MMCR: Benchmarking Cross-Source Reasoning in Scientific Papers

Yang Tian[1*] Zheng Lu[1,2*] Mingqi Gao[1,3*] Zheng Liu[4] Bo Zhao[1†]
[1] School of AI, Shanghai Jiao Tong University [2] SCUT [3] SEU [4] BAAI
{yangtian6781, bo.zhao}@sjtu.edu.cn
https://github.com/yangtian6781/MMCR

Abstract

Fully comprehending scientific papers by machines reflects a high level of Artificial General Intelligence, requiring the ability to reason across fragmented and heterogeneous sources of information, presenting a complex and practically significant challenge. While Vision-Language Models (VLMs) have made remarkable strides in various tasks, particularly those involving reasoning with evidence source from single image or text page, their ability to use cross-source information for reasoning remains an open problem. This work presents MMCR, a high-difficulty benchmark designed to evaluate VLMs' capacity for reasoning with cross-source information from scientific papers. The benchmark comprises 276 high-quality questions, meticulously annotated by humans across 7 subjects and 10 task types. Experiments with 18 VLMs demonstrate that cross-source reasoning presents a substantial challenge for existing models. Notably, even the top-performing model, GPT-4o, achieved only 48.55% overall accuracy, with only 20% accuracy in multi-table comprehension tasks, while the second-best model, Qwen2.5-VL-72B, reached 39.86% overall accuracy. Furthermore, we investigated the impact of the Chain-of-Thought (CoT) technique on cross-source reasoning and observed a detrimental effect on small models, whereas larger models demonstrated substantially enhanced performance. These results highlight the pressing need to develop VLMs capable of effectively utilizing cross-source information for reasoning.

1. Introduction

Scientific papers encapsulate the advanced human intelligence, and it requires human experts years of study and practical experience to fully comprehend the content of these papers. Hence, comprehending scientific papers with machines is a long-standing but largely under-explored challenge. Scientific papers mainly contain elements that challenge a model's understanding and reasoning ability, such as mathematical formulas, pseudocode and figures with specialized knowledge and cross-source clues.

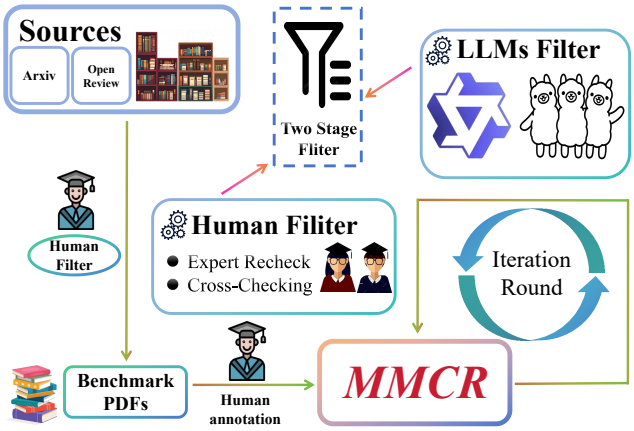

Figure 1. The construction pipeline of MMCR.

Recently, a variety of VLMs have emerged, including proprietary ones such as GPT-4o [12], Gemini-2.0 [28], and Claude-3 [1], as well as open-source ones such as InternVL2.5 [4], Qwen2.5-VL [3], mPLUG-DocOwl2.0 [10], LLaVA-OneVision [15], Molmo [6], CogVLM2 [9], and TextMonkey [20]. These models have demonstrated remarkable performance on general vision tasks, and many studies have explored their potential in document understanding, such as ChartQA [22], DocVQA [23], InfoVQA [24], DUDE [30], MP-DocVQA [29], etc. However, the evaluation of VLMs' capabilities in scientific paper understanding is still relatively limited.

Evaluating models for understanding scientific papers holds significant practical importance. Previous research in this domain has presented three primary limitations: 1) **Dependence on Model-Generated Annotations:** The high degree of specialization in scientific papers makes manual annotation both time-consuming and costly. For instance, ArxivQA [16] uses GPT-4v to generate Question-Answer (QA) pairs, while SCI-CQA [26] employs GPT-4o for the

