Understand, Think, and Answer: Advancing Visual Reasoning with Large Multimodal Models

Yufei Zhan[1,2,*], **Hongyin Zhao**[1,*], **Yousong Zhu**[1,✉], **Shurong Zheng**[1,3], **Fan Yang**[1,3], **Ming Tang**[1,2], **Jinqiao Wang**[1,2,3,4]
[1] Foundation Model Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Peng Cheng Laboratory, Shenzhen, China [4] Wuhan AI Research, Wuhan, China
{zhanyufei2021, zhaohongyin2020, zhengshurong2023, yangfan_2022}@ia.ac.cn
{yousong.zhu, tangm, jqwang}@nlpr.ia.ac.cn

Abstract

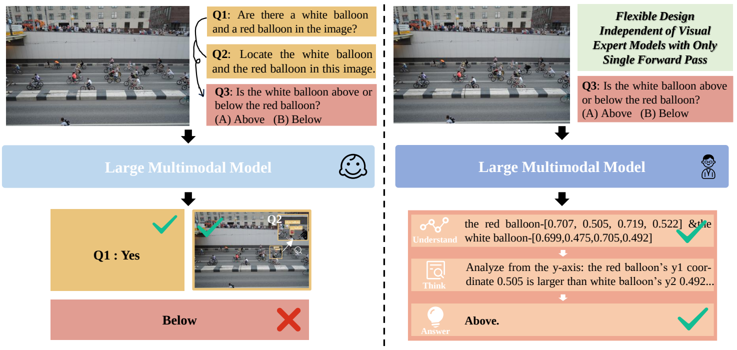
Large Multimodal Models (LMMs) have recently demonstrated remarkable visual understanding performance on both vision-language and vision-centric tasks. However, they often fall short in integrating advanced, task-specific capabilities for compositional reasoning, which hinders their progress toward truly competent general vision models. To address this, we present a unified visual reasoning mechanism that enables LMMs to solve complicated compositional problems by leveraging their intrinsic capabilities (e.g. grounding and visual understanding capabilities). Different from the previous shortcut learning mechanism, our approach introduces a human-like understanding-thinking-answering process, allowing the model to complete all steps in a single pass forwarding without the need for multiple inferences or external tools. This design bridges the gap between foundational visual capabilities and general question answering, encouraging LMMs to generate faithful and traceable responses for complex visual reasoning. Meanwhile, we curate 334K visual instruction samples covering both general scenes and text-rich scenes and involving multiple foundational visual capabilities. Our trained model, Griffon-R, has the ability of end-to-end automatic understanding, self-thinking, and reasoning answers. Comprehensive experiments show that Griffon-R not only achieves advancing performance on complex visual reasoning benchmarks including VSR and CLEVR, but also enhances multimodal capabilities across various benchmarks like MMBench and ScienceQA. Data, models, and codes will be release at `https://github.com/jefferyZhan/Griffon/tree/master/Griffon-R` soon.

1 Introduction

Inspired by the success of Large Language Models like ChatGPT [60] and Gemini [76], the vision field has been seeking to equip these models with visual understanding capabilities, aiming to replicate similar achievements in visual tasks. Currently, Large Multimodal Models (LMMs) [44, 34, 12, 37, 40] adopt a paradigm in which images are encoded and projected into a textual embedding space, then combined with language input to generate responses via the LLM [11, 14]. Trained with millions of high-quality data, LMMs demonstrate advancing performance across various vision-language tasks, such as visual question answering (VQA) [18] and image captioning [10], and become increasingly

[*] Equal Contribution.

Preprint. Under review.


Q1: Are there a white balloon and a red balloon in the image?
Q2: Locate the white balloon and the red balloon in this image.
Q3: Is the white balloon above or below the red balloon?
(A) Above (B) Below
Large Multimodal Model
Q1 : Yes
Q2
Below
Flexible Design Independent of Visual Expert Models with Only Single Forward Pass
Q3: Is the white balloon above or below the red balloon?
(A) Above (B) Below
Large Multimodal Model
Understand
the red balloon-[0.707, 0.505, 0.719, 0.522] &the white balloon-[0.699,0.475,0.705,0.492]
Think
Analyze from the y-axis: the red balloon's y1 coordinate 0.505 is larger than white balloon's y2 0.492...
Answer
Above.

(a) Based on shortcut learning mechanism

(b) Based on the proposed unified visual reasoning mechanism

Figure 1: Enabled by the proposed unified mechanism, Griffon-R naturally connects the reasoning processes for locating each balloon with answering the spatial relationship question. It effectively analyzes their y-axis coordinates and provides the correct answer in a single pass.

proficient in fine-grained visual tasks like visual grounding [63] and object detection [39], even surpassing specialized vision expert models [68, 27] in certain domains.

Despite significant progress across a wide range of tasks, LMMs still fall short in visual reasoning tasks. Existing open-source LMMs mainly follow the shortcut learning paradigm[15] and are trained to directly generate the final answer based on the question. As shown in Fig. 1(a), LMMs do well in visual foundational tasks that follow the shortcut paradigm like object recognition(Q1) and visual grounding(Q2). Though this paradigm benefits the LMMs a lot, it also hinders them from inferring based on the foundational visual capabilities, which are crucial for tackling compositional and complicated visual reasoning tasks. As indicated in Fig. 1(a), when directly asking the LMMs about the relative position of the red and white balloons, they respond with the unfaithful but confident answer, *i.e.* hallucination [72].

To address these challenges, recent studies have shown that incorporating multimodal Chain-of-Thought (CoT) [70, 36, 58] to encourage LMMs to reason step by step can improve the reason quality. However, these CoT methods are specifically designed for target types of questions or domains and may require multiple times of forwarding. Other toolkit-based methods [65, 24, 83] enhance LMMs by directly supplementing details needed for specific tasks using the visual toolkit with the format of structured text or feature-based prompts. Though the visual toolkit can provide accurate information, calling it brings significant computational load and latency. Also, with the task complexity increasing, the parameters scale up exponentially.

