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

Jiaxing Chen[◇1*], Yuxuan Liu[◇1], Dehu Li[1]
Xiang An[2], Ziyong Feng[2], Yongle Zhao[2], Yin Xie[2]

[1]Peking University, Beijing, China
{jiaxing.chen, yx.liu, dehuli}@stu.pku.edu.cn
[2]DeepGlint, Beijing, China

Abstract

The surge of Multimodal Large Language Models (MLLMs), given their prominent emergent capabilities in instruction following and reasoning, has greatly advanced the field of visual reasoning. However, constrained by their non-lossless image tokenization, most MLLMs fall short of comprehensively capturing details of text and objects, especially in high-resolution images. To address this, we propose P^{2}G, a novel framework for plug-and-play grounding of reasoning in MLLMs. Specifically, P^{2}G exploits the tool-usage potential of MLLMs to employ expert agents to achieve on-the-fly grounding to critical visual and textual objects of image, thus achieving deliberate reasoning via multimodal prompting. We further create P^{2}GB, a benchmark aimed at assessing MLLMs' ability to understand inter-object relationships and text in challenging high-resolution images. Comprehensive experiments on visual reasoning tasks demonstrate the superiority of P^{2}G. Noteworthy, P^{2}G achieved comparable performance with GPT-4V on P^{2}GB, with a 7B backbone. Our work highlights the potential of plug-and-play grounding of reasoning and opens up a promising alternative beyond model scaling[2].

1 Introduction

The rise of large language models (LLMs) [35, 28, 36] demonstrates its strong potential in becoming a unified backbone for almost any task in language modality, given their promising emergent capabilities like in-context learning [4, 39], instruction following [29], reasoning [34], and beyond. To extend LLMs to a general interface beyond language modality, a surging trend of work [46, 19, 13, 2, 38, 7] extends them into Multimodal Large Language Models (MLLMs), through incorporating each modality as a foreign language [13, 42]. Being able to perceive multimodal input and leverage instruction following and in-context learning, MLLMs achieve significant results in the realm of visual reasoning.

However, despite these promising achievements, there remain significant limitations for MLLMs in visual reasoning. One crucial drawback rooted within the high demand for high-quality, large-scaled annotated data for vision instruction tuning [46, 19]. Compared to pure language modality, it is conceivably harder to collect annotated multimedia training examples or generate synthesized ones. Worse, the demand for multimodal instruction tuning data poses a greater challenge to scaling of MLLMs.

Another limitation lies in the challenges of capturing details comprehensively, especially when dealing with high-resolution images or those containing complex textual information due to the rich

[*]Work done during Jiaxing's internship at DeepGlint. ◇: Equal contribution.
[2]Project page: bpvovqd.github.io/p2g.

Preprint. Work in progress.

details these images entail, resulting in hallucinations and/or incorrect reasoning solutions. And the non-lossless tokenization of images would inevitably overlook critical semantic details within images, due to fixed, and normally small input resolutions.

To overcome these limitations, successor works explore strategies for grounding reasoning in MLLMs. Particularly, to ground reasoning in semantic objects, KOSMOS-2 [31] finetunes MLLM to generate bounding boxes for visual occurrences in context, a training strategy that has also been applied in later works like CogVLM [38]. For another challenging scenario, text-rich visual reasoning, recent works like LLaVAR [45] and TGDoc [40] augment instruction tuning data with OCR-based textual clues and corresponding bounding boxes. However, a common problem faced by these methods is the need for large amounts of instruction data and training costs, which significantly limits their application.

To achieve grounding, the above methods invariably train MLLMs to equip them with this capability from scratch, which is undoubtedly challenging and less efficient. Many recent studies have shown that LLMs can effectively utilize external tools and agents [32, 47]. Drawing inspirations from the above, we propose P^{2}G, a novel framework that achieves Plug-and-Play Grounding of reasoning in MLLMs. Beyond supervised fine-tuning of MLLM itself, we leverage state-of-the-art, lightweight proxy models as agents for obtaining critical clues for reasoning. Particularly, we propose an OCR agent (via PaddleOCR [1]) and a visual grounding agent (via Grounding-DINO [21]), aiming at challenging text-rich and high-definition images. We then instruct MLLMs to generate specific queries in need of supporting textual or visual clues, based on the intrinsic complexity of the reasoning task.

