Plug-and-Play Grounding of Reasoning in Multimodal Large Language Models

Anonymous ACL submission

Abstract

The rise of Multimodal Large Language Models (MLLMs), renowned for their advanced instruction-following and reasoning capabilities, has significantly propelled the field of visual reasoning. However, due to limitations in their image tokenization processes, most MLLMs struggle to capture fine details of text and objects in images, especially in high-resolution samples. To overcome this limitation, we introduce P^{2}G, a novel framework for plug-and-play grounding in MLLMs. P^{2}G utilizes the tool-usage potential of MLLMs to employ expert agents for on-the-fly grounding of reasoning into critical visual and textual elements in images, thereby enabling deliberate reasoning through multimodal prompting. Additionally, we develop P^{2}GB, a benchmark designed to evaluate MLLMs' proficiency in understanding inter-object relationships and textual content in challenging high-resolution images. Extensive experiments on visual reasoning tasks demonstrate the superiority of P^{2}G, achieving performance comparable to GPT-4V on P^{2}GB with a 7B backbone. Our work underscores the potential of grounding reasoning with external agents in MLLMs, presenting a promising alternative to mere model scaling.

1 Introduction

Large language models (LLMs) (Touvron et al., 2023a; OpenAI, 2023; Touvron et al., 2023b) have shown strong potential as a unified backbone for various language tasks, including in-context learning (Brown et al., 2020; Wang et al., 2023b), instruction following (Ouyang et al., 2022), and reasoning (Sun et al., 2023; Wang et al., 2023d).

Extending LLMs to multimodal capabilities, researchers have developed Multimodal Large Language Models (MLLMs) (Zhu et al., 2023; Liu et al., 2023b; Huang et al., 2023; Alayrac et al., 2022; Wang et al., 2023a; Dai et al., 2023), treating each modality as a foreign language (Huang et al., 2023; Wu et al., 2023). These MLLMs show significant results in the field of visual reasoning.

Despite these advancements, MLLMs face limitations in visual reasoning due to the high demand for large-scale annotated data for vision instruction tuning (Zhu et al., 2023; Liu et al., 2023b). Collecting annotated multimedia training examples is challenging, and multimodal instruction tuning data is even harder to scale. Another limitation is capturing details in high-resolution images or those with complex textual information, often leading to hallucinations or incorrect reasoning. Non-lossless tokenization of images can also overlook critical semantic details.

To address these challenges, successor works have explored grounding reasoning in MLLMs. KOSMOS-2 (Peng et al., 2024) and CogVLM (Wang et al., 2023a) generate bounding boxes for visual occurrences. LLaVAR (Zhang et al., 2023) and TGDoc (Wang et al., 2023c) augment instruction tuning data with OCR-based textual clues and bounding boxes. However, these methods require large amounts of data and training costs.

Inspired by recent studies showing LLMs' effective use of external tools and agents (Shen et al., 2023; Zhuang et al., 2023), we propose P^{2}G, a novel framework for plug-and-play grounding of reasoning in MLLMs. Instead of training MLLMs from scratch, we leverage lightweight proxy models as agents to obtain critical clues for reasoning. We use an OCR agent (via PaddleOCR (pad, 2022)) and a visual grounding agent (via Grounding-DINO (Liu et al., 2023c)) for text-rich and high-definition images. MLLMs generate specific queries for these agents based on the complexity of the reasoning task.

To evaluate P^{2}G, we introduce P^{2}GB, a challenging Visual Question Answering (VQA) benchmark designed to assess MLLMs' visual grounding, especially in high-resolution and text-rich scenarios. Our experiments on visual reasoning tasks, including P^{2}GB, demonstrate the superiority of P^{2}G. Notably, P^{2}G achieved comparable performance to GPT-4V on P^{2}GB with a 7B backbone. Our work highlights the potential of plug-and-play grounding of reasoning as an alternative to model scaling. Our contributions are three-fold:

1) We propose P^{2}G, a framework for plug-and-play grounding of reasoning in high-resolution and text-rich visual scenarios using agents.

2) We introduce P^{2}GB, a VQA benchmark to assess MLLMs' reasoning capability in text-rich and high-definition image queries.

3) We conduct extensive experiments on challenging reasoning datasets, demonstrating P^{2}G's superior performance with a 7B MLLM backbone, surpassing similarly scaled or larger models.

2 Methods

Our proposed framework, which we refer to as P^{2}G, primarily addresses the challenge of visual reasoning tasks that involve high-resolution natural images and text-rich images. Our goal is to enhance the model's ability to interpret and analyze these complex visual inputs effectively, thereby improving its performance on visual reasoning that requires a nuanced understanding of both visual and textual elements in detail.

2.1 Overall Design of P^{2}G

Figure 1 illustrates the proposed P^{2}G: **P**lug-and-**P**lay **G**rounding of Reasoning in large vision language models. The key objective of P^{2}G lies in enhancing the groundedness and factualness of reasoning from multimodal language models (MLLMs), without relying on heavily supervised (instruction) fine-tuning on extensive annotated data. And to achieve this objective, we harness the emergent capabilities like *in-context learning* (Dong et al., 2023), *instruction following* (Longpre et al., 2023) and *tool-usage* (Shen et al., 2023) capability of large language models. Below, we introduce the procedure of P^{2}G in detail.

