PROREASON: MULTI-MODAL PROACTIVE REASON ING WITH DECOUPLED EYESIGHT AND WISDOM

Anonymous authors

Paper under double-blind review

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.*, insufficient and irrelevant visual descriptions, and limited multi-modal capacities). We then decompose visual reasoning process into two stages: visual perception (*i.e.*, eyesight) and textual reasoning (*i.e.*, wisdom), and introduce a novel visual reasoning framework named PROREASON. This framework features multi-run proactive perception and decoupled vision-reasoning capabilities. 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 both existing multi-step reasoning frameworks and passive peer methods on a wide range of benchmarks for both open-source and closed-source models. In addition, with the assistance of LLMs, PROREASON achieves a performance improvement of up to 15% on MMMU benchmark. 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.

029 030

004

010 011

012

013

014

015

016

017

018

019

021

025

026

027

028

031

033

1 INTRODUCTION

In recent years, large language models (LLMs) (Dubey et al., 2024; Team et al., 2023; Jiang et al., 2023) have experienced explosive growth in their capabilities, driving significant advancements 035 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) (Chen et al., 2024a; Bai 037 et al., 2023), which, like LLaVA (Li et al., 2024b), have achieved remarkable performance in multimodal understanding tasks. However, state-of-the-art (SOTA) LVLMs still struggle to integrate visual understanding with textual reasoning simultaneously due to inherent modality differences and 040 deficient training data. For example, Ghosh et al. (2024) demonstrate that LVLMs often rely more on 041 their prior language knowledge, neglecting visual information in multi-modal reasoning tasks such 042 as visual chart understanding and math reasoning, resulting in performance degradation. Figure 2 043 illustrates a typical case of this issue, where the reasoning process remains irrelevant to the image.

044 To address the challenges, a promising solution is to extract visual information from images into textual form to assist LVLMs in reasoning (Yin et al., 2023b; Mitra et al., 2024). For example, Ghosh 046 et al. (2024) instruct LVLMs to generate fine-grained captions to facilitate the subsequent reasoning 047 process. However, these existing methods exhibit two primary limitations: (i) The visual extraction 048 process is question-agnostic and reasoning-free, that the image description is not targeted for a given question, and reasoning is not involved to deduce extra information for better descriptions. This drawback, termed as "passive", results in irrelevant or insufficient information, and subsequent 051 performance degradation. (ii) The reasoning process entangles visual understanding with textual reasoning abilities of a single LVLM. As mentioned above, the limited capabilities of LVLMs do 052 not always successfully handle both abilities and produce high-quality reasoning. Hence, both of these limitations hinder the performance of LVLMs in multi-modal reasoning tasks.

054 To address both shortcomings, we propose a novel multi-step multi-modal reasoning framework 055 named PROREASON, which features proactive (*i.e.*, question-oriented and reasoning-involved) vi-056 sual extraction and disentangled vision-reasoning capabilities. Specifically, we first decouple multi-057 modal reasoning capacity into two sub-tasks: visual perception (*i.e.*, eyesight) and textual reason-058 ing (i.e., wisdom). The former focuses on understanding images and converting visual details into textual form, while the latter integrates the gathered information to draw final conclusions. As illustrated in Figure 1, our framework can be divided into three stages: Action, Judgment, and 060 Summary, with each involving different agents. In the Action stage, a Dispatcher selectively en-061 gages a Vision Expert to capture additional visual information, or a Reasoning Expert to derive 062 more insights. Unlike passive methods, all three agents operate based on the given question and 063 current descriptions, effectively avoiding information redundancy or insufficiency. Subsequently, a 064 Referee in the Judgment stage evaluates the sufficiency of known information, deciding whether to 065 return to the Action stage or proceed to the Summary stage, where all information is consolidated 066 for a Summarizer to generate the final answer. Notably, the disentangled vision-reasoning does not 067 require vision-irrelevant roles (*i.e.*, Reasoning Expert, Referee, and Summarizer) to be performed 068 by LVLMs, enabling the seamless integration of existing LLMs with proven strong reasoning abilities (Chang et al., 2024), thereby addressing the limitations of LVLMs. 069

Empirically, we evaluate PROREASON across four challenging visual reasoning benchmarks with 071 both open-source and closed-source models. Extensive experiments demonstrate that PROREASON 072 achieves significant and consistent performance improvement over both existing reasoning frame-073 works and state-of-the-art passive methods across all the benchmarks, with a peak enhancement 074 as 13.2%. Besides, our comparative analysis highlights the necessity and relative importance of 075 each sub-agent, as well as the advantages of proactive information acquisition. With the verified 076 superiority over the simultaneous inherent usage of both capabilities, the decoupled perspective not only demonstrates that textual reasoning capabilities outweigh visual understanding for multi-modal 077 reasoning tasks, but also exemplifies the feasibility of LLM-assisted LVLM reasoning, exhibiting performance improvements of up to 15% on the MMMU benchmarks (Yue et al., 2023). 079

The main contributions of this work are threefold:

- We propose a novel multi-modal reasoning framework named PROREASON, featuring iterative proactive perception and decoupled vision-reasoning capabilities.
- Empirically, we identify the drawbacks of existing solutions (*i.e.*, insufficient and irrelevant visual descriptions, and limited multi-modal capacities), and validate the successful mitigation of these issues with PROREASON by the superior performance over peer methods.
- PROREASON showcases the remarkable feasibility of integrating LLMs for boosted visual reasoning, illuminating the potential for LLM-assisted LVLM reasoning in future research.

2 PRELIMINARY OBSERVATIONS

We first analyze the behaviors of several existing methods on the MMMU (Yue et al., 2023) dataset, a challenging multi-modal benchmark requiring comprehensive college-level knowledge and finegrained reasoning abilities. All experiments are conducted with three recent LVLMs for robustness: Llama3-LLaVA-NeXT-8B (Li et al., 2024a), LLaVA-OneVision-Qwen2-7B-OV (Li et al., 2024b) and Qwen-VL-Chat (Bai et al., 2023). Further details are provided in Sec. 4.1.

097 098 099

100

080

081 082

083 084

085

087

090 091

092 093

094

095

096

2.1 CHALLENGES IN VISUAL REASONING: LIMITATIONS OF LVLMS

Chain-of-Thought(CoT) (Wei et al., 2022) has been extensively verified to enhance the performance of LLMs. Here, we explore its impact on the reasoning performance of LVLMs. Similarly, we require the models to "think step by step" before generating final answers. As shown in Table 1, a counter-intuitive phenomenon is observed: compared to "Direct" answering method, the introduction of CoT consistently incurs slight performance degradation across all three models. Inspired by the findings of Ghosh et al. (2024) that LVLMs often rely on prior language knowledge yet neglecting visual information, we hypothesize that the degraded performance might be caused by the image-irrelevant reasoning of CoT.