To better assess P^{2}G on aforementioned text-rich and/or high-definition images, we introduce P^{2}GB, a challenging Visual Question Answering (VQA) benchmark, which is expressly designed to assess MLLM's visual grounding, especially identifying multiple objects of the same category within high-resolution images and to enhance the comprehension of textual content in text-rich scenarios. Our comprehensive experiments on visual reasoning tasks including P^{2}GB demonstrate the superiority of P^{2}G. Noteworthy, P^{2}G achieved comparable performance with GPT-4V on P^{2}GB, with a 7B backbone. Our work highlights the potential of plug-and-play grounding of reasoning and opens up a promising alternative beyond model scaling. In summary, our contributions are three-fold:

1) We propose P^{2}G, a novel framework for plug-and-play grounding of reasoning in high-resolution natural and text-rich visual reasoning scenarios, through leveraging agents for enhanced textual and visual grounding and perception.

2) We introduce P^{2}GB, a VQA benchmark designed to thoroughly assess MLLM's reasoning capability under text-rich and high-definition image queries.

3) We conduct extensive experiments on text-rich reasoning datasets to verify the superior performance of P^{2}G. Empowered with P^{2}G, we surpass similar scaled (7B) or even larger model (13B) with a 7B MLLM backbone, demonstrating the significance of P^{2}G.

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 require 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

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.

emergent capabilities like *in-context learning* [8], *instruction following* [23] and *tool-usage* [32] capability of large language models. Below, we introduce the procedure of P^2G 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, it is crucial for MLLMs to know their *don't-knows* [5] ahead. To mitigate this issue, we propose Deliberate Reasoning in P^2G, 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 and straightforward visual query, P^2G generates the correct answer directly, while for challenging cases, P^2G 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) [11] and tool-usage [32, 17] 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 [18, 19, 43] 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 auto-encoding compression refrains MLLM from generating grounded, accurate reasoning. Latest works either finetune on more VQA data [45], or prepend OCR texts into context [40, 22], which does not essentially mitigate this core limitation.

As a step forward, we propose *Plug-and-Play* Grounding in P^2G, 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 sementic 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 an final comprehensive reasoning answer. Such plug-and-play design enables us to leverage SOTA text (PaddleOCR [1]) or image (Grounding DINO [21]) processing tools, mitigating the demand to 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

In this subsection, we dive deeper into the architectural implementation and design of P^2G. Specifically, our MLLM is composed of four components: an LLM, a vision encoder, a projection module, and textual (OCR) and visual grounding agents, which collectively enhance the model's capability to process and interpret complex multimodal data.

In this work, our LLM is powered by Vicuna-7B [6], which has been trained on approximately 400K high-quality, instruction-following samples, incorporating 150K conversations between users and GPT [6]. We have chosen the CLIP ViT-L/14 as the vision encoder, with inputs resized and padded to 224^2. We apply this encoder to process not only the original images, but also the specific regions cropped from the image that contain detected objects (from Grounding Agent).

In the process of mapping visual characteristics to the hidden space of the LLM, we utilize two types of projection modules: an MLP and a Resampler. The MLP preserves the count of visual tokens as outputted by the vision encoder, whereas the Resampler, functioning on cross-attention principles, diminishes the token quantity (i.e. 256 to 32).

To efficiently manage the sequence length of the LLM when processing various content and visual features, we've devised a straightforward strategy that allows for seamless toggling between two projection modules. For inputs consisting solely of initial image features without cropped areas, the MLP is utilized to map all visual tokens. In instances where 1 to 6 critical objects are detected, indicating a need for the model to concentrate on these specific targets, we apply the MLP to the visual features of these objects and use the Resampler to downsample the overarching image features. Conversely, when more than 6 objects are detected, the Resampler is employed across all visual features to mitigate computational demands. The Grounding Agent employs Grounding DINO [21] to identify and extract objects from images that are relevant to the query, while the OCR Agent utilizes PaddleOCR[1] to discern and retrieve potential textual information embedded within the image query.

