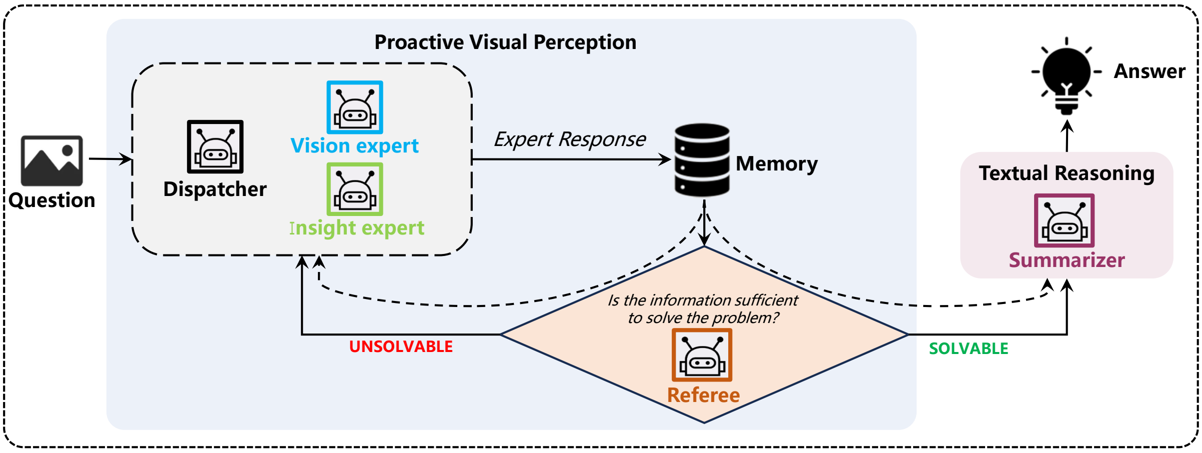
Formally, given an image I and its corresponding textual question Q, the Dispatcher decides to consult the Vision Expert or Insight Expert, based on the analysis of Q and the known information in the Memory (if not empty). The Dispatcher then generates a query q for the chosen expert. If the Vision Expert is selected, it takes the image I and query q as input, and generates an answer A_{vision}, which is then stored in the Memory. When the Insight Expert is selected, it provides a response consisting of the inference process and the final answer $A_{insight}$ based on the query q and known information in the Memory, before only $A_{insight}$ is stored in Memory. The Referee then evaluates the available information in the Memory concerning the question Q. If the Memory contains adequate information to answer the question Q, the Referee outputs the identifier "SOLVABLE"; otherwise, it outputs "UNSOLVABLE". If the Referee's output is SOLVABLE, the workflow precedes to the Textual Reasoning phase. Conversely, if the output is UNSOLVABLE, the above process will be re-executed to gather more necessary information.

In the Proactive Visual Perception phase, the Dispatcher collaborates with the Vision Expert and Insight Expert to achieve question-oriented and reasoning-involved visual information extraction, while the Referee ensures informational completeness and prevents omissions. These four sub-agents work closely together, thereby overcoming the limitations of passive methods. Notably, the Memory component allows ProReason to keep compact information, and avoids lengthy reasoning traces like CoT (Wei et al., 2022) and ReAct (Yao et al., 2022), thereby suffering less from redundant information (for a detailed analysis, refer to Section 4.4 and Table 3).

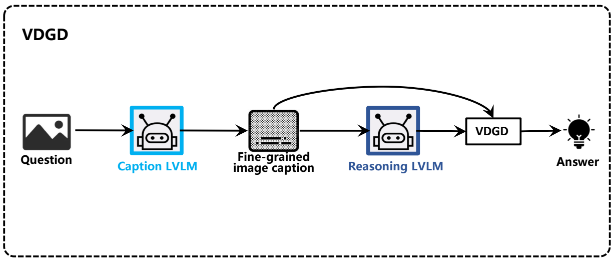
3.2 Textual Reasoning Step

The Textual Reasoning step focuses on integrating the available information in the Memory, and providing the final answer to the question Q. This step is mainly powered by a sub-agent called Summarizer. Once the Referee confirms that the Memory contains sufficient information to address question Q and outputs the SOLVABLE identifier, the Summarizer will be invoked to perform a detailed reasoning and generate a final answer based on the question Q and Memory. This final answer is then

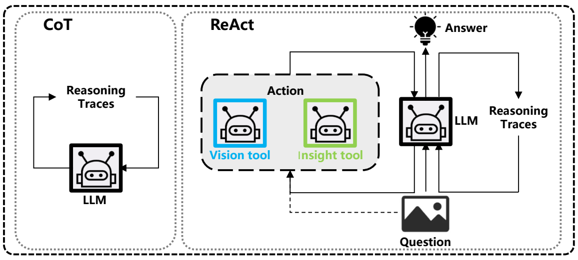
Figure 1: Overview and comparison of ProReason, VDGD and ReAct. Unlike existing works (*e.g.*, VDGD and ReAct), our proposed method decouples visual perception and textual reasoning while allowing the model to actively acquire necessary information from the images, achieving superior performance.



(1) ProReason



(2) Passive Visual Reasoning Enhancement Technique



(3) Multi-step Reasoning Framework for LLMs

evaluated using performance metrics, while the Summarizer's high-quality reasoning outputs can be utilized to train downstream models, enhancing their visual reasoning capabilities.

3.3 Advantages of ProReason

Decoupled Visual Reasoning. In ProReason, the multi-modal reasoning process is decomposed into visual perception and textual reasoning stages, each executed by separate agents. These agents are then effectively organized through a designated pipeline, significantly enhancing the ability of LVLMs to tackle visual reasoning tasks.

LLM-assisted multi-modal reasoning. Decoupled vision-reasoning eliminates the necessity for LVLMs to manage vision-irrelevant sub-agents, allowing the seamless integration of existing LLMs with established strong reasoning abilities, thereby endowing ProReason with superior visual reasoning performance.

Downstream task model enhancement. By generating high-quality data, ProReason effectively distills its superior capabilities into downstream task models, enabling them to demonstrate exceptional performance in complex visual reasoning tasks.

Reduced information Mission or Redundancy. Through sub-agent collaboration in Proactive Visual Perception, ProReason extracts essential visual details via question-oriented manner, preventing information omission or redundancy. Meanwhile, the Memory component retains only Vision Expert's observations and Insight Expert's conclusions, creating compact representation that minimizes irrelevant information. These advantages effectively reduce the overhead when ProReason functions as a reasoning framework.

4 Experiments

In this section, we first evaluate the performance of our ProReason framework against recent baselines on multiple benchmarks, followed by an in-depth ablation analysis of different components.

4.1 General Setup

Datasets. To comprehensively validate the performance of our framework, we conduct experiments across four benchmarks: Multi-modal Large Language Model Evaluation (MME) (Yin et al., 2023) [2] , Massive Multi-discipline Multi-modal Understanding and Reasoning (MMMU) (Yue et al., 2023), MathVista (Wang et al., 2024), and HallusionBench (Liu et al., 2023a). All of them require visual reasoning capabilities to complete the tasks correctly, and are introduced briefly in Section A.2.

Base Models. We employ `GPT-4o-mini`, `Llama3-LLaVA-NeXT-8B` and `Qwen2.5-VL-7B-Instruct` (Bai et al., 2025)[3] to drive PROREASON, and `Qwen2.5-72B-Instruct` (Yang et al., 2024) and `Qwen3-32B` (Team, 2025) are selected to drive the text-only sub-agents, enabling LLM-assisted visual reasoning. In addition `Qwen2.5-VL-7B-Instruct` is also utilized as a downstream task model, which is trained with the data generated by PROREASON. These models are chosen for their leading performance and popularity.

Baselines. In addition to the basic method where models are instructed to answer questions directly, we compare PROREASON with the following peer methods. First, to explore the effectiveness of directly migrating LLM solutions to LVLMs, we select two multi-step reasoning frameworks from LLMs: Chain of Thought (CoT) (Wei et al., 2022) and ReAct (Yao et al., 2022). Second, to evaluate the advantages of proactive information extraction in PROREASON, we examine two passive visual reasoning frameworks, VDGD (Ghosh et al., 2024) and CCoT (Mitra et al., 2024), which assist LVLM reasoning by extracting image information into text. Additionally, we consider `R1-Onevision-7B` (Yang et al., 2025), an LVLM based on `Qwen2.5-VL-7B-Instruct` with deep thinking capabilities similar to `Deekseek-R1` (Guo et al., 2025) and `OpenAI o1`[4]. We compare it with `ProReason-Q3`, which also exhibits deep thinking abilities. Further details on the baselines and implementation are provided in Section A.3.

Implementation Details. The prompt templates for all methods are shown in Figures 9, 10 and 11. Specifically for PROREASON, to prevent infinite loops, if the Dispatcher selects the Vision Expert or Insight Expert to obtain information up to 5 consecutive times, and the Referee still determines that the existing knowledge in Memory remains insufficient to resolve the question, the Memory will be cleared to restart the information acquisition process. If the system fail to break the loop after 5 attempts, the proactive visual perception phase will be immediately terminated. Subsequently, the Summarizer will generate the final answer based on the available information in Memory from the last attempted iteration. This setup, refined through multiple trials, is the most effective.

4.2 Main Results

PROREASON exhibits significant and consistent performance enhancement over baselines across all the benchmarks. As listed in Table 1, despite better performance than the direct method on MME dataset, VDGD and CCoT fail to demonstrate consistent improvements on the other datasets. In contrast, PROREASON consistently surpasses all other baselines across all benchmarks for every base model, enhancing the average performance of `GPT-4o-mini` by 7.9%, demonstrating the superiority and task robustness of PROREASON. Section A.5.7 and Table 13 evaluate PROREASON on additional benchmarks, including reasoning and VQA tasks, further proving its strong performance and effectiveness across various multi-modal tasks.