Method		Model	
u	Llama3-LLaVA-NeXT-8B	LLaVA-OneVision-Qwen2-7B-OV	Qwen-VL-Chat
Direct	41.8	49.1	36.2
CoT	41.5	46.5	35.4
VDGD	42.3	49.7	37.1

Table 1: Performance of three recent LVLMs on MMMU dataset with different assisting techniques.

Table 2: Image Relevance Score of Chain-of-Thought reasoning traces for "True" and "False" responses of three LVLMs, respectively.

Model	Image Relevance Score		
	True	False	
Llama3-LLaVA-NeXT-8B	3.70	3.36	
LLaVA-OneVision-Qwen2-7B-OV	4.27	3.67	
Qwen-VL-Chat	3.63	2.80	

Table 3: 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-LLa	VA-NeXT-8B
	True	False
Detail Level	4.43	3.93
Question Relevance	3.87	3.30
Reasoning Effective Info Inclusion	3.91	3.57

To validate the hypothesis, we design a pipeline referencing Liu et al. (2023c) to measure the correlation between model responses and image contents. Specifically, given a question, we take the reasoning process generated by GPT-40-mini¹, whose superiority on MMMU task is verified in Sec. 4.2, as the golden solution. Subsequently, a more powerful judge (i.e., GPT-4 (Achiam et al., 2023)) is employed to evaluate the relevance of responses of candidate models (e.g., Llama3-LLaVA-NeXT-8B) and this golden solution with the prompt template in Figure 8, and gives a comprehensive score from 1 to 5, with 5 indicating the highest relevance. Since the golden solution contains the essential information needed to solve the problem within the image, a higher correlation indicates a stronger relevance to the image content. As shown in Table 2, the reasoning chains of incorrect responses are significantly less relevant to the images than those of correct answers. An exemplary case is shown in Figure 2, where Llama3-LLaVA-NeXT-8B at-tempts to use the right-hand rule to solve the task, but its reasoning process is almost unrelated to the image, ultimately leading to an incorrect conclusion. This highlights the limitation of CoT for multi-modal reasoning tasks, that image-irrelevant reasoning process provokes performance degradation, since the limited capabilities of LVLMs cannot effectively handle both image and text information simultaneously.

2.2 DRAWBACKS OF PASSIVE INFORMATION EXTRACTION

Passive visual reasoning techniques suffer insufficient and irrelevant visual information, de-spite mitigating the oversight of images by converting them into detailed captions, like Visual De-scription Grounded Decoding (VDGD) (Ghosh et al., 2024). To support this claim, we generate fine-grained image captions using GPT-40-mini with the prompt shown in Figure 9. We then in-corporate these captions into the prompts for LVLMs to facilitate the reasoning process. As listed in Table 1, while these image descriptions improve the performance of LVLMs, the gains are marginal,

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



(2) Passive Visual Reasoning Enhancement Tech- (3) Multi-step Reasoning Framework for LLMs nique

Figure 1: Overview and comparison of workflows of PROREASON, VDGD and ReAct. Unlike existing works (e.g., VDGD and ReAct), our proposed method enables the model to actively acquire necessary information from the images, making it more suitable for LVLMs.

187 188

183

185

189

consistently amounting to less than 1%. This underscores the limited utility of captions generated by passive methods.

192 For further demonstration, we analyze the generated captions along three dimensions: Detail Level, 193 Question Relevance, and Reasoning Effective Info Inclusion, measuring the richness of detail, rele-194 vance to the given question, and the inclusion of information necessary for reasoning, respectively. 195 Similar to Sec. 2.1, we pair the image captions with the reasoning process of GPT-40-mini, 196 and employ GPT-4 as a judge to rate the captions from 1 to 5 across these dimensions. The 197 prompt template is shown in Figure 8. As shown in Table 3, the captions for correct responses of Llama3-LLaVA-NeXT-8B receive higher scores across all three criteria, highlighting the im-199 portance 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 200 Level dimension, indicating that while the captions are detailed, they often lack relevance to the 201 questions. Figure 2 shows a case where Llama3-LLaVA-NeXT-8B utilizes fine-grained captions 202 to solve a question from the MMMU benchmark. As illustrated, although the caption exhaustively 203 describes the image content, it incorrectly describes the wires in the image as octagons, and misses 204 information about the locations of these wires. This information is irrelevant to the target question, 205 thus offering minimal assistance to LVLMs. In summary, our analysis highlights the drawbacks of 206 passive visual reasoning enhancement techniques in terms of information insufficiency and redun-207 dancy, due to their question-agnostic property.

208

209 210

3 Method

211 212

As illustrated in Figure 1, PROREASON consists of five functionally distinct yet inter-cooperative sub-agents, along with a Memory component. The entire workflow comprises three steps: Action, Judgment, and Summary, effectively integrating the proactive visual understanding (*i.e.*, eyesight) and textual reasoning (*i.e.*, wisdom) capacities for enhanced multi-modal reasoning performance.

216 3.1 ACTION STEP

217

Action step is the core of question-oriented visual information extraction, driven by three sub-agents: Dispatcher, Vision Expert, and Reasoning Expert. The Dispatcher orchestrates the workflow, selectively directing the Vision Expert to capture specific visual information, or instructing the Reasoning Expert to analyze known information to derive more. The responses from both the Vision Expert and the Reasoning Expert are stored in a textual Memory component.

223 Formally, given an image I and its corresponding textual question Q, the Dispatcher decides to con-224 sult the Vision Expert or Reasoning Expert, based on the analysis of Q and the known information 225 in the Memory (if not empty). The Dispatcher then generates a query q for the chosen expert. If 226 the Vision Expert is selected, it takes the image I and query q as input, and generates an answer 227 A_v , which is then stored in the Memory. When the Reasoning Expert is selected, it provides a response consisting of the reasoning process and the final answer A_r based on the query q and known 228 information in the Memory, before only A_r is stored in Memory. Notably, the Memory component 229 allows PROREASON to keep compact information, and avoids lengthy reasoning traces like CoT and 230 ReAct, thereby suffering less from redundant information. 231

- 232
- 233 3.2 JUDGMENT STEP

235 Judgment step is conducted by a sub-agent called Referee, which evaluates whether the information 236 stored in Memory is sufficient to answer the question Q. The input to the Referee includes the 237 question Q and available information in the Memory. If the Memory contains adequate information to answer the question Q, the Referee outputs the identifier "SOLVABLE"; otherwise, it outputs 238 "UNSOLVABLE". If the Referee's output is SOLVABLE, the workflow precedes to the Summary 239 phase. Conversely, if the output is UNSOLVABLE, the Action phase is re-executed to gather more 240 necessary information. This allows the Referee to collaborate closely with the three sub-agents in the 241 Action step, enabling the framework to proactively acquire the necessary information and prevent 242 omissions, thereby overcoming the drawbacks of passive methods. 243

- 244
- 245 3.3 SUMMARY STEP 246

The Summary 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 determines that the information in the Memory is sufficient to address the question Q, and outputs the identifier SOLVABLE, the Summarizer will be called to draw a conclusion, based on the Memory information. This conclusion represents the answer to the question Q of the whole framework, and will be evaluated by the performance metrics.