2.2.1 Reasoning with Plug-and-Play Grounding

In this subsection, we elaborate in detailed procedures for plug-and-play grounding of reasoning in P^{2}G. As illustrated in Figure 1, we first perform dedicated reasoning with MLLMs to identify whether additional visual or textual clues are needed for reasoning. In scenarios involving straightforward images (e.g. clear main objects, headline texts only, etc.), the model directly output its reasoning. Conversely, when confronted with high-resolution natural images or images abundant with detailed textual content, the MLLM would generate dedicated query responses, calling *OCR Agent* or *Grounding Agent*. This deliberate reasoning capability is gained through instruction fine-tuning, which we will elaborate in Section 2.3.

Sorry, I cannot answer the question. Some visual information about the following objects is missing or unclear: $\text{object}_1, \ldots, \text{object}_n$.

Table 1: Query response for calling *Grounding Agent* for detailed visual clues.

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

Table 2: Query response for calling *OCR Agent* for detailed textual clues.

For instances involving high-resolution natural images, the model's initial response is documented in Table 1. This demonstrates that the MLLM may not effectively perceive the presence of certain objects or details within these high-resolution images. To address this, we employ Grounding DINO to detect and crop these objects. The cropped images are then magnified relative to their original size, which facilitates a more focused analysis. An example of how these crops are incorporated into prompts for a subsequent round of inference is illustrated in Table 3. This tailored approach enables MLLM to provide more accurate and grounded answers given high-resolution images.

This operation is formalized through a detection function, denoted F_d, which processes an input image I and a specified set of target objects $\{object_1, \ldots, object_n\}$, resulting in a set of image crops P:

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

where the set $P = \{p_1, p_2, \ldots, p_m\}$ consists of the image crops identified by Grounding DINO. The relationship between the total number of objects and the individual quantities of each type of object is given by the equation $\sum_{i=1}^{n} x_i = m$, where n represents the total number of object types and x_i denotes the quantity of the i-th object.

Likewise, when dealing with high-resolution text-rich images, the model's call to *OCR Agent* is shown in Table 2. Subsequently, we employ PaddleOCR to extract textual elements from the images. The extracted OCR tokens, along with their corresponding bounding boxes and the posed questions, are then cohesively integrated, as exemplified in Table 4. By incorporating the OCR tokens and

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

<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, 0.71] (Object features and their positions)

Are all buttons in the image larger than the paper clips?
Answer the question using a single word or phrase. (Original question)

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

<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)

By whom is this letter written? (Original question)

Table 4: Example prompt for the model's second round of reasoning with additional textual clues from *OCR Agent*.

their bounding boxes into the prompts presented to the MLLM, we enhance the model's capability to recognize the presence and positions of text within the given images.

Given additional textual clues $\mathcal{T}$ and visual clues $\mathcal{P}$ obtained via external agents, we obtain our 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 demote image and text query, respectively. By conditioning on both image q_i and enriched information $\mathcal{T}$ and $\mathcal{P}$ on semantics, we thereby achieve plug-and-play grounding of reasoning, leveraging in-context learning and instruction following capabilities of MLLMs.

2.3 Instruction-following Data for Fine-tuning

The instruction tuning process of P^2G is structured into two distinct phases, namely multimodal instruction tuning and learning of deliberate reasoning.

2.3.1 Multimodal Instruction Tuning

In the first phase, we equip our backbone LLM (Vicuna-7B [6]) with fundamental multimodal capabilities, following procedures in LLaVA [19]. Specifically, we apply a 120K sampling from LLaVA instruction data, following the exact same procedure and splits of V* [41].