Decoupling the visual perception and textual reasoning capabilities of an LVLM outperforms their simultaneous inherent usage. Table 1 illustrates that CoT, utilizing both capabilities simultaneously, does not exhibit consistent performance enhancements over the "Direct" method when applied to the `Llama3-LLaVA-NeXT-8B` and `Qwen2.5-VL-7B-Instruct` models. In contrast, despite the same models, PROREASON alternates between visual information acquisition and textual reasoning processes, allowing to leverage each capability more effectively. This enables PROREASON to consistently outperform CoT with both `Llama3-LLaVA-NeXT-8B` and `GPT-4o-mini` across

[2]Due to our emphasis on visual reasoning, we select the cognition-relevant tasks, including Commonsense Reasoning, Numerical Calculation, Text Translation, and Code Reasoning. To facilitate the comparison across different benchmarks, the results for MME benchmark are calculated by the percentage of correct answers out of the total answers.

[3]The model is deployed using CUDA `https://developer.nvidia.com/cuda-toolkit` on a NVIDIA A100 graphics cards.

[4]`https://openai.com/index/introducing-openai-o1-preview/`

Table 1: Performance of multiple approaches with three base models across four visual reasoning benchmarks. “Hallu.” is the abbreviation of HallusionBench. The abbreviations 4o-mini, Qwen72B, and Qwen3 refer to GPT-4o-mini, Qwen2.5-72B-Instruct, and Qwen3-32B, respectively. "Assisted" stands for LLM-assisted reasoning, which involves replacing the textual sub-agents within frameworks with corresponding LLMs, as detailed in Table 15. Based on the performance of the direct method, red and blue signify the improvement and degradation, respectively.

Model	Method	Dataset				
		MME	MMMU	MathVista	Hallu.	Average
GPT-4o-mini	Direct	79.2	48.4	53.0	56.0	59.2
	VDGD	82.3 (+3.1)	51.4 (+3.0)	51.2 (-1.8)	52.4 (-3.6)	59.3 (+0.1)
	CCoT	80.8 (+1.6)	54.2 (+5.8)	53.6 (+0.6)	56.7 (+0.7)	61.3 (+2.1)
	CoT	87.8 (+8.6)	58.5 (+10.1)	53.8 (+0.8)	56.3 (+0.3)	64.1 (+4.9)
	ReAct	87.3 (+8.1)	54.8 (+6.4)	49.3 (-3.7)	51.1 (-4.9)	60.6 (+1.4)
	ProReason	91.9 (+12.7)	61.6 (+13.2)	54.9 (+1.9)	59.9 (+3.9)	67.1 (+7.9)
Llama3-LLaVA-NeXT-8B	Direct	61.5	41.8	37.1	45.8	46.6
	VDGD	68.8 (+7.3)	42.3 (+0.5)	36.1 (-1.0)	44.2 (-1.6)	47.8 (+1.2)
	CCoT	68.9 (+7.4)	40.5 (-1.3)	36.8 (-0.3)	37.4 (-8.4)	45.9 (-0.7)
	CoT	58.8 (-2.7)	41.5 (-0.3)	35.9 (-1.2)	43.1 (-2.7)	44.8 (-1.8)
	ReAct	68.5 (+7.0)	46.7 (+4.9)	31.7 (-5.4)	43.6 (-2.2)	47.6 (+1.0)
	+ 4o-mini Assisted	73.6 (+12.1)	48.4 (+6.6)	36.2 (-0.9)	46.7 (+0.9)	51.2 (+4.6)
	+ Qwen72B Assisted	71.0 (+9.5)	50.4 (+8.6)	34.6 (-2.5)	40.4 (-5.4)	49.1 (+2.5)
	ProReason	71.5 (+10.0)	50.5 (+8.7)	38.8 (+1.7)	50.9 (+5.1)	52.9 (+6.3)
	+ 4o-mini Assisted	84.7 (+23.2)	54.5 (+12.7)	41.7 (+4.6)	53.1 (+7.3)	58.5 (+11.9)
	+ Qwen72B Assisted	81.3 (+19.8)	56.8 (+15.0)	48.8 (+11.7)	52.3 (+6.5)	59.8 (+13.2)
Qwen2.5-VL-7B-Instruct	Direct	74.2	51.8	63.3	53.8	60.8
	VDGD	74.6 (+0.4)	52.3 (+0.5)	62.1 (-1.2)	53.9 (+0.1)	60.7 (-0.1)
	CCoT	82.7 (+8.5)	52.2 (+0.4)	61.7 (-1.6)	55.8 (+2.0)	63.1 (+2.3)
	CoT	72.3 (-1.9)	53.6 (+1.8)	63.7 (+0.4)	55.9 (+2.1)	61.4 (+0.6)
	ReAct	81.9 (+7.7)	51.6 (-0.2)	60.4 (-2.9)	52.9 (-0.9)	61.7 (+0.9)
	+ Qwen72B Assisted	83.1 (+8.9)	52.5 (+0.7)	62.2 (-1.1)	54.5 (+0.7)	63.1 (+2.3)
	ProReason	83.8 (+9.6)	57.0 (+5.2)	63.2 (-0.1)	56.4 (+2.6)	64.6 (+3.8)
	+ Qwen72B Assisted	92.7 (+18.5)	64.6 (+12.8)	64.0 (+0.7)	60.6 (+6.8)	70.5 (+9.7)
	+ Qwen3 Assisted	90.7 (+16.5)	66.2 (+14.4)	67.2 (+3.9)	59.8 (+6.0)	71.0 (+10.2)

all benchmarks, validating the effectiveness of capability decoupling.

Proactive information acquisition surpasses peer passive methods, especially in complex visual reasoning tasks. Specifically, compared to MME, MathVista and HallusionBench present higher image complexity and question difficulty, and thus require stronger visual understanding and textual reasoning capabilities. This leads to performance degradation of passive methods (*i.e.*, VDGD and CCoT), highlighting their limited applicability to complex visual reasoning tasks. In contrast, ProReason achieves notable performance improvements, up to 5.1%, by proactively acquiring visual information from images rather than generating question-agnostic captions. This aligns with out previous obervations in Section 2 that passive methods introduce substantial information redundancy or omission, misleading subsequent reasoning processes.

Text-only LLMs can be effectively integrated into ProReason for dramatically enhanced performance. As mentioned in Section 3.3, the decoupled visual perception and textual reasoning capabilities facilitate the seamless integration of text-only LLMs. To demonstrate the utility of this advantage, we fix the Vision Expert and replace the textual sub-agents in ProReason with text-only LLMs, according to the configuration in Table 15. As listed in Table 1, with the assistance of powerful existing LLMs, the Llama3-LLaVA-NeXT-8B Vision Expert receives remarkable performance boost across all benchmarks, particularly by 15% on MMMU and 11.7% on MathVista, compared to directly providing answers. In addition, by configuring the Summarizer to Qwen3-32B, ProReason acquires the same kind of deep reasoning capability as Deekseek-R1 and OpenAI o1, resulting in a 10.2% average performance improvement for the Qwen2.5-VL-7B-Instruct Vision Expert. In contrast, ReAct gains a much smaller improvement. This highlights the unique advantage of ProReason in leveraging existing text-only LLMs for enhanced performance. Notably, this advantage may open new avenues for continuously pushing the performance limits of LVLMs with the assistance of existing powerful LLMs.

Table 2: Performance of different models across two visual benchmarks. Based on the performance of the base model(*i.e.*, `Qwen2.5-VL-7B-Instruct`), red and blue signify the improvement and degradation, respectively.

Dataset	Model			
	`Qwen2.5-VL-7B-Instruct`	`R1-Onevision-7B`	`ProReason-VL`	`ProReason-Q3`
MME	72.3	88.1 (+15.8)	91.2 (+18.9)	90.8 (+18.5)
MMMU(val.)	53.6	53.1 (-0.5)	65.4 (+11.8)	67.4 (+13.8)
Average	63.0	70.6 (+7.6)	78.3 (+15.3)	79.1 (+16.1)

Table 3: Average token and time consumption of multiple approaches with `GPT-4o-mini` model on the MME and MathVista benchmarks.

Dataset	Method	`GPT-4o-mini`					
		Direct	VDGD	CCoT	CoT	ReAct	PROREASON
MME	Input	393.9	1020.9	1024.3	403.9	1645.0	1286.8
	Output	5.9	155.6	254.3	103.4	197.0	327.2
	Time(s)	6.1	12.3	12.8	7.3	18.9	18.4
MathVista	Input	368.4	955.5	961.2	375.4	3092.8	2238.6
	Output	51.8	263.3	307.3	479.7	845.1	788.6
	Time(s)	4.2	11.5	12.3	12.0	28.8	24.8

4.3 Downstream Task Model Enhancement

High-quality visual reasoning data produced by PROREASON significantly improves the performance of downstream task models. We select the test set that has no answer in the MMMU dataset and gather reasoning processes paired with corresponding answers on this set, generated by two configurations detailed in Table 15: PROREASON + `Qwen72B` Assisted and PROREASON + `Qwen3` Assisted. After eliminating samples with inconsistent answers between these two configurations, we obtain two filtered datasets each containing 5,980 entries. These datasets are then used to fine-tune `Qwen2.5-VL-7B-Instruct` (Section A.4), yielding two models: `ProReason-VL` and `ProReason-Q3`. Notably, `ProReason-Q3` inherits the deep reasoning capabilities of `Qwen3-32B`, which incorporates a reasoning process containing a <think>... </think> part.

Table 2 illustrates that while `R1-Onevision-7B` features deep reasoning capabilities and shows performance enhancements on MME, it fails to demonstrate improvement on the MMMU validation set. In contrast, both `ProReason-VL` and `ProReason-Q3` achieve performance gains exceeding 10% on the MMMU validation set and also exhibit significant improvements on MME. This indicates that PROREASON effectively transfers the exceptional reasoning capabilities of LLM-integrated decoupled systems to downstream task models, highlighting the potential for leveraging the robust abilities of existing LLMs to continuously enhance the performance of LVLMs.