2.1.1 Deliberate Reasoning

To ground the reasoning procedure of MLLMs, one key challenge is the hallucination of reasoning paths. In other words, MLLMs must know their *don't-knows* (Cheng et al., 2024) ahead. To mitigate this issue, we propose Deliberate Reasoning in P^{2}G, which encourages the MLLMs to first assess their current ability to solve the provided question, before moving forward on reasoning.

As illustrated in Figure 1, for a simple visual query, P^{2}G generates the correct answer directly, while for challenging cases, P^{2}G autonomously assesses its current capability, and poses demand on support from external agents (experts) on specific textual or visual supporting clues (in the form of natural language query). By introducing this *deliberate reasoning* process before moving on to the reasoning problem, we could thereby empower the MLLM with external agents for concise textual or visual understanding, which is generally challenging for large vision language models, especially for nuanced but important details high-definition images. The capability of deliberate reasoning ahead is attained through dedicated instruction tuning, which we will elaborate on in Sec. 2.3.

2.1.2 Plug-and-Play Grounding

The surging works in the field of retrieval augmented generation (RAG) (Gao et al., 2023b) and tool-usage (Shen et al., 2023; Liang et al., 2023) inspired us on leveraging external experts (agents) in grounding multimodal reasoning with rich textual and visual facts and clues. One major challenge for MLLMs in reasoning (Liu et al., 2023a,b; Ye et al., 2023) is the expressiveness of image representation, where an *only representation* (visual tokens) is provided for reasoning, which hinders the comprehensiveness of encompassed visual information, especially under high-definition or text-rich scenarios. The information loss during such autoencoding compression refrains MLLM from generating grounded, accurate reasoning. The latest works either fine-tune on more VQA data (Zhang et al., 2023), or prepend OCR texts into context (Wang et al., 2023c; liu), which does not essentially mitigate this core limitation.

As a step forward, we propose *Plug-and-Play* Grounding in P^{2}G, to mitigate the limitation above by providing both rich textual and visual clues, leveraging external agents (experts). As illustrated in Figure 1, based on the specific query on semantic details from MLLMs, we correspondingly call *1) OCR Agent* to collect text pieces, or *2) Grounding Agent* to fetch visual patches corresponding to the crucial semantic objects requested by the MLLM. Beyond fetching these semantic premises, we also incorporate their relevant position in the image into a multi-modal question prompt, before obtaining a

Figure 1: Illustration of our proposed P^{2}G for grounding visual reasoning. Given a multi-modal query including an image and its corresponding question, (1) P^{2}G first deliberately decide whether to seek additional clues (anticipated text and/or visual objects) from dedicated textual and/or visual grounding agents, or provide a direct answer for simple and confident cases. For challenging cases, (2) additional text or visual clues are then obtained via OCR Agent (*text*) or Grounding Agent (*image*) according to MLLM's request. Specifically, we include OCR texts and their relative positions for textual clues, and for visual clues, we detect and locate all objects for each requested class. Finally, we incorporate these clues into a multi-modal prompt for obtaining a grounded reasoning answer.

final comprehensive reasoning answer. Such plug-and-play design enables us to leverage SOTA text (PaddleOCR (pad, 2022)) or image (Grounding DINO (Liu et al., 2023c)) processing tools, mitigating the demand for dedicated tuning of backbone MLLMs. By providing dedicated textual and visual clues, we significantly improve the correctness and groundedness of MLLM's reasoning. Details are described in Sec. 2.2.

2.2 Model Structure

2.2.1 Architectural Designs

P^2G integrates four main components: an LLM, a vision encoder, a projection module, and textual (OCR) and visual grounding agents. These components work jointly to enhance the model's ability to process and interpret complex multimodal data.

We use Vicuna-7B-V1.3 (Zheng et al., 2024) as our LLM, which trained Llama on approximately 125K conversations collected from ShareGPT.com. The vision encoder is CLIP ViT-L/14, which processes inputs resized and padded to 224^2. This encoder handles both the original images and specific regions containing detected objects.

To map visual semantics to the LLM's hidden space, we use two types of projection modules: an MLP module, and a cross-attention-based Resampler (Alayrac et al., 2022). The MLP maintains the count of visual tokens and only reshapes its dimension, while the Resampler (one-layer cross attention)[1] also reduces the token quantity (from 256 to 32) to ensure an efficient context.

To maintain an adequate count of visual tokens, we toggle between the two projection modules. For inputs with only initial (global) image features, the MLP maps all visual tokens. For inputs with 1 to 4 critical objects, we employ an MLP to map the visual features of these objects and utilize the Resampler to downsample the global image. When more than 4 objects are detected via Grounding Agent, the Resampler handles all visual features of objects to ensure an efficient context size.

The Grounding Agent uses Grounding DINO (Liu et al., 2023c) to identify and extract relevant objects, while the OCR Agent utilizes PaddleOCR[2] to retrieve textual information.

[1]The resampler is implemented as a single-layer cross-attention, following Alayrac et al. (2022).