2.3.2 Learning of Deliberate Reasoning

In the second phase, we tune MLLM to utilize agents with deliberate reasoning ahead, as elaborated in Section 2.2. For both agents, we meticulously curated a set of challenging Visual Question Answering (VQA) inquiries to serve as both positive and negative samples. For the negative samples, the model explicitly acknowledges its inability to provide an answer, necessitating the provision of additional detailed information. Conversely, positive samples, whether they are simpler VQA questions or complex inquiries, are directly answered, with the latter being supplemented by additional detailed information. Following the procedures outlined in Section 2.2, we prepared the initial agent interactions (*round 1*), followed by multimodal prompt instructions (*round 2*). Details on dataset selection are listed as follows.

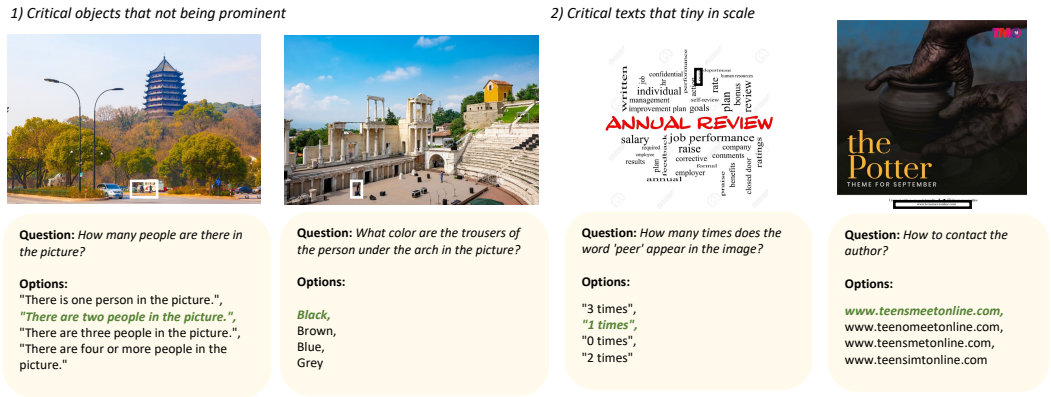

Figure 2: Illustration of our proposed P^{2}GB benchmark. In P^{2}GB, we consider two challenging visual reasoning scenarios: comprehensive image understanding and text-rich visual reasoning. For the former, we delicately collect high-definition image samples where the critical object is not prominent (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.

Negative data (54k) & Multimodal instruction data (26k) for text-rich image reasoning In the context of certain simplistic text-based image scenarios, such as identifying the title of a book from its cover, straightforward answers can be directly provided. However, for more complex text-based image queries, where the text within the image is abundant and the questions cannot be answered solely based on the text present in the image, one must leverage the powerful zero-shot or inferential capabilities of LLMs to respond further. In these instances, the direct responses from an MLLM are insufficient, necessitating additional supplementary information to effectively address the query. Building on the aforementioned complexities of text-rich image analysis, we curated a subset of data from ChartVQA, DOCVQA, and TextVQA to serve as positive and negative samples. For scenes that presented a greater challenge to address, we selected images with resolutions exceeding 500 pixels and then employed PaddleOCR to extract text from these images. Textually sparse images were filtered out using a predetermined threshold. For the remaining dataset, 70% was designated as negative samples, *where the model explicitly acknowledges its inability to answer the query based solely on the image content*, indicating a necessity for further textual information from within the image. The remaining 30% of the data was augmented with OCR-extracted text as input for the second iteration of inferential VQA processing. Data filtered out through the initial two processes were directly employed as VQA data for the first inference process, ensuring a structured approach to handling varying complexities of text-based image queries.

Negative data (100k) & Multimodal instruction data (167k) for visual objects grounding in reasoning These two data segments were obtained from V* [41]; however, we have transformed them to improve the model's understanding of both quantitative relationships and spatial arrangements between objects by integrating the extracted number of objects and their ordered bounding boxes into the training dataset.

3 P^{2}GB Benchmark

To quantitatively assess the visual reasoning capabilities of MLLMs under text-rich or high-resolution scenarios, we constructed a challenging benchmark test named P^{2}GB. In order to quantitatively evaluate the abilities of Multimodal Large Language Models (MLLMs) to process images containing rich and complex information scenes, we devised a novel benchmark test that includes two challenging tasks: Comprehensive Image Understanding with Fine-grained Recognition (101 samples), and Image Text Content Understanding (50 samples), with a total of 151 samples.