4.4 Efficiency and Complexity Analysis

PROREASON improves both efficiency and performance compared to the baseline method . Table 3 presents the evaluation of average token consumption and time expenditure of different methods on the MME and MathVista datasets. Notably, PROREASON requires significantly fewer tokens than ReAct, while achieving superior performance over ReAct as analyzed in Section 4.2. Moreover, compared to the visual reasoning frameworks VDGD and CCoT, which also involve multiple image inputs, PROREASON's token consumption on MME is only about 20% higher, yet it achieves a 12.7% performance improvement. Regarding time efficiency, PROREASON achieves 11.1% higher performance than VDGD and CCoT on MME with comparable time consumption (18.4s/sample vs. VDGD's 12.3s and CCoT's 12.8s), while ReAct is slower than PROREASON across datasets, highlighting PROREASON's efficiency advantage. Furthermore, as discussed in Section 4.3, PROREASON's robust visual reasoning capabilities can be transferred to downstream task models, further ensuring the performance and efficiency of our approach.

PROREASON achieves a balance between performance and efficiency through its Memory component and adaptive mechanisms . The Memory component enables PROREASON to keep compact information representations, avoiding lengthy reasoning traces like ReAct, reducing token use and boosting efficiency. Additionally, PROREASON dynamically adjusts Proactive Visual Perception iterations based on question difficulty, minimizing overhead for simple tasks while enhancing complex problem-solving. As shown in Table 12, MathVista's greater challenge prompts more iterations compared to MME, resulting in higher token usage and longer reasoning times. This aligns with the difficulty levels of the datasets, demonstrating the adaptive nature of PROREASON.

4.5 Ablation Study and Further Analysis

Relative Importance of Sub-agents. To evaluate the importance of each sub-agent in PROREASON, in Section A.5.1, we replace Dispatcher, Vision Expert, Insight Expert, Referee, and Summarizer individually with the less capable `Llama3-LLaVA-NeXT-8B`(which demonstrates weaker visual understanding and textual reasoning capabilities), while keeping the other sub-agents as `GPT-4o-mini`. The performance degradation on the MME and MMMU benchmarks is then used to measure the significance of each sub-agent. Results indicate that the Summarizer is the most critical sub-agent, closely followed by Referee.

Which One is More Crucial: Visual Understanding or Textual Reasoning In Section A.5.2, we perform comparative experiments by substituting the vision expert and text sub-agents in the PROREASON with `Llama3-LLaVA-NeXT-8B` and `GPT-4o-mini` respectively. Our findings indicate that while both visual understanding and textual reasoning capabilities are essential for multimodal tasks, textual reasoning ability holds greater significance in visual reasoning tasks. This result is consistent with our earlier analysis in Section A.5.1, which identifies the Summarizer and Referee as the most critical sub-agents.

The Critical Implication of Decoupling. In Section A.5.3, to validate the necessity of decoupling visual perception and textual reasoning in PROREASON, we systematically merge sub-agents through three configurations while preserving identical prompts and procedures, with observed performance degradation quantitatively demonstrating the critical role of decomposed processing in enhancing capabilities. The experimental results indicate that Decoupling serves as a crucial mechanism for improving PROREASON's performance in complex visual reasoning tasks

Reasoning Process Evaluation of PROREASON. In Section A.5.4, we evaluate the responses generated by PROREASON using LLMs. The analysis reveals that, compared to CoT, PROREASON produces more relevant answers with reduced redundancy and deficiency, consistent with its enhanced performance.

Referee's Dispel of Hallucinations. In Section A.5.5, adhering to the settings in Section 4.1, we assess PROREASON (powered by `GPT-4o-mini`) on MMMU and HallusionBench with different attempt allowances (1/3/5). When attempts are unsuccessful, systematic Memory clearance is triggered (determined by the Referee's lack of sufficient information). As the number of attempts increases, the Referee has more opportunities for information filtering. The observed performance variations illustrate the crucial influence of the Referee's decision - making and filtering efficiency on the system's capabilities. Experimental results indicate that the Referee module effectively filters hallucinated information to improve the visual comprehension capabilities of our framework.

Frequency of selection of various experts. In Section A.5.6, we assess how often the Dispatcher selects the Vision Expert or Insight Expert across both MME and MMMU benchmarks. The experimental results demonstrate that PROREASON adaptively adjusts the frequencies of expert selection, leading to consistent performance improvements.

5 Conclusion

In this paper, we first validate that existing multimodal reasoning approaches still suffer insufficient and irrelevant visual descriptions, as well as limited multi-modal capacities. To address these issues, we decompose the visual reasoning process into visual perception and textual reasoning stages, and introduce a novel visual reasoning framework named PROREASON, featuring decoupled vision-reasoning capabilities and multi-run proactive perception. Empirically, extensive experiments demonstrate the superiority of PROREASON over both passive image information acquisition methods and multi-step reasoning frameworks for text-only LLMs across multiple visual reasoning benchmarks with both open-source and closed-source models. Notably, our method showcases the remarkable feasibility of integrating LLMs for multi-modal reasoning with dramatically improved performance, highlighting the great potential for LLM-assisted LVLM reasoning in future research.

6 Limitations

In this section, we analyze the limitations of the proposed method based on typical errors made by `GPT-4o-mini`-driven PROREASON, as exemplified in Figures 6 and 7, to gain further understanding and identify potential research directions.

Cumulative Errors. As illustrated in Figure 6, the vision expert mistakenly perceives the clock as 6:25, which misguides the reasoning of subsequent agents and ultimately leads to an incorrect conclusion. More broadly, similar misperceptions occur frequently in errors made by PROREASON. This indicates that, with the assistance of LLMs, PROREASON has effectively addressed the reasoning deficiencies in multi-modal tasks, while the vision expert plays a significant role for further improvement of multi-modal capabilities.

Contradictory Information among Agents. Considering that multiple agents are engaged in the answering process, we try to find instances where contradictory information is provided by different agents, especially the vision expert and Insight Expert. However, as shown in Figure 6 and 7, when one agent (*e.g.*, the vision expert) makes an error and the referee even hints at a possible mistake, other agents (*e.g.*, the Insight Expert) tend to adhere to the available information instead of questioning it. This tendency results in a failure to find cases with contradictory information, and also highlights the importance of a reflection mechanism (Ji et al., 2023) in agent collaboration, which is left for future exploration.

We also examined existing multi-step reasoning frameworks, such as ReAct (Yao et al., 2022), ToT (Yao et al., 2024), and Insight-V (Dong et al., 2024), aiming to find solutions for resolving accumulated errors and contradictory information. However, we found that these approaches may also fail to identify effective solutions and **do not discuss the impact of the aforementioned drawbacks**. We sincerely WELCOME any constructive discussions regarding accumulated errors and contradictory information! **Besides, ProReason demonstrates significantly improved overall performance, suggesting fewer errors made by ProReason in the whole task distribution.**

7 Ethics Statement

We adhere strictly to the ACL Code of Ethics throughout our research. To our knowledge, the methods we introduce pose no foreseeable risks. We provide comprehensive details of the computing infrastructure used for all computational experiments in the paper, along with transparent statistics on our results and a detailed configuration of our experimental setup, including the optimal hyperparameter values. Furthermore, we will release the code upon publication to facilitate easy public reproducibility.

8 Acknowledgements

We want to thank our anonymous AC and reviewers for their feedback. This work was supported in part by Hong Kong Innovation and Technology Commission's Innovation and Technology Fund (Award No. ITS/269/22FP), Hong Kong RGC grants C7004-22G (CRF) and CRS_PolyU501/23 (CRS).

References

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*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. 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*, pages 24185–24198.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2024. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*.

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*.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. Vdgd: Mitigating lvlm hallucinations in cognitive prompts by bridging the visual perception gap. *arXiv preprint arXiv:2405.15683*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *Preprint*, arXiv:2401.14196.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

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

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning.

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

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *arXiv preprint arXiv:2206.01718*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2025. Qwen3.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

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

LCT Xiaomi and Core Team. 2025. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024b. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Zhehao Zhang, Ryan Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, and Nedim Lipka. 2024c. Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use. *arXiv preprint arXiv:2410.16400*.

A Appendix

A.1 Related work

Large Visual-Language Model. Recently, large vision-language models (LVLMs) (Yang et al., 2024; Bai et al., 2023; Chen et al., 2023; Liu et al., 2024b) have garnered widespread attention and demonstrated remarkable advancements in understanding and generating multi-modal contents. In the open-source domain, numerous LVLMs, like LLaVA (Liu et al., 2023c,b, 2024a; Li et al., 2024a,b) and InternVL (Chen et al., 2024c) families, have been extensively developed. In the closed-source domain, proprietary models such as GPT-4o[5] and Gemini Pro 2.5[6] have also achieved significantly success. Additionally, multi-agent frameworks like VipAct (Zhang et al., 2024c) have been developed to improve LVLMs' perception of visual details. Despite these advancements, existing LVLMs still encounter challenges in effectively integrating visual understanding with textual reasoning capabilities simultaneously. This limitation is particularly evident in their diminished attention to image content during visual reasoning process, such as chart interpretation and visual math reasoning, leading to degraded performance (Liu et al., 2023a; Ghosh et al., 2024) and motivating more effective solutions.

Multi-step Reasoning Framework. Multi-step reasoning frameworks improve LLM performance by breaking down complex tasks. Chain-of-Thought (CoT) (Wei et al., 2022) enhances reasoning via explicit intermediate steps, demonstrating effectiveness in both textual and visual tasks (Zhang et al., 2024b), while Tree-of-Thoughts (ToT) (Yao et al., 2024) extends this by evaluating multiple reasoning paths. ReAct (Yao et al., 2022) integrates dynamic knowledge retrieval during reasoning. In visual reasoning, several frameworks assist LVLM by extracting image information into text. Visual Description Grounded Decoding (VDGD) (Ghosh et al., 2024) describes the image and appends this description to the prompt, aiding LVLMs in visual reasoning tasks. Compositional Chain-of-Thought (CCoT) (Mitra et al., 2024) guides LVLMs to create scene graphs (SGs) that link visual and textual domains, supporting subsequent tasks. However, these methods often use a question-agnostic, reasoning-free visual extraction process, resulting in irrelevant or redundant information. Insight-V (Dong et al., 2024) trains LVLM with a multi-agent system, but all agents rely on a single LVLM, blending visual understanding with textual reasoning and failing to address the challenge of effectively managing both. Unfortunately, this method does not release its code and prompts, resulting in low reproducibility and making comparison difficult. In response to these drawbacks, we introduce PROREASON, which decouples visual reasoning tasks into proactive visual perception (i.e., eyesight) and textual reasoning (i.e., wisdom), and makes all prompts available. By leveraging the strengths of the decoupled system, PROREASON effectively integrates existing powerful LLMs to achieve high-performance visual reasoning and successfully transfers this capability to downstream task models.