In this paper, we propose a unified visual reasoning mechanism to enable LMMs to harness advanced intrinsic visual foundational capabilities for compositional visual reasoning in a single forward pass. Inspired by the human reasoning process[2, 56]—where individuals begin by contemplating how to answer a question, gather sufficient information from their environment, think with their experiences, and ultimately formulate an answer—our unified mechanism integrates this progressive "understand-think-answer" approach. Following this proposed process, the model first plans the necessary information acquisition for answering the question and generates structured instructions to autonomously gather relevant information, ensuring a thorough understanding. Considering the contextual understanding and extensive knowledge base of LLMs, rather than designing question-specific reasoning paths, the model is self-prompted to engage in contextual thinking after obtaining a comprehensive understanding. This design allows for greater flexibility across various question types. Finally, the model generates an answer, marking the conclusion of the visual reasoning process. This entire process operates without manual intervention or the need for external tools, achieving both efficiency and adaptability in a single forward pass. To implement this mechanism, we introduce a semi-automatic expert-supervised data engine and curate a dataset of 334K visual reasoning samples, encompassing both natural and textual scenes. This dataset is annotated progressively by employing AI[78, 81] and human experts in a streamlined pipeline. Ultimately, we train the Griffon-R model using this curated data to achieve the unified mechanism.

To validate our design, we conduct comprehensive experiments with Griffon-R across a range of visual reasoning and multimodal benchmarks. The results demonstrate that empowered by our design mechanism, Griffon-R achieves advancing performance on complex visual reasoning tasks VSR[41] and CLEVR[25] and surpasses advanced LMMs on multimodal tasks including MMbench [46], ScienceQA [50], *etc.*, highlighting its enhanced general capabilities. Our key contributions are as follows:

- We propose a unified visual reasoning mechanism inspired by the human reasoning process, enabling LMMs to handle diverse compositional tasks by leveraging advanced capabilities through an "understand-think-answer" process in a single forward pass.
- We curate 334K multi-scene visual reasoning data by the introduced semi-automatic expert-supervised data engine and further present Griffon-R, a general LMM that is skilled in solving complicated compositional problems.
- We conduct extensive experiments on a wide range of visual reasoning and multimodal benchmarks. Griffon-R achieves advancing performance in compositional VSR and TallyQA, while further boosting performance on the multimodal benchmark including MMBench, ScienceQA, *etc.*.

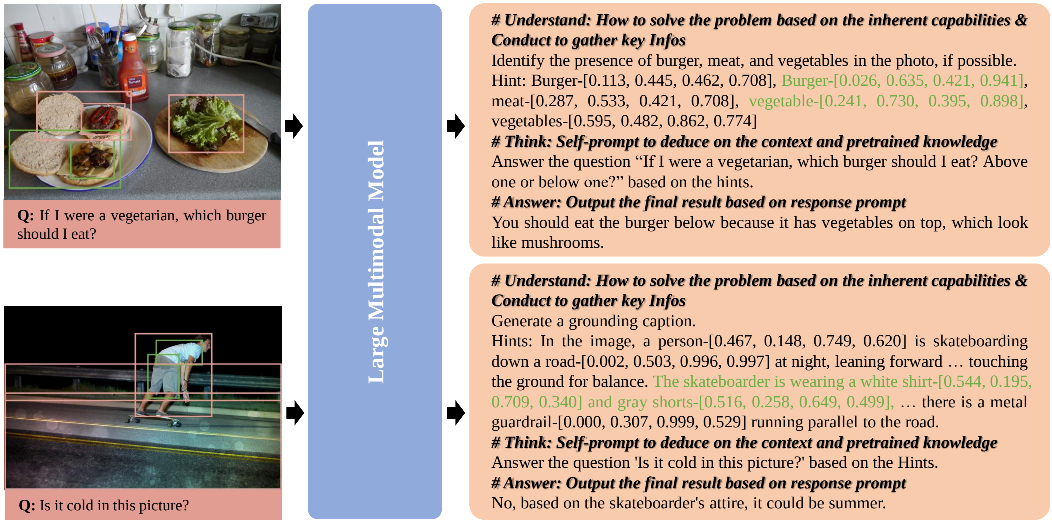
2 Related Works

2.1 Multimodal Chain of Thought

The CoT approach [82, 30, 94, 86, 4] is a series of prompting techniques that improve the ability of LLMs to solve complex reasoning tasks. With the rise of LMMs, these methods have gradually been incorporated into the multimodal domain to enhance model performance on complex reasoning tasks. ScienceQA [50] pioneeringly proposes Multimodal CoT to combine image captioning for reasoning, thus enabling models to handle complex question-answering tasks in science. Later works further decompose complex visual reasoning tasks into sequential steps, incorporating diverse prompts such as bounding boxes [8, 36, 70], textual descriptions [95], and scene graphs[58], allowing models to reduce intuitive errors by following a structured reasoning path. However, these multimodal CoT methods usually design specific paths for considered tasks and thereby limiting the model's ability to generalize across diverse question types and visual tasks. In this way, methods like Visual CoT[70] tend to provide rough prompts when questions are not included in these designed patterns. Also, some of them[70, 36] re-forward regions during inference to enhance region understanding. In contrast, our approach keeps the streamlined model structure and inference process with a single forward pass, which is more efficient. Meanwhile, our mechanism analyzes the question to automatically leverage intrinsic capabilities to understand the image precisely, facilitating a more flexible pattern for better generalization.

2.2 Toolkit-Based Visual Reasoning

Unlike multimodal CoT methods that convert various prompts into text for guidance, visual-tool-using approaches[32, 85, 24, 65, 83, 22, 19, 75] directly input information in different formats or modalities into the model, encouraging comprehensive understanding and reasoning. Methods like SoM [85] and Scaffolding [32] initially incorporate additional structured information within images, using these as anchors to prompt GPT-4V in visual reasoning. In contrast to these methods specifically designed for GPT-4V, Vcoder[24] takes a different approach by projecting depth and segmentation maps into the text embedding space, thereby improving reasoning accuracy by enhancing model comprehension. These methods typically rely on one or two fixed visual tools, limiting the richness of the information provided. Consequently, other approaches [65, 83, 22, 19, 75] have been developed to construct visual toolkits specifically for visual reasoning tasks in LMMs, allowing LLMs to generate executable program calls to specialist visual models based on the question. Building on this, CogCom [65] further integrates certain visual manipulations with the model's internal capabilities, enabling it to generate and execute an operation chain to progressively complete reasoning tasks. In comparison to these methods, our method leverages intrinsic capabilities and supports knowledge sharing across contexts, minimizing the risk of errors from isolated task execution. Additionally, with an end-to-end manner, our approach reduces the optimization difficulty and the latency.



Q: If I were a vegetarian, which burger should I eat?