Comprehensive Image Understanding with Fine-grained Recognition This task involves a deep analysis of high-resolution images from high-quality datasets, which depict rich and complex

scenes with multiple objects of the same or different types. The challenge for the model is to identify the presence and characteristics of various objects within the image, which includes but is not limited to the type, location, and possible interactions of the objects. This not only tests the model's capability in recognizing and distinguishing objects within the image but also requires the model to understand how these objects collectively form a scene in the image, especially when objects may be difficult to identify due to their small size or their color blending into the background.

Image Text Content Understanding Here, the model is challenged to identify and understand small textual content within images and to answer questions related to this information. The model must be capable of discerning fine text in high-resolution images and engage in logical reasoning to correctly respond to inquiries based on this textual content.

For the quantitative comparison of these tasks, we designed multiple-choice answers for each question. These options have been carefully crafted and manually reviewed to ensure the validity and fairness of the test and to eliminate potential ambiguities between choices.

4 Experiments

4.1 Experimental Setup

Models and Baselines For MLLMs, we select Vicuna-7B [6] as the language backbone, and follow LLaVA to train an MLLM backbone for P^{2}G[3]. To build up two agents for visual and textual grounding, we select Grounding DINO [21] for obtaining visual clue (i.e., objects) and PaddleOCR [1] for screening texts within the image query. We compare P^{2}G against multiple similar-scaled, instruction-tuned MLLMs, including vanilla LLaVA [19], MiniGPT-4 [46], mPLUG-OWL [43], and Instruct-BLIP [7]. In addition, we compare P^{2}G against MLLMs dedicated optimized for semantic-rich reasoning, i.e., SEAL [41], LLaVAR [45], and TGDoc [40]. Finally, we include the most capable MLLM so far, GPT-4V [28] on our challenging benchmark P^{2}GB.

Datasets Following previous works, we test P^{2}G on a variety of visual reasoning benchmarks. For text-rich visual reasoning, we select DocVQA [26] and ChartVQA [25], and GQA [14], SEED [15], MM-VET [44], and MME [15] for semantic-rich and general visual reasoning. Beyond existing benchmark, 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 instruction-finetuning the 7B version of LLaVA model. Specifically, we employ general visual instruction tuning data proposed in LLaVA [19], VQA data from train sets of benchmarks, and our self-curated Negative data for learning on deliberate reasoning (generating calls to grounding agents). Details are elaborated in Sec. 2.3. We finetune our models on 8 A100 GPUs, with a learning rate of $2e^{-5}$, batch size of 16, for one epoch, with cosine scheduler and Adam optimizer. Pre-training was executed on the 558K image subset from LAIONCC-SBU, as utilized in LLaVA, with subsequent fine-tuning performed on a 467K dataset, comprising 154K negative samples, as delineated in Section 2.3.

4.2 Results

4.2.1 Performance on Visual Reasoning

The performance of P^{2}G on visual reasoning benchmarks are presented in Table 6. 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 InstrtuctBLIP, demonstrating the superiority of P^{2}G.

[3]Due to compute resource constraints, we train Vicuna-7B on a 120K subset of LLaVA instructions following [41]. Details are elaborated in Section 2.3

Model	Size	DocVQA	ChartVQA	GQA	SEED	MMVET	MME
MiniGPT-4 [46]	7B	3.0	4.3	-	-	-	-
mPLUG-OWL [43]	7B	6.9	9.5	-	-	-	-
LlaVAR [45]	7B	11.6	8.0	-	-	-	-
TGDoc [40]	7B	9.0	12.72	-	-	-	-
LLaVA [19]	7B	19.06	15.30	21.80	25.66	27.20	1151
Instruct-BLIP [7]	7B	-	-	49.20	-	26.20	-
LLaVA [19]	13B	31.77	25.70	17.82	17.01	33.90	1224
Instruct-BLIP [7]	13B	-	-	49.50	-	25.60	-
LLaVA + P^{2}G (Ours)	7B	**61.44**	**37.20**	**59.87**	**27.47**	**30.40**	**1223**

Table 5: Experimental results of P^{2}G and baselines on visual reasoning benchmarks. The best performing 7B-scaled MLLM is marked in **bold**.