A.2 Dataset

To thoroughly assess the performance of our framework, we have carried out experimental evaluations using four benchmark datasets: the Multi-modal Large Language Model Evaluation (MME) (Yin et al., 2023), the cross-disciplinary Massive Multimodal Understanding and Reasoning benchmark (MMMU) (Yue et al., 2023), the visual mathematical reasoning assessment MathVista (Wang et al., 2024), and the multimodal illusion detection benchmark HallusionBench (Liu et al., 2023a). Each of these benchmarks necessitates strong visual reasoning capabilities for successful task completion, and we provide concise descriptions below:

- **MME** is an inclusive benchmark that encompasses 14 subtasks, designed to evaluate perceptual and cognitive abilities. Given our focus on visual reasoning, we select the cognition-relevant tasks, including Commonsense Reasoning, Numerical Calculation, Text Translation, and Code Reasoning.

- **MMMU** evaluates multi-modal models with multidisciplinary tasks that require college-level domain-specific knowledge and detailed reasoning. It comprises 11,500 questions across 30 disciplines and 183 sub-fields, emphasizing advanced perception and domain-specific reasoning.

- **MathVista** focuses on more challenging mathematical reasoning tasks that demand precise

[5] https://openai.com/index/hello-gpt-4o/

[6] https://deepmind.google/technologies/gemini/pro/

visual recognition and compositional reasoning. It includes 6,141 examples from 31 multimodal mathematics datasets.

- **HallusionBench** evaluates models' ability to reason with images such as statistical charts, emphasizing nuanced visual understanding. It consists of 346 images paired with 1,129 questions, meticulously crafted by experts.

A.3 Baselines

- **Direct.** As indicated by the name, models are required to answer questions directly without dedicated prompts. This baseline is set to evaluate the initial performance of base models.

- **CoT.** CoT is an advanced prompting method that encourages LLMs to break complex tasks down into a series of easy steps, which has been applied broadly and verified to boost the reasoning performance remarkably (Chu et al., 2023).

- **ReAct.** ReAct is an LLM-specific agent framework, which performs tasks by alternating between reasoning and execution behaviors. To extend it to multi-modal domain, we use two LVLMs to perform both steps, and rename them as the Vision and Insight Experts, respectively. This aligns with our notions for easy understanding, and is shown in Figure 1.

- **VDGD.** VDGD involves two main steps: initially, LVLMs generate detailed image captions, which are then incorporated into prompts to aid inference. During the inference process, VDGD also utilizes a formula based on Kullback-Leibler divergence to select tokens that minimally deviate from the description, thereby enhancing the relevance of the model's reasoning to the image[7].

- **CCoT.** Given an image and the question, CCOT first generates a scene graph of the image with LVLMs, and then extracts the answer by prompting the LVLMs with the graph.

- **R1-Onevision.** `R1-Onevision-7B` is an advanced multimodal reasoning model based on `Qwen2.5-VL-7B-Instruct` that has deep thinking capabilities akin to Deepseek-R1 by transforming images into structured textual representations and employing a training framework that merges supervised fine-tuning with reinforcement learning.

[7] Since we cannot obtain the tokens output by `GPT-4o-mini`, we omit the step of selecting the token with the smallest deviation from the image description when implementing VDGD for `GPT-4o-mini`.

A.4 Model Training

we use Supervised Fine-Tuning(SFT) and employ the parameter-efficient fine-tuning method **LoRA** (Hu et al., 2021). Specifically, we uniformly set the learning rate to 1×10^{-4}, `lora_dropout` $= 0$, and train the `ProReason-VL` for 1 epoch and `ProReason-Q3` for 2 epoch. These parameters are the optimal values obtained after multiple attempts.

A.5 Supplementary Results and Analysis

Table 4: Performance of three recent LVLMs on MMMU dataset with different assisting techniques.

Model	Method		
	Direct	CoT	VDGD
`Llama3-LLaVA-NeXT-8B`	41.8	41.5	42.7
`Qwen2.5-VL-7B-Instruct`	51.8	52.2	52.6

Table 5: Effectiveness evaluation of passive captions along Detail Level, Question Relevance, and Reasoning Effective Info Inclusion. "True" and "False" denote the response correctness of `Llama3-LLaVA-NeXT-8B`.

Score	`Llama3-LLaVA-NeXT-8B`	
	True	False
Detail Level	4.43	3.93
Question Relevance	3.87	3.30
Reasoning Effective Info Inclusion	3.91	3.57

A.5.1 Relative Importance of Sub-agents

To assess the importance of each sub-agent within the ProReason framework for visual reasoning tasks, we design five scenarios where `Llama3-LLaVA-NeXT-8B` acts as Dispatcher, Vision Expert, Insight Expert, Referee, or Summarizer, respectively, while the other sub-agents are powered by `GPT-4o-mini`. Given that `Llama3-LaVA-NeXT-8B` exhibits weaker visual understanding and textual reasoning capabilities than `GPT-4o-mini`, the more significant the performance drop incurred by replacing a sub-agent with `Llama3-LaVA-NeXT-8B` is, the more important that sub-agent is. Here we primarily consider the MME and MMMU benchmarks due to their comprehensive question coverage. The experimental results are presented in Table 6.

Summarizer is the most crucial sub-agent, closely followed by Referee. The replacement of Summarizer results in the most notable performance decline on both MME and MMMU tasks, reaching 6.2% and 10.6%, respectively. This highlights the critical function of the Summarizer in integrating all available information to conclude final answers. Besides, the substitution of Referee leads to a 10.1% reduction on MMMU. Given that MMMU is more challenging than MME, this finding underscores the essential role of the Referee in assessing the sufficiency of information, particularly in more complex visual reasoning tasks. The analysis in Section A.5.5 also demonstrates that the Referee plays a crucial role in enabling PROREASON to accurately interpret visual detail information.

Relatively, Dispatcher and Insight Expert are the least essential sub-agents. Specifically, despite a decline, these two sub-agents exhibit significantly less performance degradation than other sub-agents. This can be attributed to the easier task of the Dispatcher, which requires minimal textual reasoning capabilities, and the infrequent calls of the Insight Expert, which is only activated when additional information needs to be inferred—a situation that is rare in current benchmarks. Besides, both sub-agents operate within the acquisition loop, allowing for greater error tolerance. Even if some error occurs, subsequent iterations can compensate for the missing information.

In summary, each sub-agent contributes to the performance of PROREASON, underscoring their necessity. Relatively, the Summarizer and Referee are the most critical sub-agents, while the Dispatcher and Insight Expert have the least impact.

A.5.2 Which One is More Crucial: Visual Understanding or Textual Reasoning?

PROREASON effectively decouples the visual understanding and textual reasoning capabilities of LVLMs. However, it remains unclear which of these two capacities is more critical for visual reasoning tasks. To answer this question, we conduct comparative experiments of the following three scenarios:

- **`Llama3-LLaVA-NeXT-8B` as All Sub-Agents**. All sub-agents within PROREASON framework are performed by `Llama3-LLaVA-NeXT-8B` model.

Table 6: Performance of PROREASON across five scenarios for sub-agent assessment on visual reasoning tasks. For each scenario, one sub-agent is replacing with `Llama3-LLaVA-NeXT-8B`, while the others are performed by `GPT-4o-mini`. The blue text indicates the performance decline compared to the scenario with all agents performed by `GPT-4o-mini`.

Model	Agent	Dataset	
		MME	MMMU
`GPT-4o-mini`		90.4	61.6
`Llama3-LLaVA-NeXT-8B`	Dispatcher	88.8 (-1.6)	60.9 (-0.7)
	Vision Expert	84.7 (-5.7)	54.5 (-7.1)
	Insight Expert	88.7 (-1.7)	60.2 (-1.4)
	Referee	89.6 (-1.1)	51.5 (-10.1)
	Summarizer	84.2 (-6.2)	51.0 (-10.6)

- **`GPT-4o-mini` as Vision Expert**. Based on the above scenario, we implement the Vision Expert with `GPT-4o-mini`, while keep the other textual sub-agents unchanged.
- **`GPT-4o-mini` as Textual Sub-Agents.** Reversely, we utilize `Llama3-LLaVA-NeXT- 8B` as the Vision Expert, and `GPT-4o-mini` for the other vision-irrelevant sub-agents.

Textual reasoning capabilities outweigh visual understanding for multi-modal reasoning tasks, although both are important. As shown in Table 7, replacing either the Vision Expert or the other agents with the more capable `GPT-4o-mini` achieves consistent performance enhancement, highlighting the significance of both capabilities. However, substituting the textual sub-agents with `GPT-4o-mini` results in a more substantial performance boost compared to replacing the Vision Expert. This underscores the greater importance of textual reasoning over visual understanding for multimodal reasoning tasks, aligning with our previous analysis in Section A.5.1 that identifies the Summarizer and Referee as the most crucial sub-agents.

A.5.3 The Critical Implication of Decoupling

To validate the critical implication of decoupling visual perception and textual reasoning in PROREASON while eliminating potential confounding factors from prompt engineering and multiple CoT implementations, we systematically integrate sub-agents through three configurations:

- Merge the Vision Expert and Insight Expert into a single sub-agent to examine the neces-

Table 7: Performance of ProReason with different configurations for the relative importance assessment between visual understanding and textual reasoning capabilities on visual reasoning tasks. The red text highlights the performance improvements brought about by the introduction of GPT-4o-mini.