[2]https://github.com/PaddlePaddle/PaddleOCR

> Sorry, I cannot answer the question. Some visual information about the following objects is missing or unclear: $object_1, \ldots, object_n$.

Figure 2: Calling *Grounding Agent* for visual clues.

> <image> (Original image)
>
> Additional visual information to focus on:
> 3 button(s) <object>, <object>, <object> at location [0.25, 0.63, 0.26, 0.64], [0.47, 0.59, 0.48, 0.60], [0.52, 0.62, 0.53, 0.63]
>
> 1 paper clip <object> at location [0.65, 0.70, 0.66 ...
> (Object features and their positions)
>
> [object class] not existent in the image ... (Objects that not detected by Grounding Agents)
>
> Are all buttons in the image larger than the paper clips? Answer the question using a single word or phrase. (Original question)

Figure 3: Example prompt for the model's second round of reasoning, with visual clues from *Grounding Agent*.

2.2.2 Deliberate Reasoning and Plug-and-Play Grounding

We detail the plug-and-play grounding of reasoning in P^2G. As shown in Figure 1, the model first determines if additional visual or textual clues are needed. For straightforward ones, the model directly outputs its reasoning. For high-resolution images or those with detailed text, the model generates query responses, calling the OCR or Grounding Agent. Such capability is attained through instruction fine-tuning, detailed in Section 2.3.

For high-resolution images, the model's initial response may miss certain objects or details, as shown in Figure 2. Grounding DINO detects and crops these objects, magnifying them for focused analysis. These crops are incorporated into prompts for a second round of inference, as illustrated in Figure 3, enabling the model to provide more accurate answers. This process is formalized with a detection function F_d, which processes an image I and a set of target objects $\{object_1, \ldots, object_n\}$, resulting in image crops P:

$$P = F_d(I, \{object_1, \ldots, object_n\}), \quad (1)$$

where $P = \{p_1, p_2, \ldots, p_m\}$ are the image crops identified by Grounding DINO. The total number of objects and individual quantities of each type are related by $\sum_{i=1}^{n} x_i = m$, where n is the total number of object types and x_i is the quantity of

Sorry, I cannot answer the question. Some visual information about the following objects is missing or unclear: text in the image.

Figure 4: Calling *OCR Agent* for textual clues.

<image> (Original image)

Additional visual information to focus on:
Text in the image: 'May311918' at location [0.66, 0.043, 0.931, 0.077]; '3379Bark Jane Rd' at location [0.545, 0.103, 0.921, 0.131]. (Text and their positions)

Please focus on providing an answer to the question without considering any challenges related to the clarity or presence of text in the image.
(Add this segment when no text detected in image)

By whom is this letter written? (Original question)

Figure 5: Example prompt for the model's second round of reasoning with textual clues from *OCR Agent*.

the i-th object. As illustrated in Figure 3, we also inform MLLMs of the objects not being detected, indicating their potential absence from the image.

For text-rich images, the model's call to the OCR Agent is shown in Figure 4. PaddleOCR extracts textual elements, which are integrated with bounding boxes and questions, as shown in Figure 5. This enhances the model's recognition of text presence and positions. Given additional textual clues $\mathcal{T}$ and visual clues $\mathcal{P}$ from external agents, we obtain the final visual reasoning results via:

$$\mathcal{R} = \text{MLLM}(q_i, q_t, \mathcal{T}, \mathcal{P}), \tag{2}$$

where q_i and q_t denote image and text queries, respectively. By conditioning on both image q_i and enriched information $\mathcal{T}$ and $\mathcal{P}$, we achieve plug-and-play grounding of reasoning, leveraging MLLMs' in-context learning and instruction-following capabilities.

2.3 Training of P^{2}G

We outline the training process to equip P^{2}G with multimodal capabilities and deliberate reasoning. It consists of two stages: multimodal instruction tuning and learning of deliberate reasoning, each designed to progressively build the P^{2}G's ability to handle complex visual and textual inputs.

2.3.1 Multimodal Instruction Tuning

The first stage focuses on equipping our base LLM (Vicuna-7B-V1.3 (Zheng et al., 2024)) with fundamental multimodal capabilities. We follow the procedures established in LLaVA (Liu et al., 2023b). We employ a 80K sample from LLaVA instruction data, following the procedures and splits used in V* (Wu and Xie, 2023). This stage brings fundamental multimodal capabilities to LLMs.

2.3.2 Learning of Deliberate Reasoning

Our second stage aims to refine P^{2}G's ability to reason deliberately, using agents to gather additional clues when needed. It involves two key steps: (1) *Identifying Need for Additional Information*. The model learns to differentiate between straightforward and complex queries: Simple queries are answered directly, while complex queries trigger the use of OCR and grounding agents to gather additional textual or visual information. (2) *Learning to incorporate Additional Information*. We curate a set of challenging VQA queries, consisting of both *positive* and *negative* samples. Negative samples train the model to recognize its deficiency and generate agent calls. Positive samples (including both straightforward and complex queries) help the model to utilize additional clues from agents effectively.