# Understand: How to solve the problem based on the inherent capabilities & Conduct to gather key Infos
Identify the presence of burger, meat, and vegetables in the photo, if possible.
Hint: Burger-[0.113, 0.445, 0.462, 0.708], Burger-[0.026, 0.635, 0.421, 0.941], meat-[0.287, 0.533, 0.421, 0.708], vegetable-[0.241, 0.730, 0.395, 0.898], vegetables-[0.595, 0.482, 0.862, 0.774]
# Think: Self-prompt to deduce on the context and pretrained knowledge
Answer the question “If I were a vegetarian, which burger should I eat? Above one or below one?” based on the hints.
# Answer: Output the final result based on response prompt
You should eat the burger below because it has vegetables on top, which look like mushrooms.

Q: Is it cold in this picture?

# Understand: How to solve the problem based on the inherent capabilities & Conduct to gather key Infos
Generate a grounding caption.
Hints: In the image, a person-[0.467, 0.148, 0.749, 0.620] is skateboarding down a road-[0.002, 0.503, 0.996, 0.997] at night, leaning forward … touching the ground for balance. The skateboarder is wearing a white shirt-[0.544, 0.195, 0.709, 0.340] and gray shorts-[0.516, 0.258, 0.649, 0.499], … there is a metal guardrail-[0.000, 0.307, 0.999, 0.529] running parallel to the road.
# Think: Self-prompt to deduce on the context and pretrained knowledge
Answer the question 'Is it cold in this picture?' based on the Hints.
# Answer: Output the final result based on response prompt
No, based on the skateboarder's attire, it could be summer.

Figure 2: Detailed illustration of the unified visual reasoning mechanism with the “Understand-Think-Answer” process. The key information related to the answer is highlighted or visualized with the green color. We illustrate the details of the designed process in bold, which will not be generated or trained.

3 Methodology

In this section, we start with our designed novel unified visual reasoning mechanism, specifically the “Understand-Think-Answer” process which bridges the intrinsic visual foundational capabilities and VQA answering to allow the LMMs to reason accurately in a single pass in Sec. 3.2. Then, we detail the semi-automatic expert-supervised data engine on how it curates the 334K high-quality visual reasoning data aligned with the mechanism in Sec. 3.3. Finally, we present the Griffon-R, an visual reasoning improved LMM built on the proposed unified mechanism and data.

3.1 Preliminary

Current LMMs typically adopt an auto-regressive approach to generate response via next token prediction. Specifically, for an LMM M with parameters θ, given an input image I and user instruction Q, the model maximizes the sequence probability to output the response sequence X:

$$p(X) = \prod_{i} p(x_i | I, Q, X_{<i}) \tag{1}$$

where $X_{<i}$ represents the sequence before the current prediction x_i in the output sequence. Current LMMs employ a shortcut learning paradigm, training the model to directly generate the final answer X_{ans}. When simplifying the probability calculation process, the process can be denoted as:

$$X_{ans} = M_\theta(I, Q) \tag{2}$$

Though it effectively solves straightforward question-answering tasks, it struggles with more challenging compositional visual reasoning tasks, as illustrated in Fig. 1. Therefore, we propose the followed unified visual reasoning paradigm that advances the model to perform compositional reasoning leveraging intrinsic capabilities in a single forward pass through “understand-think-answer” process.

3.2 Unified Visual Reasoning Mechanism

Current LLMs[77, 78] trained on billions of data have acquired rich knowledge and experience, while LMMs through image-text alignment[34, 66] and instruction tuning[12, 44] further enhance their ability to perceive and interpret information by themselves. However, they still struggle to imitate how an educated person answers compositional questions by problem analyzing with past experience, relevant information gathering, and reaching a conclusion from thinking[2, 56]. Therefore, we propose the unified visual reasoning paradigm to imitate this approach to enable LMMs to sequentially and continuously perform the "understand-think-answer" process, accurately arriving at the answer in a single pass without introducing any toolkit.

Understand. Understanding is a key step that directly impacts the accuracy of the response. Unlike direct answering, the models first analyze the question and image to determine how to approach the problem and which intrinsic abilities to use to extract relevant information. Based on this analysis, the model employs the appropriate capabilities to gather key information, achieving a thorough understanding of the image in relation to the question. As shown in Fig. 2, we combine question analysis with intrinsic capability-based information retrieval planning and generate instructions for acquiring the required information. Then, they automatically guide the model to gather relevant visual cues. For the above sample in Fig. 2, the model identifies that solving the question requires knowing the location of the burger and distinguishing which items contain meat or vegetables, as indicated in the instruction by the terms "burger," "meat," and "vegetables". The "Indentify the presense" process involves applying these capabilities to gather the necessary information. This understanding process leverages common capabilities of LMMs, including caption, grounded caption, visual grounding, text recognition, *etc.*, instead of relying on any fixed capability like scene analysis[58] or REC[70]. Moreover, when no relevant information is found, the model outputs none instead of providing vague clues[70, 83], avoiding potential misleading.

Think & Answer. Instead of designing a specific reasoning path, we adopt a self-prompt approach. Previous studies show current models[61] can effectively respond to questions based on contextual cues. Additionally, modern LMMs[87, 91] excel at understanding coordinates-format object references in the context without needing additional forward passes[36, 70] for object perception. After achieving a deep understanding, we allow the model to generate the instruction to encourage the model to engage in self-thinking based on the visual cues, as indicated in Fig. 2. Ultimately, the model generates the final reasoning output according to the response template in the user's instruction. By now, the model performs question-customized deep understanding of the image in a single forward pass, engages in self-prompted thinking based on the generated cues, and outputs the final answer. This process can be summarized as:

$$p([X_U, X_T, X_{ans}]) = \prod_i p(x_i|I, Q, X_{<i})$$

$$X_{ans} = M_\theta(I, Q, X_U, X_T)$$

where X_U denotes the context generated during the understanding and X_T denotes the context generated during the thinking. The sequence inside the square brackets [] is generated with a left-to-right order of generation. Compared to shortcut learning-based methods, our unified visual reasoning mechanism integrates intrinsic capabilities to provide more accurate solutions to combinatorial visual reasoning tasks. In contrast to other CoT and toolkit-based approaches, it offers both flexibility and efficiency, which eliminates the need for multiple forward passes.

3.3 Expert-Supervised Data Engine

Beyond the design of the mechanism, training LMMs for accurate visual reasoning remains a significant challenge due to data limitations. To address this challenge and support the implementation of our designed visual reasoning mechanism, we introduce a semi-automatic expert-supervised annotation engine in this section, which is designed to generate high-quality visual reasoning data with visual cue annotations following our mechanism. As shown in Fig. 3, this semi-automated approach, supported by experts, enables us to generate high-quality data that enhances the model's visual reasoning capabilities through the unified reasoning mechanism.