253 254

255

3.4 ADVANTAGES OF PROREASON

Reduced information Mission or Redundancy. With the close collaboration of the Dispatcher,
 Vision Expert, Reasoning Expert, and Referee agents, PROREASON can proactively (*i.e.*, question orientedly and reasoning-involvedly) extract the necessary and sufficient visual details from images,
 effectively avoiding information omission or redundancy. Meanwhile, the Memory component re tains only the image captions from the Vision Expert and the reasoning results from the Reasoning
 Expert, providing a compact textual descriptions. This contributes to minimize the interference of
 irrelevant information on the subsequent Summarizer.

263

LLM-assisted Multi-modal Reasoning. In PROREASON, the complete 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 designate pipeline to complete the tasks. This decomposition-before-integration approach not only circumvents the inherent differences between modalities, but also allows visually irrelevant sub-agents to be performed by text-only LLMs. Consequently, the extensively verified powerful reasoning capabilities of LLMs can be seamlessly synergized to power multi-modal reasoning process for performance enhancement.

270 4 EXPERIMENTS

272

273

274

284

287

289

291

293

295

296

312

313

314

315

316

317

318

319

320

321

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.

275 4.1 GENERAL SETUP 276

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., 2023a), 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 as follows:

- **MME** is an inclusive benchmark that encompasses 14 subtasks, designed to evaluate perceptual and cognitive abilities. Due to our emphasis on visual reasoning, we select the cognition-relevant tasks, including Commonsense Reasoning, Numerical Calculation, Text Translation, and Code Reasoning.²
- **MMMU** evaluates multi-modal models with multidisciplinary tasks that require collegelevel domain-specific knowledge and detailed reasoning. It comprises 11,500 questions across 30 disciplines and 183 sub-fields, emphasizing advanced perception and domainspecific reasoning.
 - **MathVista** focuses on more challenging mathematical reasoning tasks that demand precise visual recognition and compositional reasoning. It includes 6,141 examples from 31 multi-modal 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.

Base Models. We employ GPT-40-mini, Llama3-LLaVA-NeXT-8B and Qwen2.5-72B-297 Instruct (Team, 2024) as our base models, due to their excellent performance, accessibility, 298 and representativeness of model sources. As one of the most performant LVLMs, GPT-40-mini 299 demonstrates significant advancements in visual reasoning capabilities, and provides cheap and fast 300 API. In contrast, Llama3-LLaVA-NeXT-8B is a fully open-sourced LVLM developed by the 301 LLaVA team based on the Llama-3-8B LLM (Dubey et al., 2024) and CLIP vision encoder (Rad-302 ford et al., 2021). Qwen2.5-72B-Instruct is a robust LLM designed to deliver high-quality 303 instruction following and handle complex tasks. It will be utilized for achieving LLM-assisted 304 multi-modal reasoning. 305

Baselines. Besides the most basic method where models are instructed to answer questions di rectly, we compare PROREASON with two categories of peer methods. To determine the benefits of
 proactive information extraction in PROREASON, we first consider two SOTA passive visual reason ing methods, VDGD (Ghosh et al., 2024) and CCoT (Mitra et al., 2024). Additionally, we choose
 two multi-step reasoning frameworks of LLMs, Chain of Thought (CoT) (Wei et al., 2022) and
 ReAct (Yao et al., 2022), to demonstrate the effects of directly migrating LLM solutions to LVLMs.

- **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.
 - **VDGD.** With image caption prefixed to text instruction, VDGD selects tokens that deviate the least from the description using a formula based on Kullback-Leibler divergence.
- **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.
- **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).

 ²To facilitate the comparison across different benchmarks, the results for MME benchmark are calculated
 by the percentage of correct answers out of the total answers. For the results using the official MME calculation method, please refer to Table 9, Table 10, Table 11, and Table 12.

328	Model	Method		Data	set	
329 330	110001	Wiethou	MME	MMMU	MathVista	Hallu.
331		Direct	61.5	41.8	37.1	45.8
332	Llama3-LLaVA-	VDGD	68.8 (+7.3)	42.3 (+0.5)	36.1 (-1.0)	44.2 (-1.6)
333		CCoT	68.9 (+7.4)	40.5 (-1.3)	36.8 (-0.3)	37.4 (-8.4)
334		CoT	58.8 (-2.7)	41.5 (-0.3)	35.9 (-1.2)	43.1 (-2.7)
335	NEX1-0B	ReAct	68.5 (+7.0)	46.7 (+4.9)	31.7 (-5.4)	43.6 (-2.2)
336		PROREASON	71.5 (+10.0)	52.5 (+10.7)	38.8 (+1.7)	50.9 (+5.1)
337		Average	66.30	44.22	36.06	44.16
339		Direct	79.2	48.4	53.0	56.0
340		VDGD	82.3 (+3.1)	51.4 (+3.0)	51.2 (-1.8)	52.4 (-3.6)
341		CCoT	80.8 (+1.6)	54.2 (+5.8)	53.6 (+0.6)	56.7 (+0.7)
342	GPT-40-mini	CoT	87.8 (+8.6)	58.5 (+10.1)	53.8 (+0.8)	56.3 (+0.3)
343		ReAct	87.3 (+8.1)	54.8 (+6.4)	49.3 (-3.7)	51.1 (-4.9)
344		PROREASON	90.4 (+11.2)	61.6 (+13.2)	54.9 (+1.9)	58.9 (+2.9)
345		Average	84.63	54.82	52.63	55.23
346		U				

Table 4: Performance of multiple approaches with two base models across four visual reasoning
benchmarks. "Hallu." is the abbreviation of HallusionBench. Based on the performance of the
direct method, red and blue signify the improvement and degradation, respectively.

Table 5: Performance of Llama3-LLaVA-NeXT-8B Vision Expert, assisted by different text-only LLMs, on various benchmarks under ReAct and PROREASON frameworks. The red and blue texts indicate the improvements and reductions relative to the case of using the same method without the assistance of text-only LLMs, respectively.