Model	Agent	Dataset	
		MME	MMMU
GPT-4o-mini	Textual Sub-Agents	84.7 (+13.2)	54.5 (+2.0)
	Vision Expert	77.8 (+6.3)	53.4 (+0.9)
Llama3-LLaVA-NeXT-8B	All Sub-Agents	71.5	52.5
	COT	58.8	41.5
	Direct	61.5	41.8

sity of modality decoupling during sub-task execution.

- Integrate the Dispatcher, Vision Expert, Insight Expert, and Referee as a unified agent to verify the essentiality of the process design for the Proactive Visual Perception stage.
- Fully consolidate all five original sub-agents to demonstrate the pivotal role of decomposing visual reasoning tasks into distinct Proactive Visual Perception and Textual Reasoning phases.

The merged agents preserve identical prompts and maintain the same task execution procedures as their original counterparts, with their prompts shown in Figures 13, 14, and 15. Therefore, the performance degradation caused by agent merging quantitatively demonstrates the critical implication of decoupled processing in enhancing ProReason's capabilities.

Decoupling serves as a crucial mechanism for improving ProReason's performance in complex visual reasoning tasks . As shown in Table 8, merging Vision and Insight Experts results in a 3% performance drop on the MMMU benchmark, while combining Proactive Visual Perception with Textual Reasoning leads to a more significant 4.8% reduction (56.8 vs. CoT's 58.5). Although agent merging also causes performance declines on MME, the merged versions still outperform CoT. Notably, given MMMU's substantially higher complexity compared to MME, these findings reveal that the decoupling of visual-textual processing fundamentally drives its performance gains in complex scenarios.

Table 8: Impact of decoupling visual perception and textual reasoning on ProReason performance with results shown from sub-agent consolidation experiments. The blue text highlights the performance degradation due to sub-agent integration.

GPT-4o-mini	Dataset	
	MMMU	MME
ProReason	61.6	91.9
CoT	58.5	87.8
Vision & Reasoning Expert Integration	58.6 (-3.0)	90.4 (-1.5)
Dispatcher, Vision & Insight Expert, & Referee Integration	57.8 (-3.8)	90.8 (-1.1)
All Sub-Agents Integration	56.8 (-4.8)	89.2 (-2.7)

A.5.4 Reasoning Process Evaluation of ProReason

Inspired by Liu et al. (2023c), we design a pipeline using LLMs to further analyze the performance of ProReason. Given that GPT-4o achieved a 70.3% score on MMMU, significantly surpassing GPT-4o-mini's 58.5% and ProReason driven by GPT-4o-mini's 61.6%, we adopt GPT-4o's CoT-based answers on MMMU as the standard answer. Using this standard answer, we evaluate the reasoning process of GPT-4o-mini (CoT), the Memory of ProReason powered by GPT-4o-mini (ProReason-Memory), and the reasoning process of ProReason's Summarizer (ProReason-Summarizer) on MMMU. This assessment focuses on three key metrics: the relevance to standard answer (RE ↑), the degree of redundant information (RI ↓), and the extent of missing information (MI↓), where arrows indicate the directions of improvement. The evaluation process is driven by GPT-4o and the prompt is shown in Figure 12.

Specifically, compared to GPT-4o-mini (CoT), **ProReason-Summarizer produces more relevant answers with less redundancy and deficiency**, aligning with its improved performance. Compared to ProReason-Summarizer, ProReason-Memory exhibits the same RE, higher RI and lower MI scores. This suggests that ProReason allows some redundancy to prevent information loss in memory, as the former typically leads to more serious consequences than the latter. Subsequently, Summarizer can leverage its powerful reasoning capabilities to select the most relevant memory.

A.5.5 Referee's Dispel of Hallucinations

Following the implementation details outlined in Section 4.1, we evaluate the PROREASON system powered by `GPT-4o-mini` on both MMMU and HallusionBench under varying attempt allowances (*i.e.*, 1, 3, and 5 attempts). Every unsuccessful attempt reflects the Referee's persistent determination that the information stored in Memory is insufficient to solve the problem. Before each new attempt, Memory is systematically cleared, ensuring the removal of information deemed irrelevant by the Referee's assessment. Consequently, increased attempt allowances essentially empower the Referee with enhanced opportunities for information filtration. The observed performance variations of PROREASON across different attempt quotas demonstrate the critical impact of Referee's decision-making mechanism and information filtering efficacy on system capability.

The Referee module effectively filters hallucinated information to enhance the visual comprehension capabilities of our framework . As demonstrated in Table 10, the performance improvement on HallusionBench (2.8%) significantly outpaces that on MMMU (1.5%) as attempt opportunities increase from 1 to 5. Given HallusionBench's dual emphasis on reasoning proficiency and precise evaluation of visual details/hallucination control compared to MMMU, these results suggest that the Referee mechanism can effectively identify erroneous or irrelevant visual information, thereby strengthening PROREASON's capacity for meticulous visual understanding.

Table 9: Performance of PROREASON driven by `GPT-4o-mini` assessed by LLMs compared to CoT on MMMU benchmark. Mainly includes three key metrics: the relevance to standard answers (RE ↑), the degree of redundant information (RI ↓), and the extent of missing information (MI ↓), where arrows indicate the directions of improvement.

`GPT-4o-mini`	Metrics		
	RE ↑	RI ↓	MI ↓
COT	4.67	3.33	1.40
PROREASON-Memory	4.83	3.66	1.17
PROREASON-Summarizer	4.83	2.88	1.33

Table 10: Impact on HallusionBench and MMMU performance across different attempt allowances.

Dataset	Attempt Allowances		
	1	3	5
MMMU	60.1	59.9	61.6
HallusionBench	57.1	58.9	59.9

A.5.6 Frequency of selection of various experts

As listed in Table 11, we evaluate the frequency of the Dispatcher choosing the Vision Expert or Insight Expert on both MME and MMMU benchmarks, with MMMU requiring higher visual and reasoning abilities. Specifically, compared to MME, the frequencies for both the Vision and Insight Experts are higher on the MMMU benchmark, aligning with their difficulty levels. Together with the results in Table 1 of our submission, **PROREASON can adaptatively increase the frequencies of experts, and provide consistent performance improvements** (*i.e.*, 11.2% and 13.2%). Despite the lower frequency of the Insight Expert, the significant performance enhancement highlights the importance of LLM-assisted reasoning capabilities for reasoning-essential questions. Additionally, the frequency of the Vision expert exceeding 1 underscores the importance of referees, which controls the loop to call experts multiple times, alleviating the issue of insufficient information.

Table 11: Frequency of the Dispatcher choosing the Vision Expert or Insight Expert on both MME and MMMU benchmarks.

Dataset	`GPT-4o-mini`	
	Vision Expert	Insight Expert
MME	1.16	0.12
MMMU	1.64	0.38

A.5.7 Performance Evaluation of PROREASON on Supplementary Datasets

To further assess the effectiveness of PROREASON across diverse visual tasks, we expand upon the four datasets introduced in Section 4.1 by incorporating three additional benchmarks: MathVerse (Zhang et al., 2024a), MMStar (Chen et al., 2024a), and A-OKVQA (Schwenk et al., 2022).

MathVerse focuses on visual reasoning within mathematical contexts, while MMStar highlights visual dependency and data reliability. **A-OKVQA, in contrast, serves as a knowledge-based VQA dataset requiring minimal reasoning.** For MathVerse, we utilize the Text Lite subset, which prioritizes visual reasoning by minimizing textual content. In MMStar, subsets related to coarse and fine-grained perception that are unrelated to reasoning are excluded. Table 13 presents PROREASON's performance on each dataset, showcasing significant improvements across various forms of multi-modal tasks and **highlighting the effectiveness and generalizability of the perception and reasoning decoupling approach**.

A.5.8 The Impact of Iterative Reasoning on PROREASON's Performance

As shown in the Table 14, we evaluated the performance of PROREASON, powered by `GPT-4o-mini`, on the MME, MMMU, and MathVista datasets, given different maximum loop iterations. The performance of PROREASON is observed to enhance, as the maximum iterations increase up to three, after which the performance tends to stabilize. This observation indicates that most questions in the MME, MMMU, and MathVista benchmarks can be effectively addressed within three cycles of question-oriented visual information extraction. **In cases of a few difficult questions, PROREASON can adaptively extend the number of iterations**, proactively gathering more information from the image to arrive at the solution.

Table 12: Number of iterations required for each problem during the Proactive Visual Perception phase on MME and MathVista benchmarks.

Model	Dataset	
	MME	MathVista
`GPT-4o-mini`	1.28	2.13

Table 13: Performance of multiple approaches with GPT-4o-mini across 7 visual benchmarks. “Hallu.” is the abbreviation of HallusionBench. Based on the performance of the direct method, red and blue signify the improvement and degradation, respectively.

Model	Method	Dataset						
		MME	MMMU	MathVista	Hallu.	MathVerse	MMStar	A-OKVQA
GPT-4o-mini	Direct	79.2	48.4	53.0	56.0	28.2	46.2	78.6
	VDGD	82.3 (+3.1)	51.4 (+3.0)	51.2 (-1.8)	52.4 (-3.6)	30.1 (+1.9)	47.0 (+0.8)	79.4 (+0.8)
	CCoT	80.8 (+1.6)	54.2 (+5.8)	53.6 (+0.6)	56.7 (+0.7)	29.2 (+1.0)	45.4 (-0.8)	79.2 (+0.6)
	CoT	87.8 (+8.6)	58.5 (+10.1)	53.8 (+0.8)	56.3 (+0.3)	28.9 (+0.7)	47.2 (+1.0)	80.9 (+2.3)
	ReAct	87.3 (+8.1)	54.8 (+6.4)	49.3 (-3.7)	51.1 (-4.9)	30.4 (+2.2)	46.7 (+0.5)	80.6 (+2.0)
	PROREASON	91.9 (+12.7)	61.6 (+13.2)	54.9 (+1.9)	59.9 (+3.9)	31.6 (+3.4)	49.1 (+2.9)	81.3 (+2.7)
	Average	84.63	54.82	52.63	55.4	29.7	46.9	80.0

Table 14: Performance of ProReason with GPT-4o-mini across MME, MMMU, and MathVista benchmarks, given different maximum loop iterations. The numbers in parentheses signify the improvement or degradation over the direct method, with red for improvement and blue for degradation.