Particularly, we adopt a two-round approach: the first stage for direct answering or generating agent calls (*round 1*), and the second stage for utilizing multimodal clues (*round 2*). (1) For **text-rich image reasoning**, we select data from train sets of ChartVQA, DOCVQA, and TextVQA, focusing on images with resolutions over 500 pixels and critical texts smaller than 20 pixels. We pre-extract texts with PaddleOCR. The data was then split into negative samples (indicating the need for additional text) and positive samples. (2) For **visual object grounding**, we adapt data from V* (Wu and Xie, 2023) to improve the model's understanding of quantitative relationships and spatial arrangements between objects by incorporating the number of objects and their bounding boxes. Our two-stage training process ensures P^{2}G handle both simple and complex multimodal queries, leveraging additional information when necessary to provide accurate, grounded answers.

3 P^{2}GB Benchmark

To quantitatively assess the visual reasoning capabilities under text-rich or high-resolution scenarios, we constructed a challenging benchmark P^{2}GB. It includes Comprehensive Image Understanding with Fine-grained Recognition (2080 samples) and Image Text Content Understanding (50

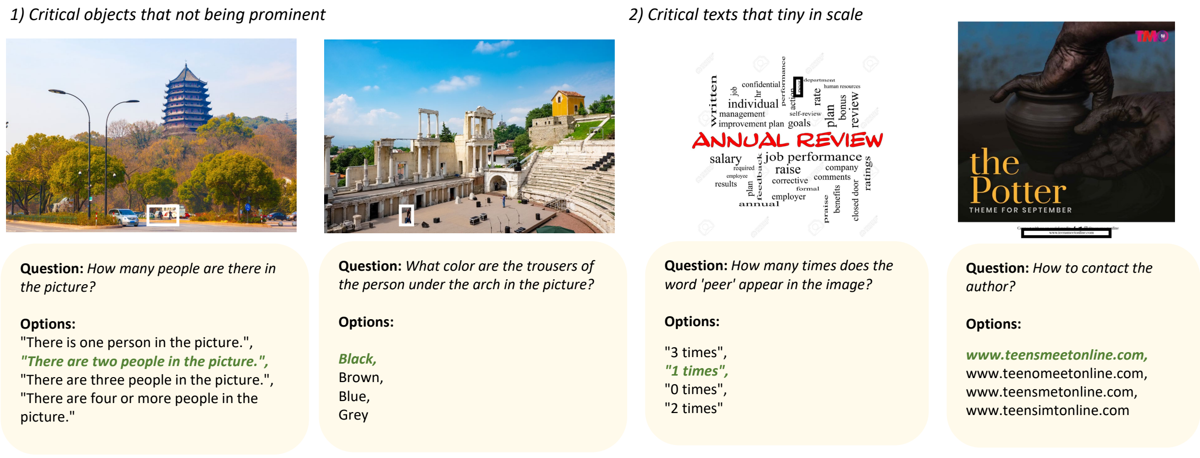


Figure 6: Illustration of our proposed P²GB benchmark. In P²GB, we consider two challenging visual reasoning scenarios: comprehensive image understanding and text-rich visual reasoning. For the former, we delicately collect high-definition image samples where the critical object is not prominent (i.e., tiny in scale) and challenging to identify, while for the latter we include samples in which crucial textual parts are tiny as well.

samples), totaling 2130 samples (pair of an image and multiple-choice question)[3].

(1) *Comprehensive Image Understanding with Fine-grained Recognition* involves analysing high-resolution images with complex scenes containing multiple objects that the model must identify and describe, including their types, locations, and interactions, to test its ability to recognize and distinguish objects within the scene. For this task, we randomly select images from SA-1B (Kirillov et al., 2023) dataset and adopt EVA-02-L (Fang et al., 2023) detector to extract small object (detection boxes) from the images. For each image, the top 5 boxes are retained based on their scores. A detection box is considered a small object if its area is less than 1/10 of the full image. We use GPT-4o as a candidate for generating questions for each image. In each image, a red visual box is used to mark the object that needs to be questioned. GPT-4o generates a question based on the red box, with four answer options and one correct answer. The questions, options, and answers are all manually reviewed subsequently for accuracy, clarity, and does not contain biased or toxic contents.

(2) *Image Text Content Understanding* involves identifying and understanding small textual content within high-resolution images and answering related questions. This task tests the model's ability to discern fine text and engage in logical reasoning based on the text. As illustrated in Figure 6, we design multiple-choice answers for each question that carefully crafted and manually reviewed to ensure validity, fairness, and eliminate ambiguities. To construct this benchmark, we adapt the PowerPoint images and questions from (Wang et al., 2023c), and manually select challenging samples that wider than 1,000 pixels, contains tiny crucial texts, and paired with difficult questions.

[3]The proposed benchmark will be released publicly.

4 Experiments

4.1 Experimental Setup

Models and Baselines For MLLMs, we select Vicuna-7B-V1.3 (Chiang et al., 2023) as the language backbone, and follow LLaVA to train an MLLM backbone for P²G. To build up two agents for visual and textual grounding, we select Grounding DINO (Liu et al., 2023c) for obtaining visual clues (i.e., objects) and PaddleOCR (pad, 2022) for screening texts within the image query. We compare P²G against multiple similar-scaled, instruction-tuned MLLMs, including vanilla LLaVA (Liu et al., 2023b), MiniGPT-4 (Zhu et al., 2023), mPLUG-OWL (Ye et al., 2023), and Instruct-BLIP (Dai et al., 2023). In addition, we compare P²G against MLLMs dedicated optimized for semantic-rich reasoning, i.e., LLaVAR (Zhang et al., 2023), and TGDoc (Wang et al., 2023c). Finally, we include the most capable MLLM so far, GPT-4V (OpenAI, 2023) on our challenging benchmark P²GB.