Vision Expert	Textual	Method		Dat	aset	
vision Expert	Sub-Agents	method	MME	MMMU	MathVista	Hallu.
Llama3-	Llama3-LLaVA- NeXT-8B	Direct ReAct PROREASON	61.5 68.5 71.5	41.8 46.7 52.5	37.1 31.7 38.8	45.8 43.6 50.9
LLaVA- NeXT-8B	Qwen2.5-72B- Instruct	ReAct ProReason	71.0 (+2.5) 81.3 (+9.8)	50.4 (+3.7) 56.8 (+4.3)	34.6 (+2.9) 48.8 (+10.0)	40.4 (-3.2) 52.3 (+1.4)
	GPT-40-mini	ReAct ProReason	73.6 (+5.1) 84.7 (+13.2)	48.4 (+1.7) 54.5 (+2.0)	36.2 (+4.5) 41.7 (+2.9)	46.7 (+3.1) 53.1 (+2.2)

• **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 Reasoning Experts, respectively. This aligns with our notions for easy understanding, and is shown in Figure 1.

Implementation Details. The prompt templates for all methods are shown in Figures 9, 10 and 11. Specifically for PROREASON, in order to prevent an infinite loop, if Dispatcher selects Vision Expert or Reasoning Expert to obtain information up to 5 times, and Referee still judges that the existing knowledge in Memory is not enough to solve the question, the information collection phase will be terminated immediately, and then Summarizer will directly answer the question based on the available information in Memory. This setup, refined through multiple trials, is the most effective. Additionally, since we cannot obtain the tokens output by GPT-40-mini, we omit the step of selecting the token with the smallest deviation from the image description when implementing VDGD for GPT-40-mini.

378 4.2 MAIN RESULTS 379

PROREASON exhibits significant and consistent performance enhancement over baselines
 across all the benchmarks. As listed in Table 4, 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 outperforms all the other baselines for both
 Llama3-LLaVA-NeXT-8B and GPT-40-mini model across all benchmarks, with a peak improvement as 13.2%, demonstrating the superiority and task robustness of PROREASON.

386 Proactive information acquisition surpasses SOTA passive methods, especially in complex vi-387 sual reasoning tasks. Specifically, compared to MME, MathVista and HallusionBench present higher image complexity and question difficulty, and thus require stronger visual understanding and tex-388 tual reasoning capabilities. This leads to performance degradation of passive methods (*i.e.*, VDGD 389 and CCoT), highlighting their limited applicability to complex visual reasoning tasks. In contrast, 390 PROREASON achieves notable performance improvements, up to 5.1%, by proactively acquiring vi-391 sual information from images rather than generating question-agnostic captions. This aligns with out 392 previous obervations in Sec. 2.2 that passive methods introduce substantial information redundancy 393 or omission, misleading subsequent reasoning processes. 394

Decoupling the visual perception and textual reasoning capabilities of an LVLM outperforms their simultaneous inherent usage. As listed in Table 4, when both capabilities are utilized concurrently, CoT consistently degrades performance compared to the "Direct" method across all benchmarks with Llama3-LLaVA-NeXT-8B model, consistent with the findings in Sec. 2.1. 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-40-mini across all benchmarks, demonstrating the effectiveness of capability decoupling.

402 PROREASON outperforms ReAct with even less token consumption. Given the similarity in 403 multi-step reasoning, we compare PROREASON and ReAct in terms of performance and cost. 404 Specifically, as an LLM-specific multi-step reasoning framework, ReAct only outperforms the "Di-405 rect" method on MME and MMMU, but underperforms on MathVista and HallusionBench, showing 406 inferior performance compared to the consistent improvements of PROREASON. Furthermore, we 407 compare their average token consumption of GPT-40-mini. As shown in Table 8 of Appendix A, 408 PROREASON consumes significantly fewer tokens than ReAct on both MME and MathVista tasks, 409 indicating its higher token efficiency and the importance of a compact Memory in reducing token 410 usage. Coupled with better performance, this suggests the superiority of PROREASON over LLMspecific ReAct framework. 411

412 Text-only LLMs can be effectively integrated into PROREASON for dramatically enhanced 413 performance. As mentioned in Sec. 3.4, the decoupled visual perception and textual reasoning ca-414 pabilities facilitate the seamless integration of text-only LLMs. To demonstrate the utility of this 415 advantage, we fix the Vision Expert as Llama3-LLaVA-NeXT-8B, and replace other agents with text-only LLMs. As listed in Table 5, with the assistance of either Qwen2.5-72B-Instruct or 416 GPT-40-mini, the Llama3-LLaVA-NeXT-8B Vision Expert receives remarkable performance 417 boost across all benchmarks, particularly by 15% on MMMU and 11.7% on MathVista, compared to 418 directly providing answers. In contrast, ReAct gains a much smaller improvement. This highlights 419 the unique advantage of PROREASON in leveraging existing text-only LLMs for enhanced perfor-420 mance. Notably, this advantage may open new avenues for continuously pushing the performance 421 limits of LVLMs with the assistance of existing powerful LLMs. 422

- 422 423 424
- 4.3 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, Reasoning Expert, Referee, or Summarizer, respectively, while the other sub-agents are powered by GPT-40-mini. Given that Llama3-LaVA-NeXT-8B exhibits weaker visual understanding and textual reasoning capabilities than GPT-40-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. 442

443

444

470

477

478 479

480

481

482

483

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-40-mini. The blue text indicates the performance decline
compared to the scenario with all agents performed by GPT-40-mini.

436 437	Dataset	GPT-40-mini		Llam	na3-LLaVA-NeXI	-8B	
438	Ducuser	011 10 11111	Dispatcher	Vision Expert	Reasoning Expert	Referee	Summarizer
439	MME	90.4	88.8 (-1.6)	84.7 (-5.7)	88.7 (-1.7)	89.6 (-1.1)	84.2 (-6.2)
440	MMMU	61.6	60.9 (-0.7)	54.5 (-7.1)	60.2 (-1.4)	51.5 (-10.1)	51.0 (-10.6)
441							

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-40-mini.

Dataset	GPT-40-r	nini	Llama3-LLa	VA-NeX	KT-8B
Dutuset	Textual Sub-Agents	Vision Expert	All Sub-Agents	COT	Direct
MME MMMU	84.7 (+13.2) 54.5 (+2.0)	77.8 (+6.3) 53.4 (+0.9)	71.5 52.5	58.8 41.5	61.5 41.8

453 Summarizer is the most crucial sub-agent, closely followed by Referee. The replacement of 454 Summarizer results in the most notable performance decline on both MME and MMMU tasks, 455 reaching 6.2% and 10.6%, respectively. This highlights the critical function of the Summarizer in 456 integrating all available information to conclude final answers. Besides, the substitution of Referee 457 leads to a 10.1% reduction on MMMU. Given that MMMU is more challenging than MME, this 458 finding underscores the essential role of the Referee in assessing the sufficiency of information, 459 particularly in more complex visual reasoning tasks.

Relatively, Dispatcher and Reasoning 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 Reasoning 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 Reasoning Expert have the least impact.