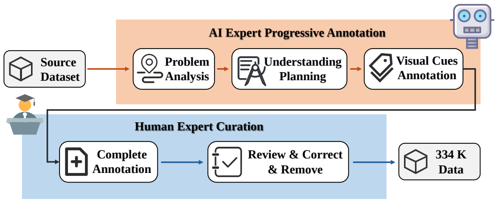


Figure 3: Illustration of the semi-automatic expert-supervised data generation engine.

Progressive Annotation with AI Expert. Due to the advanced question-answering capabilities and vast knowledge of current LMMs trained with billions of tokens, we leverage state-of-the-art LMMs to assist with annotation based on our visual reasoning mechanism. We select the Qwen2-VL-72B model[81] from the latest open-source models as our AI Expert, due to its strong task performance. For each sample, the AI Expert is responsible for designing the understanding process for solving problems and annotating tasks it excels in. First, we prompt the AI Expert to analyze the question and image to determine how to solve the problem and identify the necessary information. Then, we further instruct the expert to outlining the understanding process with specific tasks to gather key information based on the common intrinsic capabilities of LMMs. Finally, for tasks that the AI Expert specializes in, we directly use the model to generate the corresponding annotations, such as caption, text recognition, *etc*. The instructions used in this process are described in the Appendix A.

Curation with Human Expert. Although the AI Expert is skilled in analyzing and annotating tasks related to the question, its capabilities are limited in complex visual reasoning and tasks, such as visual grounding involving multiple objects. Previous works[9, 91] have also highlighted that human-involved high-quality annotations form the foundation for scaling data size in subsequent phases. Therefore, after the AI Expert's annotation, we first employ human experts to complete tasks that the AI Expert is not proficient at, primarily visual grounding in scenes with multiple objects or partial text. Then, human experts further review the annotations for quality, including evaluating whether the understanding process is sound, ensuring the accuracy of the task annotations, and eliminating overly simplistic questions.

Data Description. Leveraging our proposed semi-automatic expert-supervised data generation pipeline, we curate 334K image-question pairs from multi-task instruction-following data across multiple levels from public datasets. We mainly focus on the general scene data and text-rich scene data, which covers compositional reasoning problems. The data annotations are then completed through the above semi-automatic expert-supervised annotation process, which incorporates the AI expert progressive annotation stage and human expert curation stages. We provide more information about the source data in the Appendix A.

3.4 Griffon-R

Leveraging the proposed unified visual reasoning mechanism and the curated 334K data, we further develop the Griffon-R model, an LMM proficient at both compositional visual reasoning tasks and straightforward QA tasks. Benefiting from our design, Griffon-R maintains a streamlined architecture without the need for additional perception structures. Also, Griffon-R generates responses only through the next token prediction, rather than relying on module calling or multiple forward passes. We detail the construction process, including the architecture and training pipeline, which may also serve as a guideline to facilitate the implementation of the unified visual reasoning mechanism.

Structure. Griffon-R adopts the advanced single-branch high-resolution structure proposed in Griffon v2 [93]. Compared with other LMMs, this structure is proved to be better at fine-grained object localization, which is important for most of visual reasoning tasks. It consists of three core components: a high-resolution visual encoder, a vision-language connector, and an LLM. The high-resolution visual encoder processes image inputs up to 1K resolution without partitioning, with the connector projecting and compressing the tokens while retaining the performance.

Training Pipeline. We follow the common practice [36, 70] to train the Griffon-R model. Specifically, after the basic pertaining stage, we combine the curated visual reasoning data following our mechanism with VQA and instruction data to fine-tune the model with the whole model updated. Differently, we only use the cross-entropy loss to supervise the training without introducing task-specific losses[83]. We detail the training data and setting for each stage in the Appendix B.1.

4 Experiments

4.1 Implementation Details

We follow the Griffon v2 [93] to set the resolution to 1022, randomly initialize a down-sampling projector implemented as a 3×3 convolution (stride 2, padding 1).Then, we utilize CLIP-ViT-L / 14-336 [66] to initialize the visual encoder and further interpolate to the defined resolution. To ensure that our framework learns visual language reasoning capabilities from the scratch - without sacrificing generality - we selected Griffon-G-9B architectural paradigm and use Gemma9B [77] to initialize the LLM. For the training, we utilize the AdamW optimizer [48], setting the learning rate to 1e-3 for the first stage and 2e-5 for stage 2 and stage 3. We use the DeepSpeed zero2 [67] and a cosine learning

Table 1: Evaluation results on visual reasoning benchmarks. CLEVR focuses on generated structured scenes, while the others evaluate models in the natural scenes. We use the Spatial subset of V-Star here due to the reasoning step included, while the attribute subset primarily evaluates models' attribute perception abilities in high-resolution scenes.

Methods	VSR	CLEVR	GQA	V-Star$_{Spat.}$	TallyQA	
					Simple	Complex
Large Multimodal Models						
BLIP-2-7B [34]	50.9	-	44.7	53.9	-	-
InstructBLIP-7B [12]	52.1	-	48.3	47.4	74.3	48.7
LLaVA-1.5-7B [42]	64.2	43.7	62.0	53.9	75.9	63.8
LLaVA-1.5-13B [42]	70.4	55.8	63.3	55.3	76.9	65.4
Monkey-7B [37]	62.9	46.3	60.7	53.9	80.9	63.0
DeepSeek-VL-7B [49]	67.5	48.8	61.31	40.3	79.5	62.1
LLaVA-Next-7B [43]	63.8	51.9	64.2	63.2	80.3	66.5
Ferret v2-7B [88]	-	-	64.7	-	-	-
Large Multimodal Models with MCoTs or Toolkit-Based Enhancement						
VolCano-7B [36]	67.2	56.2	64.4	50.3	73.5	55.2
CogCom-17B [65]	-	-	**71.7**	-	84.0	70.1
SEAL-7B [83]	48.5	-	-	76.3	51.9	20.6
VPD-5B [22]	-	-	61.3	-	83.1	**70.9**
VisualCoT-7B [70]	61.4	55.5	63.1	50.3	82.4	60.2
VisualCoT-13B [70]	-	55.8	63.3	54.9	83.1	70.3
Large Multimodal Models Leveraging Intrinsic Capabilities						
Griffon-R-9B	**70.9**	**63.7**	65.1	**77.6**	**84.4**	70.4

rate strategy [47] with a warmup [17] ratio of 0.3. We train each stage for 1 epoch with the batch size of 256.All training was performed on eight NVIDIA H100 GPUs.