Datasets Following previous works, we test P²G on a variety of visual reasoning benchmarks. For text-rich visual reasoning, we select DocVQA (Mathew et al., 2021) and ChartVQA (Masry et al.,

Model	Size	DocVQA	ChartVQA	GQA	SEED	MMVET	MME
MiniGPT-4 (Zhu et al., 2023)	7B	3.0	4.3	-	-	-	-
mPLUG-OWL (Ye et al., 2023)	7B	6.9	9.5	-	-	-	-
LlaVAR (Zhang et al., 2023c)	7B	11.6	8.0	-	-	-	-
TGDoc (Wang et al., 2023c)	7B	9.0	12.72	-	-	-	-
LLaVA (Liu et al., 2023b)	7B	19.06	15.30	17.09	23.50	29.10	1107
Instruct-BLIP (Dai et al., 2023)	7B	-	-	49.20	-	26.20	-
LLaVA (Liu et al., 2023b)	13B	31.77	25.70	17.09	24.01	32.70	965
Instruct-BLIP (Dai et al., 2023)	13B	-	-	49.50	-	25.60	-
LLaVA + P^{2}G (Ours)	7B	**61.44**	**37.20**	**59.87**	**27.46**	**32.90**	**1223**

Table 1: Performance of P^{2}G on visual reasoning tasks. The best performing 7B model is marked in **bold**.

Model	Size	Objects	Texts
GPT-4V (OpenAI, 2023)	>1T	50.1	68.0
LLaVA (Vicuna-1.3)	7B	40.1	8.0
LLaVA (Vicuna-1.3)	13B	40.2	8.0
LLaVA + P^{2}G (Ours)	7B	**42.5**	**50.0**
Gain (%)	-	1.06×	6.3×

Table 2: Experimental results of P^{2}G and baselines on our challenging high-resolution benchmark P^{2}GB.

Benchmark	P^{2}G	w/o Position in Prompt	w/ Weaker DINO	w/o Agents
DocVQA	61.4	71.6 (+10.2)	61.4 (0.0)	19.0 (-42.4)
ChartVQA	37.2	26.8 (-10.4)	37.2 (0.0)	15.3 (-21.9)
SEED	27.5	24.6 (-2.9)	27.4 (-0.1)	23.5 (-4.0)
MM-VET	32.9	29.1 (-3.8)	29.3 (-3.6)	29.1 (-3.8)

Table 3: Effects on removing the relative position of grounded (text and/or visual) objects in prompt (*w/o Position in Prompt*), replacing the visual grounding agent with a weaker, non-finetuned DINO (*w/ Weaker DINO*), and removing agents in P^{2}G (w/o Agents).

2022), and GQA (Hudson and Manning, 2019), SEED (Li et al., 2023a), MM-VET (Yu et al., 2023), and MME (Li et al., 2023a) for semantic-rich and general visual reasoning. Beyond existing benchmarks, we also curate a challenging benchmark P^{2}GB, which contains challenging high-definition, semantic, or text-rich visual queries.

Implementation We implement P^{2}G based on the LLM as Vicuna-7B-V1.3, and ViT 224/14, following LLaVa's architecture. We finetune our models on 8 Nvidia GPUs, with a learning rate of $2e-5$, batch size of 16, for one epoch, with a cosine scheduler and Adam optimizer.

For pre-training, we use the 558K subset from LAIONCC-SBU, following LLaVA. Subsequently, we fine-tune on a 427K dataset, comprising 130K negative (for agent call generation) and 297K positive examples[4]. Our negative data includes 110K objects from (Wu and Xie, 2023) and 20K text images[5] from DocVQA, ChartVQA, and TextVQA. The positive data consists of 80K simple questions from VQA train sets (for direct-answering training) and 217K challenging samples for agent utilization (190K object images from (Wu and Xie, 2023) and 27K text images from Doc, Chart, and TextVQA).

[4]The LLaVA 7B and 13B baselines in this work are also reproduced by fine-tuning on the 297K positive examples, following (Wu and Xie, 2023). The difference is that no extra clues from agents are provided for the 217K hard queries.

[5]Selected for their critical text dimensions < 20 pixels.

4.2 Results

4.2.1 Performance on Visual Reasoning

The performance of P^{2}G on visual reasoning benchmarks is presented in Table 2. On text-rich visual reasoning, P^{2}G significantly outperform baselines, including the vanilla LLaVA, by more than doubled (3× on DocVQA, 2.4× on ChartVQA), and also greatly surpass MLLMs that dedicated tuned for text-rich visual reasoning, e.g., LLaVAR and TGDoc, and even surpasses 13B LLaVA variants. On general visual reasoning benchmarks, P^{2}G also enjoys a consistent improvement over LLaVA and InstructBLIP, demonstrating the superiority of P^{2}G.