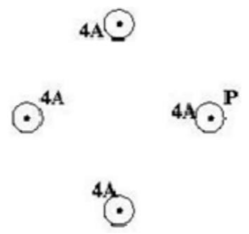
Dataset	Direct	VDGD	CCoT	CoT	ReAct	ProReason (max loop iterations)				
						1	2	3	4	5
MME	79.2	82.3 (+3.1)	80.8 (+1.6)	87.8 (+8.6)	87.3 (+8.1)	90.8 (+11.6)	91.5 (+12.3)	91.2 (+12.0)	91.5 (+12.3)	91.9 (+12.7)
MMMU	48.4	51.4 (+3.0)	54.2 (+5.8)	58.5 (+10.1)	54.8 (+6.4)	59.2 (+10.8)	60.8 (+12.4)	61.3 (+12.9)	61.1 (+12.7)	61.6 (+13.2)
MathVista	53.0	51.2 (-1.8)	53.6 (+0.6)	53.8 (+0.8)	49.3 (-3.7)	52.3 (-0.7)	53.8 (+0.8)	54.2 (+1.2)	54.5 (+1.5)	54.9 (+1.9)

Table 15: Different model configurations of Textual Sub-Agents for implementing LLM-assisted visual reasoning in PROREASON. The abbreviations 4o-mini, Qwen72B, and Qwen3 refer to GPT-4o-mini, Qwen2.5-72B-Instruct, and Qwen3-32B, respectively. "Assisted" stands for LLM-assisted reasoning.

Model	Method	Vision Expert	Summarizer	Other Textual Sub-Agents
Llama3-LLaVA-NeXT-8B	PROREASON + 4o-mini Assisted	Llama3-LLaVA-NeXT-8B	GPT-4o-mini	GPT-4o-mini
	PROREASON + Qwen72B Assisted	Llama3-LLaVA-NeXT-8B	Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct
Qwen2.5-VL-7B-Instruct	PROREASON + Qwen72B Assisted	Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct
	PROREASON + Qwen3 Assisted	Qwen2.5-72B-Instruct	Qwen3-32B	Qwen2.5-72B-Instruct

A.6 Demonstrative Examples

Question:
For the four identical current-carrying wires shown (with conventional current coming out of the plane of the page), the wire on the right is labeled P. What is the direction of the total magnetic force on the wire labeled P that is caused by the other wires?
(A) To the left (B) To the right
(C) Towards the top of the page
(D) There is no force



a. Fine-grained caption based answer:
The image consists of a diagram that features several labeled circles arranged in a specific layout. Here are the details of the contents:
1. Shapes: There are **four identical octagonal** shapes and **one circular shape**.
2. Labels:
- Four of the octagons are labeled "4A."
- One of the shapes, which appears to be a circle or another recognizable form, is labeled "P."
3. Arrangement: The shapes are positioned in a symmetrical manner. The **octagonal shapes** labeled "4A" are located in the four corners surrounding the central shape labeled "P."
The overall composition gives a sense of structure, likely serving a purpose in a scientific, mathematical, or graphical context.

Answers based on caption: The answer is **D**.

b. Answer of Chain-of-thought:
The image shows four identical wires carrying current, with the rightmost wire labeled P. To find the direction of the magnetic force on wire P, use the right-hand rule.

Point your thumb in the direction of the current in one wire, your index finger in the direction of the current in another wire, and your middle finger in the direction of the magnetic force on the third wire. Your remaining three fingers will point in the direction of the magnetic force on the fourth wire.

Applying this to the three wires not labeled P, if the thumb, index finger, and middle finger all point to the right, the magnetic force on wire P is also to the right. Thus, the correct answer is:

(**B**) To the right.

c. ProReason:

Dispatcher: Can you describe the **arrangement** and the **directions** of the currents in the four wires shown in the image?

Vision Expert: Each wire carries a current of 4 A and the arrangement and direction of the current is as follows,
-Wire A (top, currents directed out from the page)
-Wire B (left, currents directed out from the page)
-Wire C (bottom, currents directed out from the page)
-Wire D (right, currents directed out from the page)

Dispatcher: What is the resultant direction of the magnetic force acting on wire P (D) due to the influence of the other wires?

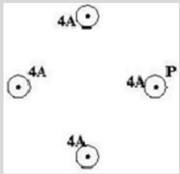
Insight Expert: Reasoning:
The force on the wires depends on the direction of the current in them:
-Wires carrying currents in the same direction attract each other.
-Wires carrying currents in opposite directions repel each other.
The current in the top wire (A) and the bottom wire (C) flows out of the page, similar to that in wire P (D), generating an attractive force directed to the left.
The current in the left wire (B) also flows out of page, producing an attractive force that acts to the left on wire P.
Considering these interactions, the net force experienced by wire P is a combination of the forces directed to the left from A, B, and C; hence, the resultant magnetic force acting on wire P points to the left.
Final Answer:
The total magnetic force acting on wire P points to the left.

Summarizer: The answer is **A**

Figure 2: An example with three reasoning frameworks: fine-grained caption, chain-of-thought, and ProReason. ProReason enables LVLMs to proactively acquire necessary information in a question-oriented manner, and predicts answers based on the collected information. Apparently, ProReason is superior to previous methods, which often describe question-irrelevant visual details, or overlook informative elements. Green indicates correct information or conclusions, while red signifies incorrect ones.

Example case

Question:
For the four identical current-carrying wires shown (with conventional current coming out of the plane of the page), the wire on the right is labeled P. What is the direction of the total magnetic force on the wire labeled P that is caused by the other wires?
(A) To the left (B) To the right (C) Towards the top of the page (D) There is no force
[Dispatcher]:
'Thought':'To determine the direction of the total magnetic force on wire P caused by the other wires, I need to visualize the arrangement and the effect of the magnetic fields created by the other wires, as well as the direction of the conventional current in each wire.',
'EXPERT name':'Vision Expert',
'Question':'Can you describe the **arrangement** and the **directions** of the currents in the four wires shown in the image? '
[Vision Expert]:

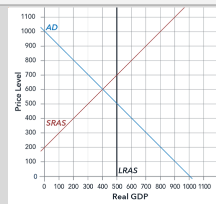


Each wire carries a current of 4 A and the arrangement and direction of the current is as follows,
-Wire A (top, currents directed out from the page)
-Wire B (left, currents directed out from the page)
-Wire C (bottom, currents directed out from the page)
-Wire D (right, currents directed out from the page)
[Referee]:
Thought: The arrangement and the directions of the currents in the four wires shown in the image are as follows:
-Wire A (top, currents directed out from the page)
-Wire B (left, currents directed out from the page)
-Wire C (bottom, currents directed out from the page)
-Wire D (right, currents directed out from the page)
To determine the total magnetic force exerted on wire P by the other three wires, it is necessary to apply the right-hand rule for analysis. Therefore, with the current information provided, it is not possible to directly obtain the answer.
Answer: UNSOLVABLE
[Dispatcher]:
'Thought':'To determine the total magnetic force on wire P due to the three surrounding wires, I need to use the right-hand rule and the fact that all currents are flowing in the same direction. Each wire will exert a magnetic force on wire P that can be calculated from their configurations.',
'EXPERT name':'Insight Expert',
'Question': 'What is the resultant direction of the magnetic force acting on wire P (D) due to the influence of the other wires? '
[Insight Expert]:
Reasoning:
The force on the wires depends on the direction of the current in them: -Wires carrying currents in the same direction attract each other. -Wires carrying currents in opposite directions repel each other.
The current in the top wire (A) and the bottom wire (C) flows out of the page, similar to that in wire P (D), generating an attractive force directed to the left.
The current in the left wire (B) also flows out of page, producing an attractive force that acts to the left on wire P.
Considering these interactions, the net force experienced by wire P is a combination of the forces directed to the left from A, B, and C; hence, the resultant magnetic force acting on wire P points to the left.
Final Answer: The total magnetic force acting on wire P points to the left.
[Referee]:
Thought: Based on the known information, the total magnetic force acting on wire P points to the left.
Answer: SOLVABLE
[Summarizer]:
Thought: Given that the total magnetic force acting on wire P points to the left, when comparing this answer with the choices: (A) To the left (B) To the right (C) Towards the top of the page (D) There is no force
It is evident that the correct choice is **(B)**.
Answer:**B**

Figure 3: A complete reasoning process of PROREASON for the case shown in Figure 2.

Examples of picture unrelated Chain-of-thought.

The graph below shows the AD-AS diagram for Spain. All numbers are in billions. <image 1> What is the size of real GDP in the short-run equilibrium?
(A) 500 (B) 450 (C) 400 (D) 600
We refer to Wikipedia articles on macroeconomics for help. The short-run equilibrium real GDP is equal to the intersection of the SRAS curve and the vertical axis. The SRAS curve is a straight line with slope 200, and the vertical axis is at y = 500. Therefore, the answer is (A).



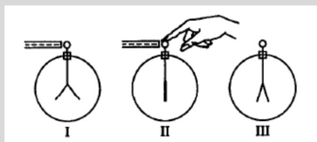
Based on <image 1>, what's the most likely diagnosis?
(A) first degree atrioventricular block (B) third degree atrioventricular block
(C) Second degree type II atrioventricular block (D) atrial flutter
We refer to Wikipedia articles on medicine for help. The given ECG shows a junctional bradycardia with a rate of 40 beats per minute, a P-R interval of 0.12 seconds, and a Q-T interval of 0.39 seconds. The PR interval is prolonged, which is consistent with second-degree type II atrioventricular block. Therefore, the most likely diagnosis is (C) Second degree type II atrioventricular block.

Figure 4: Additional examples of images-unrelated Chain-of-Thought reasoning.