4.4 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.
 - **GPT-40-mini as Vision Expert**. Based on the above scenario, we implement the Vision Expert with GPT-40-mini, while keep the other textual sub-agents unchanged.
- **GPT-40-mini as Textual Sub-Agents.** Reversely, we utilize Llama3-LLaVA-NeXT-8B as the Vision Expert, and GPT-40-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-40-mini achieves consistent performance enhancement,

highlighting the significance of both capabilities. However, substituting the textual sub-agents with GPT-40-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 Sec. 4.3 that identifies the Summarizer and Referee as the most crucial sub-agents.

- 5 RELATED WORK
- 492 493

491

Large Visual-Language Model. Recently, large vision-language models (LVLMs) (Bai et al., 494 2023; Chen et al., 2023; Liu et al., 2024b) have garnered widespread attention and demonstrated re-495 markable advancements in understanding and generating multi-modal contents. In the open-source 496 domain, numerous LVLMs, like LLaVA (Liu et al., 2023c;b; 2024a; Li et al., 2024a;b) and In-497 ternVL (Chen et al., 2024b) families, have been extensively developed. In the closed-source do-498 main, proprietary models such as GPT-4 (Achiam et al., 2023) and Gemini Pro 1.5 (Reid et al., 499 2024) have also achieved significantly success. Despite these advancements, existing LVLMs still 500 encounter challenges in effectively integrating visual understanding with textual reasoning capabil-501 ities simultaneously. This limitation is particularly evident in their diminished attention to image 502 content during visual reasoning process, such as chart interpretation and visual math reasoning, 503 leading to degraded performance (Ghosh et al., 2024) and motivating more effective solutions.

504

505 Passive Visual Reasoning. Extracting information from images into text can effectively as-506 sist LVLM in performing visual reasoning tasks. Visual Description Grounded Decoding (VDGD) (Ghosh et al., 2024) first describes the image before prefixing this description to the 507 prompt, assisting LVLMs on visual reasoning tasks. Furthermore, Compositional Chain-of-Thought 508 (CCoT) (Mitra et al., 2024) directs LVLMs to generate scene graphs (SGs) that serve as a bridge 509 between the visual and textual domains, aiding LVLMs in subsequent tasks. However, most of 510 these methods employ a question-agnostic and reasoning-free visual extraction process, where im-511 age descriptions are not tailored to specific questions, and no reasoning is applied to infer additional 512 information for improved descriptions. These "passive" approaches lead to the inclusion of irrele-513 vant or redundant information, ultimately degrading performance. In contrast, PROREASON adopts 514 question-oriented agents to collect necessary and sufficient information, effectively circumventing 515 these drawbacks.

516

517 Multi-step Reasoning Framework. Multi-step reasoning frameworks have been developed for 518 LLMs to achieve better performance by breaking down complex questions into easier ones (Pan et al., 2024). As a representative method, Chain-of-Thought (CoT) (Wei et al., 2022) enhances the 519 arithmetic and commonsense reasoning capabilities by explicitly generating intermediate reason-520 ing steps before concluding the final answers. Tree-of-Thoughts (ToT) (Yao et al., 2024) further 521 refines the CoT mechanism by allowing LLMs to consider multiple reasoning paths and do self-522 assessment before making decisions. Considering that the inherent knowledge of LLMs may not be 523 sufficient to complete tasks, ReAct (Yao et al., 2022) integrates information retrieval into the reason-524 ing chains, enabling models to pause to verify results and determine whether additional information 525 is needed before proceeding. Nevertheless, multi-step reasoning frameworks designed for text-only 526 LLMs are not fully applicable to visual reasoning tasks, and may even impair the performance of 527 LVLMs (Ghosh et al., 2024).

528 529

530

6 CONCLUSION

531 In this paper, we first validate that existing multi-modal reasoning approaches still suffer insufficient 532 and irrelevant visual descriptions, as well as limited multi-modal capacities. To address these issues, 533 we decompose the visual reasoning process into visual perception and textual reasoning stages, and 534 introduce a novel visual reasoning framework named PROREASON, featuring multi-run proactive perception and decoupled vision-reasoning capabilities. Empirically, extensive experiments demon-536 strate the superiority of PROREASON over both passive image information acquisition methods and 537 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 538 feasibility of integrating LLMs for multi-modal reasoning with dramatically improved performance, highlighting the great potential for LLM-assisted LVLM reasoning in future research.

540 REFERENCES

552

553

554

555

556

560

561

562

563

565

566

567

575

580

581

582

583

588

589

590

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
 report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.
 - Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.
- Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong,
 Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024a.
 - Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pp. 24185–24198, 2024b.
 - Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. arXiv preprint arXiv:2309.15402, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Ilama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- 572 Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Vdgd: Mitigating lvlm hallucinations in cognitive prompts by bridging the visual perception gap. *arXiv preprint arXiv:2405.15683*, 2024.
- 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. Deepseek-coder: When the large language model meets programming the rise of code intelligence, 2024. URL https://arxiv.org/abs/2401.14196.
 - Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.
- 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.
 Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.
 - Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024a. URL https://llava-vl.github.io/blog/ 2024-05-10-llava-next-stronger-llms/.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei
 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024b.

594 595 596 597	Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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. <i>arXiv preprint arXiv:2310.14566</i> , 2023a.
598 599 600	Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.
601 602	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c.
603 604 605	Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https:// llava-vl.github.io/blog/2024-01-30-llava-next/.
606 607	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024b.
609 610 611	Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of- thought prompting for large multimodal models. In <i>Proceedings of the IEEE/CVF Conference</i> <i>on Computer Vision and Pattern Recognition</i> , pp. 14420–14431, 2024.
612 613	Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. Chain-of-action: Faithful and multimodal question answering through large language models. <i>arXiv preprint arXiv:2403.17359</i> , 2024.
614 615 616 617 618	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In <i>International conference on machine learning</i> , pp. 8748–8763. PMLR, 2021.
619 620 621 622	Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean- baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gem- ini 1.5: Unlocking multimodal understanding across millions of tokens of context. <i>arXiv preprint</i> <i>arXiv:2403.05530</i> , 2024.
623 624 625	Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role- playing. <i>arXiv preprint arXiv:2310.10158</i> , 2023.
626 627 628 629	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.
630 631 632 633	Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. <i>arXiv preprint arXiv:2312.11805</i> , 2023.
634 635	Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.
636 637 638	Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multi- modal mathematical reasoning with math-vision dataset. <i>arXiv preprint arXiv:2402.14804</i> , 2024.
639 640 641	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 35:24824–24837, 2022.
642 643 644 645	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. <i>arXiv preprint arXiv:2210.03629</i> , 2022.
646 647	Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. <i>Advances in Neural Information Processing Systems</i> , 36, 2024.

- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023a.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023b.
- Kiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

702 A APPENDIX

704

705

A.1 DEMONSTRATIVE EXAMPLES



749 750

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. 758 759 760 761 762 763 Example case 764 **Ouestion:** 765 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 766 by the other wires? 767 (A) To the left (B) To the right (C) Towards the top of the page (D) There is no force 768 [Dispatcher]: 'Thought': To determine the direction of the total magnetic force on wire P caused by the other wires, I need to visualize the 769 arrangement and the effect of the magnetic fields created by the other wires, as well as the direction of the conventional 770 current in each wire.', 'EXPERT name':'Vision Expert', 771 'Question': 'Can you describe the arrangement and the directions of the currents in the four wires shown in the image? ' 772 [Vision Expert]: 44. 773 Each wire carries a current of 4 A and the arrangement and direction of the current is as follows, 774 -Wire A (top, currents directed out from the page) ()^{4A} A. -Wire B (left, currents directed out from the page) 775 -Wire C (bottom, currents directed out from the page) 776 44 -Wire D (right, currents directed out from the page) 777 [Referee]: Thought: The arrangement and the directions of the currents in the four wires shown in the image are as follows: 778 -Wire A (top, currents directed out from the page) 779 -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) 781 To determine the total magnetic force exerted on wire P by the other three wires, it is necessary to apply the right-hand rule 782 for analysis. Therefore, with the current information provided, it is not possible to directly obtain the answer. Answer: UNSOLVABLE 783 [Dispatcher]: 784 'Thought': To determine the total magnetic force on wire P due to the three surrounding wires, I need to use the right-hand 785 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 786 be calculated from their configurations.', 'EXPERT name': 'Reasoning Expert', 787 'Question': 'What is the resultant direction of the magnetic force acting on wire P (D) due to the influence of the other wires? ' 788 [Reasoning expert]: Reasoning: 789 The force on the wires depends on the direction of the current in them: -Wires carrying currents in the same direction attract 790 each other. -Wires carrying currents in opposite directions repel each other. 791 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. 792 The current in the left wire (B) also flows out of page, producing an attractive force that acts to the left on wire P. 793 Considering these interactions, the net force experienced by wire P is a combination of the forces directed to the left from A, 794 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. 795 [Referee]: 796 Thought: Based on the known information, the total magnetic force acting on wire P points to the left. 797 Answer: SOLVABLE [Summarizer]: 798 Thought: Given that the total magnetic force acting on wire P points to the left, when comparing this answer with the choices: 799 (A) To the left (B) To the right (C) Towards the top of the page (D) There is no force 800 It is evident that the correct choice is (**B**). Answer:**B** 801 802

802 803

756

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

804 805

806

807

808



864 A.2 ANALYSIS OF BAD CASES

In this section, we analyze the typical errors made by GPT-40-mini-driven PROREASON, exemplified as Figure 6 and 7, for further understanding and 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 reasoning 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 reasoning 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.

919	
920	
921	
922	
923	
924	
925	
026	
920 097	
020	
920	Bad case
930 931 932	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.', reasoning.',
933	'EXPERT name': Vision Expert',
934	[Vision Expert]:
935	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
936	past the 6.
937	Thought: Given that the clock's display of 6:25 does not correspond with any of the answers provided, it can therefore be
938	concluded that this question cannot be resolved with the available information as none of the options represent the correct
939	time the clock shows.
940	Dispatcher:
941	Thought': 'The clock shows that the hour hand is slightly past the 6 and the minute hand is pointing at the 5, which indicates
942	that the time is 6:25. However, I need to choose the closest matching answer from the provided options.',
943	'Ouestion': '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:
944	12:30) is the closest to this time?'
945	[Reasoning expert]:
946	Reasoning: (A) $1:30$ - This is 4 hours and 5 minutes away from 6:25
947	(B) 6:30 - This is 5 minutes away from 6:25. [10] 2
948	(C) 5:30 - This is 55 minutes away from 6:25.
949	(D) 4:30 - This is 1 hour and 55 minutes away from 6:25. (E) 12:20 This is 5 hours and 55 minutes away from 6:25
950	Now, the option that is the closest to the current time of 6:25 is: 6:30 $7 \frac{1}{6} 5$
951	Final Answer: (B) 6:30, which is only 5 minutes away.
952	[Referee]:
953	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
954	Answer: SOLVABLE
955	[Summarizer]:
956	Thought: The clock currently shows 6:25, meaning the hour hand is just above the 6 and the minute hand points at the 5,
057	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
551 050	Answer: B
900	
929	

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.

972	
973	
974	
975	
976	
977	Bad case
978	Question:
979	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
980	tan之bAC的但为() 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
0.021	ABC are located at the vertices of these small squares, so the value of $\tan \angle BAC$ is ().
901	Choices: (A) $\frac{4}{2}$ (B) 0.75 (C) 0.6 (D) 0.8
902	[Dispatcher]:
983	'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
984	information is represented in the image, and since I cannot see it, I should consult the Vision Expert to get the coordinates of
985	'EXPERT name':'Vision Expert'.
986	'Question':'What are the coordinates of points A, B, and C in the 5×4 grid?'
987	[Vision Expert]:
988	In a 5×4 grid, we can assign coordinates to the points based on their positions:
989	- Point A is located at the top light corner, at coordinates (0, 0).
990	- Point C is at the bottom right corner of the grid, at coordinates $(4, 0)$.
991	[Referee]:
992	Thought: While the coordinates of points A, B, and C are known, the value of $\tan \angle BAC$ remains undetermined, preventing the resolution of the problem
993	Answer: UNSOLVABLE
994	[Dispatcher]:
995	'Thought': Given the coordinates of A (4, 3), B (0, 0), and C (4, 0), the value of tan \angle BAC still needs to be calculated.',
006	EXPERT name : Reasoning Expert ', Ouestion' Given the coordinates of A (4, 3) B (0, 0) and C (4, 0) calculate the value of \tan / BAC '
990	[Reasoning expert]:
997	Reasoning:
998	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
999	First, let's look at the coordinates of the points:
1000	Point A is located at (4, 3).
1001	Point B is located at $(0, 0)$.
1002	Point C is located at (4, 0).
1003	The slope of AB, denoted as m_AB, can be found by dividing the difference in the y-coordinates by the difference in the
1004	x-coordinates, which gives us $(0 - 3) / (0 - 4) = (-3) / (-4) = 3/4$.
1005	Then, we determine the slope of line segment AC:
1006	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
1007	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
1008	right triangle, the tangent of an angle is the length of the opposite side divided by the length of the adjacent side.
1009	For angle BAC:
1010	Line AB rises 5 units vertically (from point B at $(0, 0)$ to point A at $(4, 3)$) and runs 4 units norizontally (from point B at $(0, 0)$) to point C at $(4, 0)$)
1011	Therefore, the tangent of angle BAC is the ratio of these two lengths: 3 divided by 4, which equals 0.75.
1012	Final Answer: The tangent of angle BAC is 0.75.
1012	[Referee]:
1013	Answer: SOLVABLE
1014	[Summarizer]:
CIUI	Thought: Given that $\angle BAC=0.75$, when comparing this value with the answer choices:
1016	(A) 4/3 (B) 0.75 (C) 0.6 (D) 0.8
1017	Answer: B
1018	
1019	

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.