4.2.2 Performance on P^{2}GB

On the more challenging P^{2}GB, P^{2}G achieved a significant improvement over LLaVA, demonstrating a markedly enhanced comprehension of object details in high-resolution images by over $5x$ compared with vanilla LLaVA. P^{2}G is also comparable to GPT-4V and significantly outperforms baselines on reasoning related with nuanced *Objects*, the most capable MLLM so far, and is huge in scale and training compute. These promising results further highlight the significance of P^{2}G in plug-and-play grounding. A detailed case study on P^{2}GB against GPT-4V is illustrated in Figure 8.

4.2.3 Ablation Study

We study the effect of P^{2}G in Table 3. We first remove the two agents for plug-and-play grounding (w/o Agents) by providing no additional clues, and the performance drops drastically, indicating the significance of Plug-and-Play Grounding. Upon removing the relative position vector for grounded objects and texts, we observed a performance degradation across multiple benchmarks. This decrement was more notable in structured image datasets like ChartVQA, where grounding bounding boxes are essential for the model to locate crucial text pieces[6]. We finally replaced the grounding agent with a weaker model that not being fine-tuned[7]. It drops improvements in benchmarks that require both object and text recognition, such as MM-VET, while it does not impact benchmarks focused solely on text recognition, like DocVQA.

5 Analysis

To further understand the role of deliberate reasoning in P^{2}G, we present a comprehensive analysis of this capability in P^{2}G, on SEED, which contains both text- and visual-rich samples (Li et al., 2023a).

Model	Size	Simple	Hard
LLaVA	7B	29.58	14.86
LLaVA + P^{2}G (Ours)	7B	**33.67**	**18.57**
Gain (%)	-	13.8	25.0

Table 4: Performance P^{2}G and baselines under simple and hard questions in SEED.

Performance Gain via Agent Assistance We first study the effect of deliberate reasoning, under both *simple* and *hard* visual queries. To obtain such splits, we leverage a strong, larger model LLaVa-V1.5-13B. We treat the samples whose answers are correct as simple sets, and vice versa. As listed in Table 4, our P^{2}G is able to improve performance on both easy and difficult tasks, while the improvement is greater for difficult topics. This suggests that our deliberate reasoning allows the model to answer simple questions more confidently while being able to use extrinsic agents to improve performance on complex questions.

[6]In DocVQA, we discover that removing bounding boxes unintentionally enables room for more detected texts within the maximized input token limitation (2K). We expect a positive effect of bounding boxes, given an MLLM with longer context.

[7]Both versions: `longzw1997/Open-GroundingDino`

Routing to Different Agents We further study the routing to each (OCR or Grounding) agent in P^{2}G. As illustrated in Figure 7, both two types of agents are called during inference, indicating that P^{2}G is capable of utilizing corresponding agents for reasoning in need (for text or visual clues).

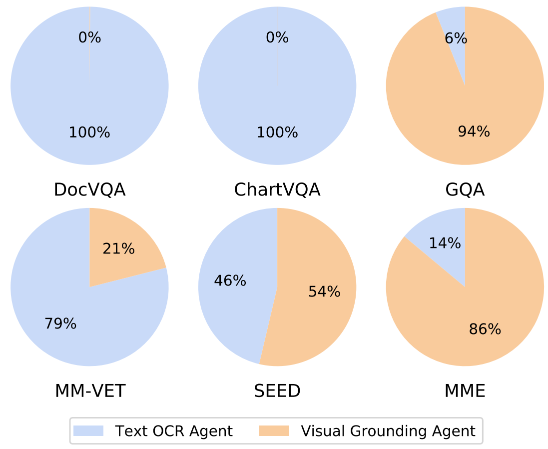


Figure 7: Agent routing of P^{2}G under various tasks.

Case Study of P^{2}G We first perform a case study on P^{2}GB, in Figure 8, where we compare rationales generated by P^{2}G and GPT-4V(ision). As illustrated in the figure, P^{2}G could generate more grounded and accurate answers, especially for text-rich and high-resolution samples. To further understand the deliberate reasoning process of P^{2}G, we provide detailed case studies in Appendix B.

6 Conclusion

In this paper, we focus on the challenge of grounding visual reasoning of multimodal large language models. To address the limitations of most existing works that heavily rely on question-answer pairs for instruction tuning, we propose P^{2}G, a novel framework for plug-and-play grounding of visual reasoning. Dedicated tuned to deliberate thinking, P^{2}G promptly generates calls on external agents for detailed text and visual clues within the image, thus performing better reasoning. Furthermore, we propose P^{2}GB, a challenging benchmark with text-rich and high-definition images to better assess reasoning capabilities. Comprehensive experiments on a variety of datasets demonstrate the superiority of P^{2}G, especially under text-rich and high-definition images. Our work provides meaningful insights into the enhancement of MLLM reasoning capabilities with tool usage and plug-and-play grounding. We provide a detailed discussion on related works to P^{2}G in Appendix A.

Limitations

In this section, we discuss the limitations of the current work in detail, outlining future directions.