Model	Size	Objects	Texts
GPT-4V [28]	>1T	36.0	**68.0**
SEAL(V*) [41]	7B	19.6	30.0
LLaVA	7B	12.3	8.0
LLaVA + P^{2}G (Ours)	7B	**39.7**	50.0
Gain (%)	-	3.2×	6.3×

Table 6: Experimental results on our challenging high-resolution benchmark P^{2}GB.

4.2.2 Performance on P^{2}GB

On the more challenging P^{2}GB, P^{2}Gachieved a significant improvement over V* and LLaVA, demonstrating a markedly enhanced comprehension of object details in high-resolution images by over $5x$ compared with vanilla LLaVA. Remarkably, P^{2}G even outperforms GPT-4V 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 3.

4.2.3 Ablation Study

We study the effect of P^{2}G in Table 7. We first remove the two agents for plug-and-play grounding (w/o Grounding 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[4]. We finally replaced the grounding agent with a model that being continually trained. It yields 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 Related Work

5.1 Multimodal LLMs

The surge of large language models (LLMs) [28, 35], especially instruction-tuned ones [23, 6, 36, 27] demonstrated a strong potential in becoming generic interface for language modality. To extend LLMs beyond language perception, recent works [46, 19, 13, 2, 38, 7] extends them into Multimodal Large Language Models (MLLMs) with instruction tuning, through incorporating each modality as

[4]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 a MLLM with longer context.

Figure 3: 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.

Benchmark	P^2G	w/o Position in Prompt	w/ Improved DINO	w/o Grounding 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)	27.5 (0.0)	25.7 (-1.8)
MM-VET	29.3	29.1 (-0.2)	30.4 (+1.1)	27.2 (-2.1)
GQA	59.9	59.8 (-0.1)	60.0 (+0.1)	21.8 (-28.0)

Table 7: Effects on removing the relative position vector for grounded (text and/or visual) objects in prompt (*w/o Position in Prompt*), replacing the visual grounding agent with a stronger, continue fine-tuned DINO (*w/ Improved DINO*), and removing the two agents in P^2G (w/o Grounding Agents).

a foreign language[13, 42]. To equip LLM with capability in image perception, pioneer works like Flamingo [2] and BLIP-2 [16] first encode image with a dedicated model (e.g.ViT [9]), then propose specific modules for aligning image and text modality. Subsequent works like LLaVA [19] and KOSMOS-1 [13] 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 [31, 38], generating contents leveraging multimodal adaptors [42, 30], leveraging parameter-efficient fine-tuning [10, 33], and scaling of multimodal instruction data and model parameters [28, 3, 24]. Despite these improvements, MLLMs so far still suffers from multiple prevailing limitations, including high-demand on quality and quantity of instruction-following data, hallucination [20], and difficulties in processing images within text-rich contexts [40] or grasping details within high-resolution images [19].

5.2 Visual Reasoning in Text-Rich Images

Zhang et al. [45] developed LLaVAR, which aims to enhance the interactive capabilities of MLLMs through improved visual instruction tuning for text-rich image understanding. Hu et al. [12] 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. [40] 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. [37] 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 finetuning struggles to effectively leverage LLM's potent generative capabilities, resulting in marginal improvements.

6 Conclusion

In this paper, we focus on the challenge in 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^2G, a novel framework for plug-and-play grounding of visual. Dedicately tuned on deliberate thinking, P^2G promptly generate calls on external agents for detailed text and visual clues within image, thus performing better reasoning. Furthermore, we propose P^2GB, a challenging benchmark with text-rich and high-definition images to better assess reasoning capabilities. Comprehensive experiments on a variety of datasets demonstrates the superiority of P^2G, 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.