4.2 Evaluation Details

To comprehensively evaluate Griffon-R's capabilities, we conduct a fair comparison with advancing LMMs and visual reasoning methods on visual reasoning benchmarks across both structured scenes like CLEVR[25] and natural scenes, encompassing VSR[41], GQA[23], TallyQA[1], and V-Star$_{Spat.}$. These benchmarks mainly require models to comprehend the multi-level information based on the image to infer the final answer. We also evaluate Griffon-R on multimodal benchmarks to demonstrate its comprehensive capabilities. We include benchmarks that contain partial visual reasoning with common sense and knowledge, like MMBench[46], ScienceQA[50], SEED[33], and LLaVA Bench[44], and also TextVQA[74], which focuses on text-scene understanding. Given that strong visual reasoning skills help mitigate model hallucinations, we also assess performance on the POPE[35] benchmark. Also, we validate our design specifically in the ablation studies, including understanding quality indicated by grounding task, mechanism, and data.

4.3 Evaluation on Compositional Visual Reasoning

As shown in Tab. 1, Griffon-R achieves outstanding performance across these visual reasoning benchmarks. It reaches advancing levels in both generated scenes task CLEVR and natural scenes including VSR, V-Star$_{Spat.}$, and TallyQA$_{Simple}$, outperforming advanced visual reasoning models like Visual CoT and VPD. Specifically, Griffon-R achieves 63.7% accuracy on CLEVR and 70.9% accuracy on the VSR, surpassing the second-place models by a large margin. Such remarkable performance demonstrates the strong visual reasoning capabilities of Griffon-R across multiple tasks and scenarios. Notably, it also surpasses the latest advanced LMMs LLaVA-Next and Ferret v2. On the high-resolution, small-object compositional reasoning benchmark V-Star whose scenes are quite challenging for LMMs, Griffon-R outperforms multiple methods based on CoT and visual search techniques. These achievements across diverse task focus further highlight Griffon-R's robust visual reasoning capabilities and validate the effectiveness of our mechanism and data design.

Table 2: Evaluation results on multimodal benchmarks. Compared with visual reasoning benchmarks, these benchmarks focus more on visual understanding with common sense and knowledge, yet incorporating simple reasoning.

Methods	MMB	ScienceQA	TextVQA	SEED-Img	LLaVA-W	POPE
Large Multimodal Models						
InstructBLIP-7B [12]	36.0	-	50.1	58.8	60.9	72.1
LLaVA-1.5-13B [42]	67.7	71.6	61.3	68.2	72.5	85.9
QwenVL-Chat-7B [3]	60.6	68.2	61.5	65.4	-	84.7
Monkey-7B [37]	61.9	69.4	67.6	67.6	53.9	82.6
DeepSeek-VL-7B [49]	71.3	-	63.7	70.4	21.6	85.8
LLaVA-NeXT-7B [43]	69.0	73.2	64.9	70.2	72.3	86.4
Ferret v2-13B [88]	-	-	62.2	61.7	69.9	88.1
Large Multimodal Models with MCoTs or Toolkit-Based Enhancement						
VolCano-7B [36]	68.1	38.3	57.4	64.5	56.5	86.5
SEAL-7B [83]	33.1	-	-	41.7	59.1	82.4
CCoT-13B [58]	70.7	69.7	-	69.7	74.9	-
VPD-5B [22]	69.0	83.1	70.9	-	-	88.6
CogCom-17B [65]	-	84.0	-	-	-	87.8
VisualCoT-7B [70]	67.3	68.3	61.0	-	49.7	86.5
VisualCoT-13B [70]	67.4	73.6	62.3	-	57.7	83.3
Large Multimodal Models Leveraging Intrinsic Capabilities						
Griffon-R-9B	**79.0**	**87.0**	**72.4**	**73.8**	**76.2**	**89.3**

4.4 Evaluation on Multimodal Benchmarks

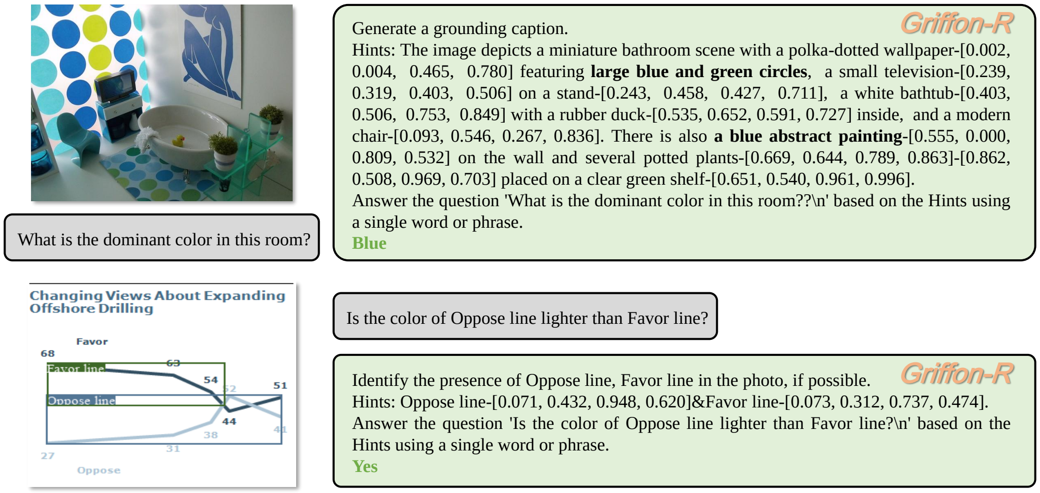
Empowered by the unified visual reasoning mechanism, Griffon-R leverages the model's inherent capabilities for visual reasoning tasks. Meanwhile, Griffon-R with streamlined structure and joint optimization with straightforward QA data can also handle these widely applied multimodal benchmarks. Therefore, we also evaluate Griffon-R on multimodal benchmarks and various VQA tasks to verify its comprehensive capabilities. As shown in Tab. 2, Griffon-R outperforms representative LMMs as well as advanced visual-reasoning-enhanced LMMs. Griffon-R achieves a score of 79.0 on the comprehensive MMBench and excels on ScienceQA, which focuses on scientific reasoning. Also, in real scenarios, Griffon-R is more proficient at challenging tasks like SEED and LLaVA-in-the-wild benchmark. In text-based scenes, Griffon-R further surpasses the program-driven VPD model, achieving 72.4% accuracy on TextVQA. These evaluations demonstrate that Griffon-R not only precisely handles complex compositional reasoning tasks but also performs well in general visual question answering.