1026 A.3 PROMPTS

1028	
1029	Prompt templates of image Relevance Score .
1030	I answered a question based on what was in a picture, and here is the standard answer:
1031	{standard answer}
1032	{answer}
1033	Standard answer effectively utilizes key information from the images, providing detailed and question-oriented image
103/	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
1035	bused on the standard answer, please of address the relevance of my answer to the content of the mage, on a seare of 1 to 5.
1036	Please base your response on the following format:
1037	Assessment process: analyze and assess here. Final answer: one of [1]: 2]: 3]: 4]: 5]
1037	
1030	Prompt templates of caption effectiveness evaluation.
1039	I answered a question based on what was in a picture, and here is the question:
1040	{question}
1041	{caption}
1042	And here is the standard answer:
1043	{standard answer} Standard answer effectively utilizes key information from the images, providing detailed and question-oriented image
10/15	descriptions.
1045	Based on the standard answer, please evaluate:
1040	1. The level of detail in the caption. 2. The relevance of the caption to the question
1047	3. The extent to which the caption includes information used in the standard answer.
1048	On a scale of 1 to 5.
1049	Please base your response on the following format:
1050	Assessment process: analyze and assess here.
1051	Final answer:
1052	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']
1053	The extent to which the caption includes information used in the standard answer: one of ['1', '2', '3', '4','5']
1054	
1055	
1055	Figure 8: Prompt templates of Relevance Score and caption effectiveness evaluation.
1057	
1058	
1059	
1060	
1001	
1062	
1063	
1064	
1065	
1066	
1067	
1068	
1069	
1070	
1071	
1072	
1073	
1074	
1075	
1076	
1077	
1078	

D	
P	rompt template of Chain-of-1 hought.
Р	lease solve the following question with step-by-step reasoning: {question}
Р	rompt template of fine-grained image captions generation.
P	lease describe the contents of this image in detail: {image}
-	
P	rompt template of Compositional Chain-of-Thought (CCoT).
F	or the provided image and its associated question, generate only a scene graph in JSON format that includes the fo
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
P	compt template of ReAct
	newer the following questions as best you can. You have access to the following tools:
in	ase description tool.
C	all this tool to interact with the Image Description Tool API.
U	tilize this tool when you require insight into the components of an image, such as identifying objects or reading te:
W	ithin it.
Pa	rameters:
[{	'name': 'image_description_query',
'de	escription': 'The input for this tool must be a question in string format. For example: The input could be, "What ite
ar 'rc	e in this picture?",
'sc	hema' {'type' 'string'}}]
Fo	brmat the arguments as a JSON object.
co	mputational_tool:
Ca	all this tool to interact with the computational tool API.
U	se this tool when you need to conduct reasoning, such as calculating the current in a device with a voltage of 4 vol
ac	ross and a resistance of 10 ohms, and similar scenarios.
Pa fí	rrameters:
L1 'de	name. computational_query, escription': 'The input for this tool must be a problem that requires calculation and reasoning. For example: The in
co	buld be. "What is the acceleration produced by a force of 10 Newtons acting on a 1-kilogram object?".
're	quired': True,
sc	hema': {'type': 'string'}}]
Fo	ormat the arguments as a JSON object.
U	se the following format:
Q T	uestion: the input question you must answer
	ction: the action to take should be one of limage description computational tool
A	ction Input: the input to the action
0	bservation: the result of the action
	(this Thought/Action/Action Input/Observation can be repeated zero or more times)
Tł	nought: I now know the final answer
Fi	nal Answer: the final answer to the original input question
P	acint
D	

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

134	
135	
136	Prompt template of Dispatcher.
137	You currently need to address the following question:
138	{question}
139	At the same time, you're not canable of complex reasoning.
140	
141	However, you can consult the following two EXPERTs for help:
142	1. VISION EXPERT: You can ask nim for information in the picture, for example, you could ask nim, "what color is the bird in the picture?"
143	2. Reasoning Expert: You can ask him to get the results of complex reasoning, e.g. you can ask him, "What is the
144	acceleration produced by a 1N force applied to a 1KG object?"
145	To solve this problem, which EXPERT do you think you should consult now?
146	
147	Use the following format:
148	'Thought': 'analyze the problem here.'.
149	'EXPERT name': The name of the EXPERT you choose should be one of Vision Expert and Reasoning Expert',
150	'Question':'Questions you want to ask the EXPERT'
151	
152	The last expert you chose was:
153	{last expert}
154	{memory}
155	
156	Figure 10: Prompt templates of Dispatcher
157	rigare 10. riompt templates of Dispateller.
158	
150	
160	
100	Prompt template of Vicion expert
161	Tompt complate of vision expert.
161 162	Please answer the following question in detail: {question}
161 162 163	Please answer the following question in detail: {question}
161 162 163 164	Please answer the following question in detail: {question} Prompt template of Reasoning expert.
161 162 163 164 165	Please answer the following question in detail: {question} Prompt template of Reasoning expert. The following is the available information:
161 162 163 164 165	Please answer the following question in detail: {question} Prompt template of Reasoning expert. The following is the available information: {memory} Please solve the following problems step by step:
161 162 163 164 165 166	Please answer the following question in detail: {question} Prompt template of Reasoning expert. The following is the available information: {memory} Please solve the following problems step by step: {question}
161 162 163 164 165 166 167	Please answer the following question in detail: {question} Prompt template of Reasoning expert. The following is the available information: {memory} Please solve the following problems step by step: {question} Use the following format:
161 162 163 164 165 166 167 168	Please answer the following question in detail: {question} Prompt template of Reasoning 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 Answerser The final engager use at when you have finished reasoning.
161 162 163 164 165 166 167 168 169	Please answer the following question in detail: {question} Prompt template of Reasoning 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.
161 162 163 164 165 166 167 168 169 170	Please answer the following question in detail: {question} Prompt template of Reasoning 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.
161 162 163 164 165 166 167 168 169 170 171	Please answer the following question in detail: {question} Prompt template of Reasoning expert. The following is the available information: {memory} Please solve the following problems step by step: {question} (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:
161 162 163 164 165 166 167 168 169 170 171 172	Please answer the following question in detail: {question} Prompt template of Reasoning 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 language in th
161 162 163 164 165 166 167 168 169 170 171 172 173	Please answer the following question in detail: {question} Prompt template of Reasoning 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)
161 162 163 164 165 166 167 168 169 170 171 172 173 174	Prompt template of Vision expert. Please answer the following question in detail: {question} Prompt template of Reasoning 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.
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175	Prompt template of Vision Cxprt. Please answer the following question in detail: {question} Prompt template of Reasoning 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:
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176	Prompt template of vision expert. Please answer the following question in detail: {question} Prompt template of Reasoning 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 L'SOLVABLE. 'UNSOLVABLE':
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177	Prompt template of Vision capert. Please answer the following question in detail: {question} Prompt template of Reasoning 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']
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178	Prompt template of Vision Expert: Please answer the following question in detail: {question} Prompt template of Reasoning 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.
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179	Prompt template of Vision capter. Please answer the following question in detail: {question} Prompt template of Reasoning 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:
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180	Please answer the following question in detail: {question} Prompt template of Reasoning 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 format: Thought: Conduct an analysis before you give me an answer. Answer: the action to take, should be one of ['SOLVABLE', 'UNSOLVABLE']
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180	Please answer the following question in detail: {question} Prompt template of Reasoning 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 format: (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)
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182	Please answer the following question in detail: (question) Prompt template of Reasoning 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: (question] Prompt template of Summarizer. My current question that needs to be addressed is: {question] Prompt template of Summarizer. My current question that needs to be addressed is: {question] Prompt template of Summarizer. My current question that needs to be addressed is: {question] Prompt template of Summarizer. My current question that needs to be addressed is: {question] The following is the known information: {memory} 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:
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183	Please answer the following question in detail: {question} Prompt template of Reasoning 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. My current question that needs to be addressed is: [question] Prompt template of Summarizer. My current question that needs to be addressed is: [question] Prompt template of Summarizer. My current question that needs to be addressed is: [question] Prompt template of Summarizer. My current question that needs to be addressed is: [question] Prompt template of Summarizer. My current question that needs to be addressed is: [question] Prompt template of Summarizer. My current question that needs to be addressed is: [question] 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.
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 177 178 179 180 181 182 183 184	Please answer the following question in detail: {question} Prompt template of Reasoning 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] 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.
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 177 178 179 180 181 182 183 184 185	Please answer the following question in detail: {question} Prompt template of Reasoning 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) Prompt template of Summarizer. My current question that needs to be addressed is: (question) The following is the known information: (memory)