1) Noise in agents. It is a shared common challenge on the capability of external agents itself (Liang et al., 2023; Shen et al., 2023) in tool-augmented (M)LLMs. While we leverage state-of-the-art agents when building P^{2}G, it is possible that it returns noisy, biased, or inaccurate results. In the future, we may propose a post-agent-call filtration strategy, or explore recent advances like self-consistency (Wang et al., 2023b).

2) Token count. To incorporate finer multimodal semantics into contexts for grounded reasoning, P^{2}G inevitably leverages a longer context of input. To accommodate more tokens, we propose novel routing strategies for MLP or resampler-based token compression mechanisms. However, we believe it is also promising to explore enhancing P^{2}G with efficient sampling approaches, e.g. KV-Caching.

3) Modality-interleaved or multi-hop reasoning. Another limitation of current work and valuable future direction is to expand P^{2}G into multi-hop and complex reasoning that involves interleaved multi-modality clues. For future studies, we may explore expanding types of agents, and adapting tree (Yao et al., 2023) or graph-structured (Besta et al., 2024) reasoning or agent calling paths for supporting these more challenging scenarios.

Ethnics Statement

This work studies enhancing smaller MLLMs on visual reasoning via leveraging external agents and deliberate reasoning, which improves the reasoning capability of smaller MLLMs and potentially makes them more helpful by improving the accuracy and groundedness of their answers.

All visual images for creating our novel benchmark P^{2}GB are from publicly accessible resources, which we adequately cited in our paper. On corresponding verbal multiple-choice questions, for ones we adapted from existing works, we cite their sources in our paper; and we leverage a publicly-accessible model (GPT-4o) to synthesize the rest and manually double-check their correctness. The proposed benchmark will be publicly released.

References

2022. PaddleOCR.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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. *CoRR*, abs/2308.12966.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can AI assistants know what they don't know? *CoRR*, abs/2401.13275.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *CoRR*, abs/2301.00234.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA-02: A visual representation for neon genesis. *CoRR*, abs/2303.11331.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023a. Llama-adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. 2024. BLIVA: A simple multimodal LLM for better handling of text-rich visual questions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 2256–2264. AAAI Press.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. 2023. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, Yun Wang, Linjun Shou, Ming Gong, and Nan Duan. 2023. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *CoRR*, abs/2303.16434.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. volume abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023c. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. *CoRR*, abs/2303.05499.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29*

July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2263–2279. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*.

Ying Shen, Zhiyang Xu, Qifan Wang, Yu Cheng, Wenpeng Yin, and Lifu Huang. 2024. Multimodal instruction tuning with conditional mixture of lora. *CoRR*, abs/2402.15896.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengze Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng-Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2023. A survey of reasoning with foundation models. *CoRR*, abs/2312.11562.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *CoRR*, abs/2401.13311.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023a. Cogvlm: Visual expert for pretrained language models. *CoRR*, abs/2311.03079.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. 2023c. Towards improving document understanding: An exploration on text-grounding via mllms. *CoRR*, abs/2311.13194.

Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023d. Democratizing reasoning ability: Tailored learning from large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1948–1966. Association for Computational Linguistics.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *CoRR*, abs/2312.14135.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal LLM. *CoRR*, abs/2309.05519.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

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

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *CoRR*, abs/2306.17107.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for LLM question answering with external tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Related Works

A.1 Multimodal LLMs

The surge of large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023a), especially instruction-tuned ones (Longpre et al., 2023; Chiang et al., 2023; Touvron et al., 2023b; Mukherjee et al., 2023) demonstrated a strong potential in becoming generic interface for language modality. To extend LLMs beyond language perception, recent works (Zhu et al., 2023; Liu et al., 2023b; Huang et al., 2023; Alayrac et al., 2022; Wang et al., 2023a; Dai et al., 2023) extends them into Multimodal Large Language Models (MLLMs) with instruction tuning, through incorporating each modality as a foreign language(Huang et al., 2023; Wu et al., 2023). To equip LLM with capability in image perception, pioneer works like Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023b) first encode image with a dedicated model (e.g.ViT (Dosovitskiy et al., 2021)), then propose specific modules for aligning image and text modality. Subsequent works like LLaVA (Liu et al., 2023b) and KOSMOS-1 (Huang et al., 2023) leverage vision tokenizers to feed image semantics as in-context tokens, thereby aligns the perception of image and language. To further advance MLLMs, recent works explored enabling grounding and reference to visual contexts (Peng et al., 2024; Wang et al., 2023a), generating contents leveraging multimodal adaptors (Wu et al., 2023; Pan et al., 2024), leveraging parameter-efficient fine-tuning (Gao et al., 2023a; Shen et al., 2024), and scaling of multimodal instruction data and model parameters (OpenAI, 2023; Bai et al., 2023; Lu et al., 2024). Despite these improvements, MLLMs so far still suffers from multiple prevailing limitations, including high-demand on quality and quantity of instruction-following data, hallucination (Liu et al., 2024), and difficulties in processing images within text-rich

Figure 8: Case study of visual reasoning on P^{2}GB, where we compare rationales generated by P^{2}G and GPT-4V(ision). The first three lines from top to bottom demonstrate cases on both text-rich and semantic-rich reasoning, and bounding boxes generated with *OCR agent* and/or *Grounding Agent* of P^{2}G, where P^{2}G (based on LLaVA-7B) demonstrates its superior capability in generating grounded reasoning leveraging additional semantic clues against GPT-4V. The last row comprises two challenging failure cases where both P^{2}G and GPT-4V fails in generating an accurate answer.