4.5 Ablation Studies

Discussion on Understanding Quality. As we have demonstrated in Sec. 3.2, understanding plays an important role in the whole mechanism for accurate visual reasoning. To quantitatively indicate the understanding quality, we choose the performance of Referring Expression Comprehension (REC) [89, 59] as the metric, which incorporates both visual localization and attribute perception that are commonly used intrinsic capabilities during understanding for lots of reasoning tasks. As shown in Tab. 4, our model outperforms the SOTA method Ferret v2 in this task and also visual reasoning methods CogCom and Visual CoT. The result verifies that our model can achieve accurate understanding with intrinsic capabilities and it further facilitates the precise visual reasoning of Griffon-R.

Figure 4: Ablation on understanding quality. With REC task covering object localization and attribute perception, we choose it to evaluate the quality of understanding in the mechanism.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	test-A	test-B	val	test-A	test-B	val-u	test-u
Expert Models								
UNINEXT [84]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
MDETR [27]	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9
G-DINO-L [45]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
Large Multimodal Models								
VistaLLM-7B [64]	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4
Ferret v2-7B [88]	92.8	94.7	88.7	87.4	92.8	79.3	89.4	89.3
CogCom-17B [65]	92.3	94.6	89.2	**88.2**	**92.8**	82.8	89.3	**90.5**
Visual CoT-7B [70]	91.8	94.3	87.5	87.5	92.1	81.2	88.4	88.4
Griffon-R-9B	**93.2**	**95.1**	**90.2**	**88.4**	92.3	**82.8**	**89.8**	89.7



What is the dominant color in this room?

Generate a grounding caption.
Hints: The image depicts a miniature bathroom scene with a polka-dotted wallpaper-[0.002, 0.004, 0.465, 0.780] featuring **large blue and green circles**, a small television-[0.239, 0.319, 0.403, 0.506] on a stand-[0.243, 0.458, 0.427, 0.711], a white bathtub-[0.403, 0.506, 0.753, 0.849] with a rubber duck-[0.535, 0.652, 0.591, 0.727] inside, and a modern chair-[0.093, 0.546, 0.267, 0.836]. There is also **a blue abstract painting**-[0.555, 0.000, 0.809, 0.532] on the wall and several potted plants-[0.669, 0.644, 0.789, 0.863]-[0.862, 0.508, 0.969, 0.703] placed on a clear green shelf-[0.651, 0.540, 0.961, 0.996].
Answer the question 'What is the dominant color in this room??\n' based on the Hints using a single word or phrase.
Blue

Is the color of Oppose line lighter than Favor line?

Identify the presence of Oppose line, Favor line in the photo, if possible.
Hints: Oppose line-[0.071, 0.432, 0.948, 0.620]&Favor line-[0.073, 0.312, 0.737, 0.474].
Answer the question 'Is the color of Oppose line lighter than Favor line?\n' based on the Hints using a single word or phrase.
Yes

Figure 5: Visualization of Griffon-R's reasoning results. Correct answers are highlighted in bold green, and the relevant information within the long text leading to the answer is bolded.

Table 3: Ablation study on the unified visual reasoning mechanism and toolkit-based mechanism.

Method	V-Star$_{Spat.}$	Time/Sample
Unified(ours)	77.6	0.336s
Toolkit-Based	76.3	4.586s

Table 4: Ablation study on the training and annotated data. UVRM denotes our mechanism.

UVRM	334k Ann.	V-Star$_{Spat.}$	POPE
		75.0	89.1
√		75.0	89.4 (+0.3)
√	√	77.6 (+2.6)	89.3 (+0.2)

Discussion on the unified visual reasoning mechanism. To validate the effectiveness of our unified visual reasoning mechanism, we compare its accuracy and inference time to the toolkit-based SEAL-7B method.We evaluate with the V-Star$_{Spat.}$ benchmarks and static the average inference time per-sample. As shown in Tab. 3, in the compositional reasoning task of V-Star, the designed mechanism outperforms the toolkit-based pipeline by 1.3 points and is 13x faster in average inference time. These results justify that our mechanism not only boost visual reasoning capabilities without relying on visual expert modules but also delivers high efficinet inference.

Discussion on training with the curated data annotations. In this section, we validate the necessity of training with curated annotations for our mechanism. The first line (baseline) and the second line results are trained on the same data with Griffon-R using the raw annotations. The second line additionally follows our mechanism but requires the model to answer step-by-step. We evaluate these methods on the V-Star$_{Spat.}$ complex visual reasoning benchmark and the POPE object existence hallucination benchmark. As shown in Tab. 4, simply providing visual cues (using UVRM) improves performance on tasks like POPE, which do not require compositional reasoning, by compensating for insufficient understanding. However, to achieve accurate visual reasoning, the model requires training on carefully curated datasets that enable it to understand the image, identify relevant cues, and reason based on different-pattern cues comprehension. This comparison emphasizes the effectiveness of our data-supported and unified visual reasoning mechanism design.

4.6 Visualization Results

We display the qualitative performance of Griffon-R through Fig. 5. The results highlight Griffon-R effectively handles compositional visual reasoning across various scenes by thoroughly understanding the problem and visual cues, then performing context-based reasoning to generate the final answers.

5 Conclusion

In this paper, we present the unified visual reasoning mechanism that empowers LMMs to tackle compositional reasoning tasks end-to-end, without external expert models or toolkits. This unified mechanism introduces a human-like understand-think-answer process to reason based on individual questions flexibly instead of utilizing fixed paths and completes all steps in a single pass forward without multiple inferences. Moreover, we design a semi-automatic expert-supervised data generation engine to produce high-quality visual reasoning data corresponding to the design mechanism. We collect public data related to visual reasoning and re-annotate them with our designed data generation engine, and curate 334K visual instruction samples. Based on the curated data and designed mechanism, we present Griffon-R. Griffon-R achieves advancing results on visual reasoning benchmarks, including VSR and CLEVR, and also demonstrates comprehensive multimodal capabilities. We hope our attempts at a unified visual reasoning mechanism will facilitate the deep exploration in this field to achieve more general LMMs. In the Appendix, we provide a detailed discussion of our method's limitations and its broader impact. In future work, we plan to explore additional reasoning paradigms and more diverse data to extend the applicability of our approach.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, 2019.

[2] John R Anderson. *Cognitive psychology and its implications*. Macmillan, 2005.

[3] Jinze Bai and et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv:2308.12966*, 2023.

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17682–17690, 2024.