1188 A.4 SUPPLEMENTARY RESULTS

Table 8: Average token consumption of PROREASON and ReAct with GPT-40-mini model on the MME and MathVista benchmarks.

Method	MI	MME MathVista		Vista
	Input	Output	Input	Output
ProReason ReACT	1286.8 1645.0	327.2 197.0	2238.6 3092.8	788.6 845.1

Table 9: Performance of multiple approaches with two base models across four visual reasoning benchmarks. Different from Table 4, MME scores are calculated with the official method here.

1202	Model	Method			Dataset	
1203	Widder	Wiethou	MME	MMMU	MathVista	HallusionBench
1205		Direct	359	41.8	37.1	45.8
1206		VDGD	374	42.3	36.1	44.2
1207	Llama3-LLaVA-	CCoT	354	40.5	36.8	37.4
1208	NeXT-8B	CoT	258	41.5	35.9	43.1
1209		ReAct	385	46.7	31.7	43.6
1210		PROREASON	416	52.5	38.8	50.9
1211		Direct	552	48.4	53.0	56.0
1212		VDGD	584	51.4	51.2	52.4
1213	CDT lo mini	CCoT	535	54.2	53.6	56.7
1214	GP1-40-mini	CoT	729	58.5	53.8	56.3
1215		ReAct	712	54.8	49.3	51.1
1216		PROREASON	779	61.6	54.9	58.9

Table 10: Performance of Llama3-LLaVA-NeXT-8B Vision Expert, assisted by different text-only LLMs, on various benchmarks under ReAct and PROREASON frameworks.Different from Table 5, MME scores are calculated with the official method here.

Vision Expert	Textual	Method		Da	ataset	
, 101011 2.1.pere	Sub-Agents	1.100100	MME	MMMU	MathVista	Hallu.
I.lama3-	Llama3-LLaVA- NeXT-8B	ReAct ProReason	385 416	46.7 52.5	31.7 38.8	43.6 50.9
LLaVA- NeXT-8B	Qwen2.5-72B- Instruct	ReAct ProReason	428 603	50.4 56.8	34.6 48.8	40.4 52.3
	GPT-40-mini	ReAct ProReason	473 613	48.4 54.5	36.2 41.7	46.7 53.1

Table 11: 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-40-mini. Different from Table 6, MME scores are calculated with the official method here.

Datastri	GPT-40-m	ini	Llama3-LLaVA-NeXT-8B			
Dutuset	01 1 10 10	Dispatcher	Vision Expert	Reasoning Expert	Referee	Sumn
MME	779	746	613	743	762	6
MMMU	61.6	60.9	54.5	60.2	51.5	5
Tabla 10.	Daufaunaanaa	of DRODE AGON	with different a	and aunations for th	a malativa	immon
Table 12:	Performance	of PROREASON	with different co	onfigurations for th	e relative n visual r	import easoni
Table 12: sessment Different	Performance between visua from Table 7,	of PROREASON l understanding MME scores are	with different co and textual reaso calculated with	onfigurations for th oning capabilities o the official method	e relative n visual r l here.	import easoni
Table 12: sessment Different	Performance between visua from Table 7,	of PROREASON l understanding MME scores are	with different co and textual reaso calculated with	onfigurations for th oning capabilities o the official method	e relative n visual r 1 here.	import easoni
Table 12: sessment Different	Performance between visua from Table 7,	of PROREASON I understanding MME scores are GPT-40-r	with different co and textual rease calculated with nini	onfigurations for th oning capabilities o the official method Llama3-LLa	e relative n visual r l here. VA-NeX	impor easoni T-8B
Table 12: sessment Different D	Performance between visua from Table 7, Pataset Text	of PROREASON Il understanding MME scores are GPT-40-r ual Sub-Agents	with different co and textual rease calculated with nini Vision Expert	onfigurations for th oning capabilities o the official method Llama3-LLa All Sub-Agents	e relative n visual r l here. VA-NeX COT	impor reasoni T-8B Direc
Table 12: sessment Different D	Performance between visua from Table 7, Pataset Text MME	of PROREASON Il understanding MME scores are GPT-40-r ual Sub-Agents 613	with different co and textual reaso e calculated with nini Vision Expert 512	onfigurations for th oning capabilities o the official method Llama3-LLa All Sub-Agents 416	e relative n visual r l here. VA-NeX COT 286	import reasonit T-8B Direct 359