References

[1] Paddleocr, 2022. https://github.com/PaddlePaddle/PaddleOCR.

[2] ALAYRAC, J.-B., DONAHUE, J., LUC, P., MIECH, A., BARR, I., HASSON, Y., LENC, K., MENSCH, A., MILLICAN, K., REYNOLDS, M., ET AL. Flamingo: a visual language model for few-shot learning. *NIPS* (2022).

[3] BAI, J., BAI, S., YANG, S., WANG, S., TAN, S., WANG, P., LIN, J., ZHOU, C., AND ZHOU, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* (2023).

[4] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *NIPS* (2020).

[5] CHENG, Q., SUN, T., LIU, X., ZHANG, W., YIN, Z., LI, S., LI, L., CHEN, K., AND QIU, X. Can ai assistants know what they don't know? *arXiv:2401.13275* (2024).

[6] CHIANG, W.-L., LI, Z., LIN, Z., SHENG, Y., WU, Z., ZHANG, H., ZHENG, L., ZHUANG, S., ZHUANG, Y., GONZALEZ, J. E., STOICA, I., AND XING, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[7] DAI, W., LI, J., LI, D., TIONG, A., ZHAO, J., WANG, W., LI, B., FUNG, P. N., AND HOI, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NIPS* (2023).

[8] DONG, Q., LI, L., DAI, D., ZHENG, C., WU, Z., CHANG, B., SUN, X., XU, J., AND SUI, Z. A survey for in-context learning. *arXiv:2301.00234* (2022).

[9] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G., GELLY, S., ET AL. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[10] GAO, P., HAN, J., ZHANG, R., LIN, Z., GENG, S., ZHOU, A., ZHANG, W., LU, P., HE, C., YUE, X., ET AL. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* (2023).

[11] GAO, Y., XIONG, Y., GAO, X., JIA, K., PAN, J., BI, Y., DAI, Y., SUN, J., AND WANG, H. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997* (2023).

[12] HU, W., XU, Y., LI, Y., LI, W., CHEN, Z., AND TU, Z. Bliva: A simple multimodal llm for better handling of text-rich visual questions. *AAAI* (2024).

[13] HUANG, S., DONG, L., WANG, W., HAO, Y., SINGHAL, S., MA, S., LV, T., CUI, L., MOHAMMED, O. K., PATRA, B., ET AL. Language is not all you need: Aligning perception with language models. *NIPS* (2024).

[14] HUDSON, D. A., AND MANNING, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR* (2019).

[15] LI, B., WANG, R., WANG, G., GE, Y., GE, Y., AND SHAN, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125* (2023).

[16] LI, J., LI, D., SAVARESE, S., AND HOI, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (2023), PMLR, pp. 19730–19742.

[17] LIANG, Y., WU, C., SONG, T., WU, W., XIA, Y., LIU, Y., OU, Y., LU, S., JI, L., MAO, S., WANG, Y., SHOU, L., GONG, M., AND DUAN, N. Taskmatrix.ai: Completing tasks by connecting foundation models with millions of apis. *Intelligent Computing* (2024).

[18] LIU, H., LI, C., LI, Y., AND LEE, Y. J. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following* (2023).

[19] LIU, H., LI, C., WU, Q., AND LEE, Y. J. Visual instruction tuning. *NIPS* (2024).

[20] LIU, H., XUE, W., CHEN, Y., CHEN, D., ZHAO, X., WANG, K., HOU, L., LI, R., AND PENG, W. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

[21] LIU, S., ZENG, Z., REN, T., LI, F., ZHANG, H., YANG, J., LI, C., YANG, J., SU, H., ZHU, J., ET AL. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499* (2023).

[22] LIU, X., TANG, W., NI, X., LU, J., ZHAO, R., LI, Z., AND TAN, F. What large language models bring to text-rich vqa? *arXiv:2311.07306* (2023).

[23] LONGPRE, S., HOU, L., VU, T., WEBSON, A., CHUNG, H. W., TAY, Y., ZHOU, D., LE, Q. V., ZOPH, B., WEI, J., ET AL. The flan collection: Designing data and methods for effective instruction tuning. *ICML* (2023).