[5] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[6] Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[7] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024.

[8] Keqin Chen and et al. Shikra:unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023.

[9] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[11] 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. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[13] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[15] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[16] Deepanway Ghosal, Yew Ken Chia, Navonil Majumder, and Soujanya Poria. Flacuna: Unleashing the problem solving power of vicuna using flan fine-tuning, 2023.

[17] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[18] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, page 398–414, 2019.

[19] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023.

[20] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023.

[21] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[22] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024.

[23] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002, 2024.

[25] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[26] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.

[27] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021.

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

[29] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[30] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, page 32–73, 2017.

[32] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

[33] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[35] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[36] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024.

[37] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024.

[38] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`, 2023.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[40] Dongyang Liu, Renrui Zhang, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, Kaipeng Zhang, et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. In *Forty-first International Conference on Machine Learning*.

[41] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

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

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

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

[45] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[46] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[47] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[49] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[50] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[51] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*, 2023.

[52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[53] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics.

[54] Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: a dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*, 2020.

[55] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[56] Mary B McVee, Kailonnie Dunsmore, and James R Gavelek. Schema theory revisited. *Review of educational research*, 75(4):531–566, 2005.

[57] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[58] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain of thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[59] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016.

[60] OpenAI. Gpt-4 technical report, 2023.

[61] OpenAI. Chatgpt: A large language model for natural language processing, 2024. Accessed via OpenAI platform. URL: `https://chat.openai.com/`.

[62] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13018–13028, 2021.

[63] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, page 74–93, 2017.

[64] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14076–14088, 2024.

[65] Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, et al. Cogcom: Train large vision-language models diving into details through chain of manipulations. *arXiv preprint arXiv:2402.04236*, 2024.

[66] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[67] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

[68] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[69] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[70] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024.

[71] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438, 2019.

[72] Yin Shukang, Fu Chaoyou, Zhao Sirui, Xu Tong, Wang Hao, Sui Dianbo, Shen Yunhang, Li Ke, Sun Xing, and Chen Enhong. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[73] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[74] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[75] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.

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

[77] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[78] Qwen Team. Qwen2.5: A party of foundation models, 2024.

[79] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

[80] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *The IEEE International Conference on Computer Vision (ICCV)*, 2023.

[81] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

[83] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024.

[84] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023.

[85] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

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

[87] Haoxuan You and et al. Ferret: Refer and ground anything anywhere at any granularity. In *ICLR*, 2024.

[88] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. Ferret-ui: Grounded mobile ui understanding with multimodal llms. *arXiv preprint arXiv:2404.05719*, 2024.

[89] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.

[90] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

[91] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28202–28211, 2024.

[92] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

[93] Yufei Zhan, Yousong Zhu, Zhiyang Chen, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon: Spelling out all object locations at any granularity with large language models. In *ECCV*, 2024.

[94] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

[95] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.

Appendix

Contents

A Details of Expert-Supervised Data Engine

In this section, we illustrate the details of data curation process with our proposed semi-automatic expert supervised data engine of Sec. 3.3. We start with the raw data collection, and then the instructions for the progressive annotation with AI expert.

A.1 Raw Data Curation

As described in Sec. 3.3, we first collect the widely used diverse types of data that related to visual reasoning following the previous practice [83, 36, 70] and curate a total of 334K visual-reasoning-oriented data using the proposed Expert-Supervised Data Generation Engine. These data are composed of two main parts: VQA-based data and caption data tailored to visual reasoning tasks.

Table 5: Annotation sources of the 334K visual-reasoning data.

Type	Num.	Source
VQA	207K	GQA[23], VAW [62], VizWiz [5], ChartQA [53], DUE_Benchmark [6], TextVQA [74]
Instruction	76K	LLaVA [44], ALLaVA [7], LVIS-Instruct4V [79]
Caption	51K	ShareGPT-4V [9]

VQA-Based Data. The VQA-based data source from the general VQA datasets, instruction-following data, and text-oriented VQA datasets. We extract visual reasoning data based on two criteria: (1) Yes/No Answers: We select questions with "yes" or "no" answers, as these often involve tasks like comparisons or attribute judgments (e.g., color differentiation). (2) Reasoning Keywords: Using the relationship and attribute keywords appearing in GQA [23], VAW[62], and DUE_Benchmark [6], which we extract with the raw keywords in the annotations or with spaCy [21] (*eg.*, "left", "right", "over", "red", and "plastic"), we identify Q&A pairs that included these keywords. As summarized in Tab. 5, this process yielded 207K data from VQA annotations and 76K entries from instruction-based dialogues. For VAW data, we specifically utilize images aligned with GQA to generate more diverse visual reasoning questions. After acquiring these visual-reasoning-oriented VQAs, we follow the proposed data generation method to annotate these data.

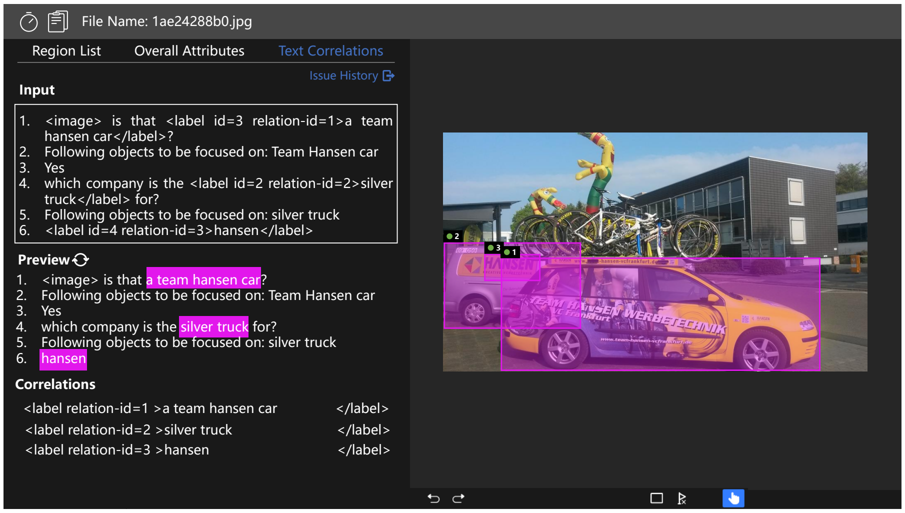
Table 6: Instruction templates for the AI expert.