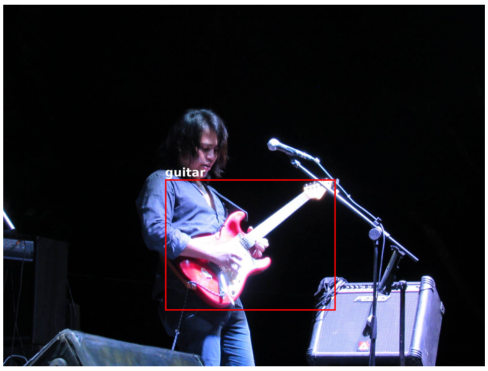
	Case #1	Case #2
Question	What is the color of the bowl on the counter? A. Blue B. Green C. White D. Silver	Is there any musical instrument seen on the stage? A. No, there isn't. B. Yes, there is a drum. C. Yes, there is a guitar. D. Yes, there is a piano.
Image Size	3264×2448	2048×1536
Agent Returns		
Final Prompt	Additional visual information to focus on: 1 bowl <object> at location [0.891,0.184,0.999,0.328] What is the color of the bowl on the counter? A. Blue B. Green C. White D. Silver Answer with the option's letter from the given choices directly.	Additional visual information to focus on: 1 guitar <object> at location [0.336,0.484,0.690,0.846] Is there any musical instrument seen on the stage?. No, there isn't. B. Yes, there is a drum. C. Yes, there is a guitar. D. Yes, there is a piano. Answer with the option's letter from the given choices directly.
Final Answer	P^{2}G (Ours): D LLaVa: B	P^{2}G (Ours): C LLaVa: B

Table 5: Two cases of Plug-and-Play grounding of P^{2}G to critical objects in high-resolution images.

contexts (Wang et al., 2023c) or grasping details within high-resolution images (Liu et al., 2023b).

A.2 Visual Reasoning in Text-Rich Images

Zhang et al. (2023) developed LLaVAR, which aims to enhance the interactive capabilities of MLLMs through improved visual instruction tuning for text-rich image understanding. Hu et al. (2024) introduce BLIVA, which employs a novel approach by integrating both learned query embeddings and image-encoded patch embeddings to enhance the multimodal LLM's understanding and processing of text-rich visual questions. Wang et al. (2023c) focus on enhancing MLLMs with text-grounding to improve document understanding, especially in text-rich scenarios. Despite employing extensive instruction fine-tuning data, the models' capability for text grounding remains limited. Wadhawan et al. (2024) emphasize the need for models to understand interactions between text and visual content in their evaluation of context-sensitive text-rich visual reasoning in large multimodal models. They primarily employ OCR tools and GPT-4 to construct instruction-finetuned datasets that enhance MLLM's visual reasoning of text-rich images; however, mere instruction fine-tuning struggles to effectively leverage LLM's potent generative capabilities, resulting in marginal improvements.

B Extended Case Study

To further understand plug-and-play grounding of reasoning in P^{2}G, we provide two case studies in Table 5 and 6. As illustrated in Table 5, P^{2}G could effectively utilize additional visual clues from Grounding Agent to improve its accuracy of answers, compared to LLaVA. As illustrated in Table 6, by providing textual clues from OCR Agent, the capability of P^{2}G in understanding tiny texts are also largely improved. These cases further highlights the effectiveness of P^{2}G's design.

	Case #3	**Case #4**
Question	How would you describe the general appearance of the buildings in the photo? A. Modern and sleek B. Colorful and unique C. Industrial and metallic D. Old and brick	How much alcohol is in this beverage?
Image Size	736 × 938	550 × 1200
Agent Returns	 *(no texts detected in the image)*	
Final Prompt	Additional visual information to focus on: *Please focus on providing an answer to the question without considering any challenges related to the clarity or presence of text in the image.* How would you describe the general appearance of the buildings in the photo? A. Modern and sleek B. Colorful and unique C. Industrial and metallic D. Old and brick Answer with the option's letter from the given choices directly. *(no texts detected in the image)*	Additional visual information to focus on: text in the image: 'CARLING' at location [0.107, 0.285, 0.658, 0.559]; 'OFTASTE AND' at location [0.156, 1.297, 0.295, 1.328]; 'ALC4.1%VOL' at location [0.177, 1.619, 0.278, 1.649]; 'ENJOY EXTRA' at location [0.177, 1.619, 0.278, 1.649]; 'COLD' at location [0.205, 1.647, 0.247, 1.67] How much alcohol is in this beverage?
Final Answer	P^2G (Ours): D LLaVa: A	P^2G (Ours): 4.1% LLaVa: 2%

Table 6: Two cases of Plug-and-Play grounding of P^2G to critical texts that tiny in its scale. *Left*: when no texts are detected by OCR agent, we inform the model and encourage it to focus on non-textual semantics. *Right*: when critical texts are detected, we incorporate them with their relative position in multimodal query.