Examples of question unrelated caption.

Question: How much do I have to pay according to the bill
Caption: The image shows an order details screen from a ride-hailing app, displaying a map of a city with two markers indicating the pickup location (green) and drop-off location (red). The total fare is calculated based on the base fare, mileage fee, and duration fee, with a discount applied resulting in the final amount paid. The image also includes a table summarizing the order details, including the base fare, mileage fee, duration fee, total cost, discount, and final amount paid. The image provides a clear overview of the ride details and pricing information for the user.

Question: When a negatively charged rod is brought near, but does not touch, the initially uncharged electroscope shown above, the leaves spring apart (I). When the electroscope is then touched with a finger, the leaves collapse (II). When next the finger and finally the rod are removed, the leaves spring apart a second time (III). The charge on the leaves is
(A) positive in both I and III (B) negative in both I and III
(C) positive in I, negative in III (D) negative in I, positive in III



Caption: The image shows three diagrams depicting the same hand holding a stick, but with different perspectives. In the first diagram, the hand is seen from above, holding the stick vertically. In the second diagram, the hand is seen from the side, holding the stick horizontally. In the third diagram, the hand is seen from the front, holding the stick at a diagonal angle. The diagrams show how perspective can affect our perception of objects and their orientation in space.

Figure 5: Additional examples of question-unrelated passive image captions.

Bad case

Question: What time does the clock show?
Choices: (A) 1:30 (B) 6:30 (C) 5:30 (D) 4:30 (E) 12:30
[Dispatcher]:
'Thought':'I need to determine what time the clock shows, which involves interpreting the image rather than complex reasoning.',
'EXPERT name':'Vision Expert',
'Question':'What time does the clock show in the image?'
[Vision Expert]:
The clock shows **6:25**. The minute hand is pointing at the 5, indicating 25 minutes past the hour, and the hour hand is slightly past the 6.
[Referee]:
Thought: Given that the clock's display of 6:25 does not correspond with any of the answers provided, it can therefore be concluded that this question cannot be resolved with the available information as none of the options represent the correct time the clock shows.
Answer: UNSOLVABLE
[Dispatcher]:
'Thought': 'The clock shows that the hour hand is slightly past the 6 and the minute hand is pointing at the 5, which indicates that the time is 6:25. However, I need to choose the closest matching answer from the provided options.',
'EXPERT name': 'Insight Expert',
'Question': 'Given that the time shows 6:25, which of the provided answer choices (A: 1:30, B: 6:30, C: 5:30, D: 4:30, E: 12:30) is the closest to this time?'
[Insight Expert]:
Reasoning:
(A) 1:30 - This is 4 hours and 5 minutes away from 6:25.
(B) 6:30 - This is 5 minutes away from 6:25.
(C) 5:30 - This is 55 minutes away from 6:25.
(D) 4:30 - This is 1 hour and 55 minutes away from 6:25.
(E) 12:30 - This is 5 hours and 55 minutes away from 6:25.
Now, the option that is the closest to the current time of 6:25 is: 6:30
Final Answer: (B) 6:30, which is only 5 minutes away.

[Referee]:
Thought: The clock currently shows 6:25. The closest time to 6:25 among the listed choices is (B) 6:30, as it is only 5 minutes away. Thus, this question can be resolved with the known information.
Answer: SOLVABLE
[Summarizer]:
Thought: The clock currently shows 6:25, meaning the hour hand is just above the 6 and the minute hand points at the 5, indicating that it is 25 minutes past 6. The closest time to 6:25 among the listed choices is (B) 6:30, as it is only 5 minutes away.
Answer:**B**

Figure 6: A typical mistake made by PROREASON. The vision expert incorrectly identifies 4:30 as 6:25, leading the other agents to base their judgments on this erroneous information, and ultimately resulting in the wrong conclusion.

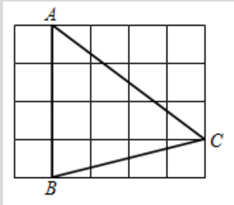
Bad case

Question:

如图，在5×4的正方形网格中，每个小正方形的边长都是1，△ABC的顶点都在这些小正方形的顶点上，则tan∠BAC的值为（）

As shown in the figure, in a 5×4 square grid where each small square has a side length of 1, the vertices of triangle ABC are located at the vertices of these small squares, so the value of tan∠BAC is ().

Choices: (A) $\frac{4}{3}$ (B) 0.75 (C) 0.6 (D) 0.8



[Dispatcher]:

'Thought':'To determine the value of tan∠BAC, I need to know the coordinates of points A, B, and C on the 5×4 grid. This information is represented in the image, and since I cannot see it, I should consult the Vision Expert to get the coordinates of these points.',
'EXPERT name':'Vision Expert',
'Question':'What are the coordinates of points A, B, and C in the 5×4 grid?'

[Vision Expert]:

In a 5×4 grid, we can assign coordinates to the points based on their positions:
- Point A is located at the top right corner of the grid, at coordinates **(4, 3)**.
- Point B is situated at the bottom left corner, at coordinates **(0, 0)**.
- Point C is at the bottom right corner of the grid, at coordinates **(4, 0)**.

[Referee]:

Thought: While the coordinates of points A, B, and C are known, the value of tan∠BAC remains undetermined, preventing the resolution of the problem.
Answer: UNSOLVABLE

[Dispatcher]:

'Thought':'Given the coordinates of A (4, 3), B (0, 0), and C (4, 0), the value of tan∠BAC still needs to be calculated.',
'EXPERT name':'Insight Expert ',
'Question':'Given the coordinates of A (4, 3), B (0, 0), and C (4, 0), calculate the value of tan∠BAC. '

[Insight Expert]:

Reasoning:
To find the value of the tangent of angle BAC, we start by figuring out the slopes of the lines AB and AC in the triangle made up of points A, B, and C.
First, let's look at the coordinates of the points:
Point A is located at (4, 3).
Point B is located at (0, 0).
Point C is located at (4, 0).
Next, we figure out the slope of line segment AB:
The slope of AB, denoted as m_AB, can be found by dividing the difference in the y-coordinates by the difference in the x-coordinates, which gives us (0 - 3) / (0 - 4) = (-3) / (-4) = 3/4.
Then, we determine the slope of line segment AC:
For AC, when we try to calculate its slope using the same method, we end up with (0 - 3) / (4 - 4), which results in division by zero. This tells us that the line AC is vertical and does not have a defined slope.
Given that AC is a vertical line and AB has a slope of 3/4, we can use this information to find the tangent of angle BAC. In a right triangle, the tangent of an angle is the length of the opposite side divided by the length of the adjacent side.
For angle BAC:
Line AB rises 3 units vertically (from point B at (0, 0) to point A at (4, 3)) and runs 4 units horizontally (from point B at (0, 0) to point C at (4, 0)).
Therefore, the tangent of angle BAC is the ratio of these two lengths: 3 divided by 4, which equals 0.75.
Final Answer: The tangent of angle BAC is 0.75.

[Referee]:

Thought: Based on the known information, the value of tan∠BAC is 0.75.
Answer: SOLVABLE

[Summarizer]:

Thought: Given that ∠BAC=0.75, when comparing this value with the answer choices:
(A) **4**/**3** (B) 0.75 (C) 0.6 (D) 0.8
It is evident that the correct choice is (B).
Answer:**B**

Figure 7: A typical mistake made by ProReason. The vision expert incorrectly provided the coordinates of the three vertices of the triangle, leading the other agents to base their judgments on this erroneous information, and ultimately resulting in the wrong conclusion.

A.7 Prompts

Prompt templates of image Relevance Score .

I answered a question based on what was in a picture, and here is the standard answer:
{standard answer}
Here is my answer:
{answer}
Standard answer effectively utilizes key information from the images, providing detailed and question-oriented image descriptions.
Based on the standard answer, please evaluate the relevance of my answer to the content of the image, on a scale of 1 to 5.

Please base your response on the following format:
Assessment process: analyze and assess here.
Final answer: one of ['1', '2', '3', '4','5']

Prompt templates of caption effectiveness evaluation.

I answered a question based on what was in a picture, and here is the question:
{question}
Here is the caption of the picture:
{caption}
And here is the standard answer:
{standard answer}
Standard answer effectively utilizes key information from the images, providing detailed and question-oriented image descriptions.
Based on the standard answer, please evaluate:
1. The level of detail in the caption.
2. The relevance of the caption to the question.
3. The extent to which the caption includes information used in the standard answer.
On a scale of 1 to 5.

Please base your response on the following format:
Assessment process: analyze and assess here.
Final answer:
The level of detail in the caption: one of ['1', '2', '3', '4','5']
The relevance of the caption to the question: one of ['1', '2', '3', '4','5']
The extent to which the caption includes information used in the standard answer: one of ['1', '2', '3', '4','5']

Figure 8: Prompt templates of Relevance Score and caption effectiveness evaluation.

Prompt template of Chain-of-Thought.

Please solve the following question with step-by-step reasoning: {question}

Prompt template of fine-grained image captions generation.

Please describe the contents of this image in detail: {image}

Prompt template of Compositional Chain-of-Thought (CCoT).

For the provided image and its associated question, generate only a scene graph 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

Prompt template of ReAct.

Answer the following questions as best you can. You have access to the following tools:
image_description_tool:
Call this tool to interact with the Image Description Tool API.
Utilize this tool when you require insight into the components of an image, such as identifying objects or reading text within it.
Parameters:
[{'name': 'image_description_query',
'description': 'The input for this tool must be a question in string format. For example: The input could be, "What items are in this picture?"',
'required': True,
'schema': {'type': 'string'}}]
Format the arguments as a JSON object.
computational_tool:
Call this tool to interact with the computational tool API.
Use this tool when you need to conduct reasoning, such as calculating the current in a device with a voltage of 4 volts across and a resistance of 10 ohms, and similar scenarios.
Parameters:
[{'name': 'computational_query',
'description': 'The input for this tool must be a problem that requires calculation and reasoning. For example: The input could be, "What is the acceleration produced by a force of 10 Newtons acting on a 1-kilogram object?"',
'required': True,
'schema': {'type': 'string'}}]
Format the arguments as a JSON object.