[24] LU, H., LIU, W., ZHANG, B., WANG, B., DONG, K., LIU, B., SUN, J., REN, T., LI, Z., SUN, Y., ET AL. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525* (2024).

[25] MASRY, A., LONG, D. X., TAN, J. Q., JOTY, S., AND HOQUE, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ACL* (2022).

[26] MATHEW, M., KARATZAS, D., AND JAWAHAR, C. Docvqa: A dataset for vqa on document images. In *CVPR* (2021).

[27] MUKHERJEE, S., MITRA, A., JAWAHAR, G., AGARWAL, S., PALANGI, H., AND AWADALLAH, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707* (2023).

[28] OPENAI. Gpt-4 technical report. *arXiv:2303.08774* (2023).

[29] OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. Training language models to follow instructions with human feedback. *NIPS* (2022).

[30] PAN, X., DONG, L., HUANG, S., PENG, Z., CHEN, W., AND WEI, F. Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations* (2023).

[31] PENG, Z., WANG, W., DONG, L., HAO, Y., HUANG, S., MA, S., YE, Q., AND WEI, F. Grounding multimodal large language models to the world. In *ICLR* (2024).

[32] SHEN, Y., SONG, K., TAN, X., LI, D., LU, W., AND ZHUANG, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *NIPS* (2024).

[33] SHEN, Y., XU, Z., WANG, Q., CHENG, Y., YIN, W., AND HUANG, L. Multimodal instruction tuning with conditional mixture of lora. *arXiv preprint arXiv:2402.15896* (2024).

[34] SUN, J., ZHENG, C., XIE, E., LIU, Z., CHU, R., QIU, J., XU, J., DING, M., LI, H., GENG, M., ET AL. A survey of reasoning with foundation models. *arXiv:2312.11562* (2023).

[35] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., ET AL. Llama: Open and efficient foundation language models. *arXiv:2302.13971* (2023).

[36] TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S., ET AL. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288* (2023).

[37] WADHAWAN, R., BANSAL, H., CHANG, K.-W., AND PENG, N. Contextual: Evaluating context-sensitive text-rich visual reasoning in large multimodal models. *arXiv:2401.13311* (2024).

[38] WANG, W., LV, Q., YU, W., HONG, W., QI, J., WANG, Y., JI, J., YANG, Z., ZHAO, L., SONG, X., ET AL. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079* (2023).

[39] WANG, X., WEI, J., SCHUURMANS, D., LE, Q. V., CHI, E. H., NARANG, S., CHOWDHERY, A., AND ZHOU, D. Self-consistency improves chain of thought reasoning in language models. In *ICLR* (2023).

[40] WANG, Y., ZHOU, W., FENG, H., ZHOU, K., AND LI, H. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv:2311.13194* (2023).

[41] WU, P., AND XIE, S. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv:2312.14135* (2023).

[42] WU, S., FEI, H., QU, L., JI, W., AND CHUA, T.-S. Next-gpt: Any-to-any multimodal llm. *arXiv:2309.05519* (2023).

[43] YE, Q., XU, H., XU, G., YE, J., YAN, M., ZHOU, Y., WANG, J., HU, A., SHI, P., SHI, Y., ET AL. mplug-owl: Modularization empowers large language models with multimodality. *CVPR* (2024).

[44] YU, W., YANG, Z., LI, L., WANG, J., LIN, K., LIU, Z., WANG, X., AND WANG, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv:2308.02490* (2023).

[45] ZHANG, Y., ZHANG, R., GU, J., ZHOU, Y., LIPKA, N., YANG, D., AND SUN, T. Enhanced visual instruction tuning for text-rich image understanding. In *NIPS Workshop* (2023).

[46] ZHU, D., CHEN, J., SHEN, X., LI, X., AND ELHOSEINY, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR* (2024).

[47] ZHUANG, Y., YU, Y., WANG, K., SUN, H., AND ZHANG, C. Toolqa: A dataset for llm question answering with external tools. *NIPS* (2024).