Type	Template
I.1	For the question:'{question}', identify and focus on the specific physical objects or entities relevant to answering it. Directly output the names of these objects or short descriptive phrases with key attributes. If the object is mentioned in the question, use the exact word from the question. If the object is not explicitly mentioned, describe it concisely using your own terms.
I.2	For the question:'{question}', based on the context, identify the task used to gather the key entities information like [*TASKs*], if the question is too abstract, respond with "Global Understanding".

Visual-Reasoning-Oriented Caption Data. Caption data provide global context and are essential for the in-depth understanding of images. To make them suitable for visual reasoning, we apply the proposed data generation method, focusing first on identifying the key objects in the image that are critical for reasoning tasks. The annotation process involved checking whether these key objects are included, fixing them if missing, and manually labeling their bounding boxes. We utilized the detailed descriptions generated by ShareGPT-4V [9]. After the above processing, we obtain the final expert-supervised visual-reasoning-oriented data, as listed in Tab. 5.

A.2 AI Expert Annotation Details

At the beginning of data generation process, we first utilize the AI expert Qwen2-VL-72B [81] to identify how to solve this problem with key objects or information listed (I.1) and plan how to gather these key information with intrinsic capabilities (I.2). We list the instructions to prompt the expert to finish these two tasks in Tab. 6 respectively. Also, as we have mentioned in Sec. 3.3, for the task annotation process that the AI expert is skilled at, we use the AI Expert to finish the annotation first. We mainly apply this strategy in tasks that require a global understanding, including caption and grounded caption. We directly used the instruction examples in the paper[81]. For generating this intermediate information at this stage, we perform batch inference using 8 NVIDIA A800 GPUs.



Input

1. <image> is that <label id=3 relation-id=1>a team hansen car</label>?
2. Following objects to be focused on: Team Hansen car
3. Yes
4. which company is the <label id=2 relation-id=2>silver truck</label> for?
5. Following objects to be focused on: silver truck
6. <label id=4 relation-id=3>hansen</label>

Preview

1. <image> is that a team hansen car?
2. Following objects to be focused on: Team Hansen car
3. Yes
4. which company is the silver truck for?
5. Following objects to be focused on: silver truck
6. hansen

Correlations

<label relation-id=1 >a team hansen car </label>
<label relation-id=2 >silver truck </label>
<label relation-id=3 >hansen </label>

Figure 6: An example UI screenshot showcasing the annotation process for a question-answer pair requiring grounding. After reviewing and confirming the understanding process, the human verifying expert highlights the text based on the knowledge generated by the AI expert and annotates the corresponding bounding box by drawing on the image for each object.

A.3 Visualization of Human Expert Annotation

As we have introduced in Sec. 3.3, initial high-quality annotation is important. Therefore, we hire ten human annotation experts to help us. We provide a screenshot example of human expert annotating platform in Fig. 6.

A.4 Discussion on Data Scaling

Several works [9, 91] have highlighted that it's more effective to scale up the annotated data based on the initial high-quality data. Therefore, we follow this insight to employ the human expert for annotation as existing advanced LMMs still fall short in tasks like multi-object visual grounding. For the visual reasoning data scaling up, we provide a brief discussion here. From the perspective of data amount, it's implementable to specifically train an expert model[9] or our model using the curated data to further annotate a large amount of data. Then, a large expert-level model like ChatGPT[61] or human experts can be used for checking. While for the scene scaling, when considering more scenes, it's possible to include more capabilities into our set and curated related data to support scenes like math, accounting, *etc*. In this paper, we mainly focus on general natural scenes leveraging intrinsic capabilities like REG, caption, and visual grounding, and curate the 334K visual reasoning data. We will follow the discussion to scale up the data and generalize our mechanism to more scenes.

B Training Details

B.1 Training Data

We list the training data used for the three training stages in Tab. 7, with the data volume and dataset type specified. As we have introduced in Sec. 3.4, we employ the direct supervised fine-tuning (SFT) training strategy to build the mechanism within the Griffon-R. Therefore, we mix the visual reasoning data with general SFT data. While for the overlapping data with the visual reasoning data, we directly remove them.

B.2 Trainable Parameter Setting

In the first stage, we freeze the visual encoder and the LLM and leave the projector trainable. Then, we pretrain the whole model in stage II and further finetune the whole model in stage III.

Table 7: Training data used in each stage. The Lang. represents the language-only instructions, the Inst. represents the vision-language instructions, the Gen. represents the general, the Text. represents text-oriented, and the Perc. represents the perception. Perceptions data include REC, REG, visual grounding and object detection, while VR stands for visual reasoning.

Stage	Vol.	Type	Training Data
I	1.2M	Caption	ShareGPT-4V [9]
II	3.0M	REC	RefCOCO/+[89], RefCOCOg[59], GRefCOCO[20]
		REG	RefCOCO/+[89], RefCOCOg[59], GRefCOCO[20], Flickr30K Entities[63], Visual Genome[31], Osprey[91]
		DET	Objects365[71], MSCOCO[10], V3Det[80], Visual Genome[31]
III	3.9M	Lang.	UltraChat[13], Flan-mini[16], OpenOrca[38], MetaMathQA[90], ShareGPT, MathInstruct[92], WizardCoder[51]
		Inst.	LLaVA[44], ALLaVA[7], LVIS-Instruct4V[79]
		Caption	ShareGPT4V[9], TextCaps[73]
		Gen. VQA	VQA v2[18], GQA[23], OK-VQA[52], A-OKVQA[69], SQA[50], VizWiz[5]
		Text. VQA	TextVQA[74], OCR-VQA[57], AI2D[28], Synthdog[29], DVQA[26], ChartQA[53], DocVQA[54], InfoVQA[55], DeepForm[6], KLC[6], WTQ[6], TabFact[6]
		Perc.	RefCOCO/+[89], RefCOCOg[59], GRefCOCO[20], Flickr30K Entities[63], Visual Genome[31], Osprey[91], Objects365[71], MSCOCO[10], V3Det[80]
		VR	**Curated Visual Reasoning Data**

C Limitations

In complex scenarios, when the question-related objects are associated, the model's output sequence will grow, which can lead to an increase in the model's response time.

Our data pipeline incorporates Qwen2-VL-72B. As such, it inherits limitations common to MLLMs, including potential inaccuracies and the propagation of misinformation, although we have taken steps to reduce these through human verification. Concerning data usage, we explicitly state that Qwen's terms of use must be followed. The data may be used freely for research purposes; however, its use for other applications is restricted or necessitates a further request for permission.

D Broader Impact

This paper presents work whose goal is to advance the field of MLLMs. There are many potential societal consequences of our work following the MLLMs, none of which we feel must be specifically highlighted here.