Use the following format:
Question: the input question you must answer
Thought: you should always think 'step by step' about what to do
Action: the action to take, should be one of [image_description,computational_tool]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can be repeated zero or more times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question

Begin!

Figure 9: Prompt templates of Chain-of-Thought, fine-grained image captions generation, Compositional Chain-of-Thought (CCoT), and ReAct.

Prompt template of Dispatcher.

You currently need to address the following question:
{question}
The information you need is in an image, but you can't see the image right now.
At the same time, you're not capable of complex reasoning.

However, you can can consult the following two EXPERTs for help:
1. Vision Expert: You can ask him for information in the picture, for example, you could ask him, "What color is the bird in the picture?"
2. Insight Expert: You can ask him to get the results of complex reasoning, e.g. you can ask him, "What is the acceleration produced by a 1N force applied to a 1KG object?"

To solve this problem, which EXPERT do you think you should consult now?

Use the following format:
{
'Thought': 'analyze the problem here.',
'EXPERT name':'The name of the EXPERT you choose should be one of Vision Expert and Reasoning Expert',
'Question':'Questions you want to ask the EXPERT’
}

The last expert you chose was:
{last expert}
And the information you know currently is as follows:
{memory}

Figure 10: Prompt templates of Dispatcher.

Prompt template of Vision expert.

Please answer the following question in detail: {question}

Prompt template of Insight expert.

The following is the available information:
{memory}
Please solve the following problems step by step:
{question}
Use the following format:
Reasoning: Perform a step-by-step process of reasoning to solve a problem.
Final Answer: The final answer you get when you have finished reasoning.

Prompt template of Referee.

My current question that needs to be addressed is:
{question}
The following is the known information:
{memory}
Return SOLVABLE if you think question can be resolved with known information. Otherwise return UNSOLVABLE.
Use the following format:
Thought: Conduct an analysis before you give me an answer.
Answer: the action to take, should be one of ['SOLVABLE', 'UNSOLVABLE']

Prompt template of Summarizer.

My current question that needs to be addressed is:
{question}
The following is the known information:
{memory}
Please solve the question using the following format:
Thought: Conduct a step-by-step analysis before you give me an answer.
Answer: The final answer you get when you have finished analysis.

Figure 11: Prompt templates of Vision Expert, Insight Expert, Referee, and Summarizer.

Prompt templates of LLMs-based ProReason evaluation.

I answered a question based on what was in a picture, and here is the standard answer:
{standard answer}
Here is my answer:
{answer}
Standard answer effectively utilizes key information from the images, providing detailed and question-oriented image descriptions.

Based on the standard answer, please evaluate:
1. The relevance of my answer to the content of the image.
2. The extent of redundant information contained in my answer.
3. Extent of missing necessary information in my answer.
On a scale of 1 to 5.

Please base your response on the following format:
Assessment process: analyze and assess here.
Final answer:
The relevance of my answer to the standard answer: a float between 1 and 5. Cannot output any other characters except numbers.
The extent of redundant information contained in my answer: a float between 1 and 5. Cannot output any other characters except numbers.
Extent of missing necessary information in my answer: a float between 1 and 5. Cannot output any other characters except numbers.

Figure 12: Prompt templates of LLMs-based PROREASON evaluation.

Prompt templates of Vision and Insight Expert Integration .

You need to act as the following two agents:
1. Vision Expert
2. Insight Expert

The prompt for each of the two agents is as follows:
1. Vision Expert
<Vision Expert Prompt Begin>
Please answer the following vision question in detail:
<The question>
<Vision Expert Prompt End>

2. Insight Expert
<Insight Expert Prompt Begin>
The following is the available information:
<Memory>
Please solve the following problems step by step:
<The question>
Use the following format:
Reasoning: Perform a step-by-step process of reasoning to solve a problem.
Final Answer: The final answer you get when you have finished reasoning.
<Insight Expert Prompt End>

where <The question>is:
<QUESTION FROM DISPATCHER>
<Memory> is:
<MEMORY>

Only one of Vision Expert or Insight Expert is selected to answer the question each time, and this time <EXPERT NAME> is selected.

Please act as the Vision Expert, and Insight Expert in the following format:
VISION EXPERT:
<The response of Vision Expert, if Vision Expert is not selected, output 'None'>
Insight Expert:
<The response of Insight Expert, if Insight Expert is not selected, output 'None'>

Figure 13: Prompt templates of Vision and Insight Expert Integration.

Prompt templates of Dispatcher, Vision Expert Insight Expert and Referee Integration.

You need to act as the following four agents:
1. Dispatcher
2. Vision Expert
3. Insight Expert
4. Referee

The prompt for each of the four agents is as follows:
1. Dispatcher
<Dispatcher Prompt Begin>
You currently need to address the following questions:
<Question>
You can consult the following two EXPERTs for help:
1. Vision Expert: You can ask him for information in the picture, for example, you could ask him, "What color is the bird in the picture?"
2. Insight Expert: You can ask him to get the results of complex reasoning, e.g. you can ask him, "What is the
acceleration produced by a 1N force applied to a 1KG object?"
To solve this problem, which EXPERT do you think you should consult now?
Use the following format:
{
'Thought': 'analyze the problem here.',
'EXPERT name':'The name of the EXPERT you choose should be one of Vision Expert and Insight Expert',
'Question':'Questions you want to ask the EXPERT'
}
The last expert you chose was <Last_Expert> and the information you know currently is as follows:
<Memory>
<Dispatcher Prompt End>

2. Vision Expert
<Vision Expert Prompt Begin>
Please answer the following vision question in detail:
<The question>
<Vision Expert Prompt End>

3. Insight Expert
<Insight Expert Prompt Begin>
The following is the available information:
<Memory>
Please solve the following problems step by step:
<The question>
Use the following format:
Reasoning: Perform a step-by-step process of reasoning to solve a problem.
Final Answer: The final answer you get when you have finished reasoning.
<Insight Expert Prompt End>

4. Referee
<Referee Prompt Begin>
My current QUESTION that needs to be addressed is:
<Question>
The information I know is:
<Memory>
Return SOLVABLE if you think question can be resolved with known information. Otherwise return UNSOLVABLE.
Use the following format:
Thought: Conduct an analysis before you give me an answer.
Answer: the action to take, should be one of ['SOLVABLE', 'UNSOLVABLE']
<Referee Prompt End>

where <Question> is:
<QUESTION>
<Memory> is:
<MEMORY>
<Last_Expert> is:
<LAST_EXPERT>

Please act as the Dispatcher, Vision Expert, and Insight Expert and Referee in the following format:
DISPATCHER:
<The response of Dispatcher>
VISION EXPERT:
<The response of Vision Expert, if Vision Expert is not selected by Dispatcher, output 'None'>
Insight Expert:
<The response of Insight Expert, if Insight Expert is not selected by Dispatcher, output 'None'>
REFEREE:
<The response of Referee>

Figure 14: Prompt templates of Dispatcher, Vision Expert Insight Expert and Referee Integration.

Prompt templates of All Five Sub-Agent Integration.

You need to act as the following five agents:
1. Dispatcher
2. Vision Expert
3. Insight Expert
4. Referee
5. Summarizer

The prompt for each of the five agents is as follows:
1. Dispatcher
<Dispatcher Prompt Begin>
You currently need to address the following questions:
<Question>
You can consult the following two EXPERTs for help:
1. Vision Expert: You can ask him for information in the picture, for example, you could ask him, "What color is the bird in the picture?"
2. Insight Expert: You can ask him to get the results of complex reasoning, e.g. you can ask him, "What is the
acceleration produced by a 1N force applied to a 1KG object?"
To solve this problem, which EXPERT do you think you should consult now?
Use the following format:
{
'Thought': 'analyze the problem here.',
'EXPERT name':'The name of the EXPERT you choose should be one of Vision Expert and Insight Expert',
'Question':'Questions you want to ask the EXPERT'
}
The last expert you chose was <Last_Expert>.
<Dispatcher Prompt End>

2. Vision Expert
<Vision Expert Prompt Begin>
Please answer the following vision question in detail:
<The question>
<Vision Expert Prompt End>

3. Insight Expert
<Insight Expert Prompt Begin>
Please solve the following problems step by step:
<The question>
Use the following format:
Reasoning: Perform a step-by-step process of reasoning to solve a problem.
Final Answer: The final answer you get when you have finished reasoning.
<Insight Expert Prompt End>

4. Referee
<Referee Prompt Begin>
My current QUESTION that needs to be addressed is:
<Question>
Return SOLVABLE if you think question can be resolved with known information. Otherwise return UNSOLVABLE.
Use the following format:
Thought: Conduct an analysis before you give me an answer.
Answer: the action to take, should be one of ['SOLVABLE', 'UNSOLVABLE']
<Referee Prompt End>

5. Summarizer
<Summarizer Prompt Begin>
My current question that needs to be addressed is:
<Question>
Please solve the question using the following format:
Thought: Conduct a step-by-step analysis before you give me an answer.
Answer: The final answer you get when you have finished analysis.
<Summarizer Prompt End>

where <Question> is:
<QUESTION>
<Last_Expert> is:
<LAST_EXPERT>

Please act as the Dispatcher, Vision Expert, and Insight Expert, Referee and Summarizer in the following format:
While 'UNSOLVABLE' in the output of Referee:
    DISPATCHER:
    <The response of Dispatcher>
    VISION EXPERT:
    <The response of Vision Expert, if Vision Expert is not selected by Dispatcher, output 'None'>
    Insight Expert:
    <The response of Insight Expert, if Insight Expert is not selected by Dispatcher, output 'None'>
    REFEREE:
    <The response of Referee>
SUMMARY:
<The response of Summarizer>

You can repeat the Dispatcher->Vision Expert or Insight Expert->Referee loop several times until Referee outputs 'SOLVABLE' then you can end the loop and use Summarizer to give the final answer.

Figure 15: Prompt templates of All Five Sub-Agent Integration.