*Co-first authors
†Corresponding author

Benchmarks	Avg. Pages	Cross-Page	Text	Chart	Table	Image	Pseudocode	Formula	CR	Annotatiom
DocVQA [23]	1.0	✗	✓	✓	✓	✓	✗	✗	✗	✓
ChartQA [22]	1.0	✗	✗	✓	✗	✗	✗	✗	✗	✔
InfoVQA [22]	1.0	✗	✗	✓	✓	✓	✗	✗	✗	✓
Charxiv [32]	1.0	✗	✗	✓	✗	✗	✗	✗	✗	✔
ArxivQA [16]	1.0	✗	✓	✓	✓	✓	✗	✗	✗	✗
MMSCI [18]	1.0	✗	✓	✓	✓	✓	✗	✗	✗	✗
DUDE [30]	5.7	✓	✓	✓	✓	✓	✗	✗	✗	✓
MP-DocVQA [29]	8.3	✗	✓	✓	✓	✓	✗	✗	✗	✔
SlideVQA [27]	20.0	✓	✓	✓	✓	✓	✗	✗	✗	✓
MMLongBench-Doc [21]	47.5	✓	✓	✓	✓	✓	✗	✗	✗	✓
MMCR	19.0	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Comparison between our benchmark and previous related datasets. **Avg. Pages:** Average Pages. **Cross-Page:** Cross-Page Questions. **Text/Chart/Table/Image/Pseudocode/Formula:** Pure Text/Chart/Table/Image/Pseudocode/Formula Questions. **CR:** Cross-Source Reasoning Questions. **Annotation:** ✓: Human Annotation, ✗: Automatic Annotation, ✔: Semi-automatic Annotation.

similar task. Although these generated QA pairs are manually reviewed, they still reflect biases inherent in the generative models, which could skew the evaluation of other models. 2) **Weak Relevance of Annotations to the Article's Content:** The specialized nature of scientific papers presents a significant challenge in dataset construction. Charxiv [32] collects papers from Arxiv but focuses solely on charts for real-world chart understanding, neglecting the specialized interpretation of these charts within the context of the articles. Similarly, MMSCI [18], which utilizes high-quality papers from *Nature Communications*, concentrates on constructing datasets around figures, without addressing the content-related questions and answers. These efforts leverage the complexity of images in scientific papers but fail to tap into the specialized knowledge embedded within the scientific paper. 3) **Lack of Cross-Source Reasoning:** In scientific papers, cross-source interactions reveal insights that are absent from a single source. The ability to integrate and reason across diverse sources is crucial, as it demonstrates a deeper and more comprehensive understanding of the paper. However, the majority of existing research [16, 26, 32] primarily focuses on understanding individual source within scientific papers. In summary, the absence of high-quality, professional benchmarks for scientific paper understanding has motivated the exploration of our work.

In this paper, we introduce MMCR, a benchmark specifically designed to evaluate the **M**ulti-**M**odality **C**ross-source **R**easoning capability of VLMs in scientific papers. Figure 1 outlines the annotation pipeline of our benchmark. The benchmark has been meticulously curated by expert annotators, with each paper averaging 19 pages across 7 different subjects. We have crafted 10 distinct categories of questions, totaling 276 high-quality, domain-specific questions. The benchmark has undergone filtering assisted by large language models (LLMs) and rigorous manually cross-checking, which guarantees that each question is firmly tied to its cross-source clues and cannot be answered through alternative sources, thereby ensuring the robustness of the benchmark.

We conducted extensive experiments on MMCR to evaluate the cross-source reasoning capability of state-of-the-art VLMs in scientific papers. A total of 18 VLMs were evaluated, comprising 15 open-source models and 3 proprietary models. For each scientific paper, we re-render it into high-quality JPEG images and then are processed by the VLMs in an end-to-end approach. In other words, we expect VLMs to be able to comprehend scientific papers purely through visual input. Our results underscore the significant challenges these models face in performing cross-source reasoning of contents in scientific papers. Among the models tested, the best performer, GPT-4o, achieved an overall accuracy of only 48.55%, while the second-best model, Qwen2.5-VL-72B, reached 39.86%. Furthermore, we conducted a manual analysis of GPT-4o's incorrect answers to identify the key bottlenecks in scientific paper understanding for current VLMs. The most frequent source of errors was perceptual error, which accounted for 27.5% of the total, revealing the perceptual defects of current models when handling multi-page scientific papers with multiple information sources. What's more, we explored the efficacy of CoT in cross-source reasoning and found that that its performance-enhancing benefits are predominantly evident in large-scale models, which reveals the limitations of CoT in enhancing cross-source reasoning for smaller models. In summary, these findings highlight that cross-source reasoning in scientific papers remains a complex and ongoing challenge.

2. Related Work

2.1. Scientific Paper Understanding

Despite the increasing significance of automating the understanding of scientific papers, the datasets of scientific papers

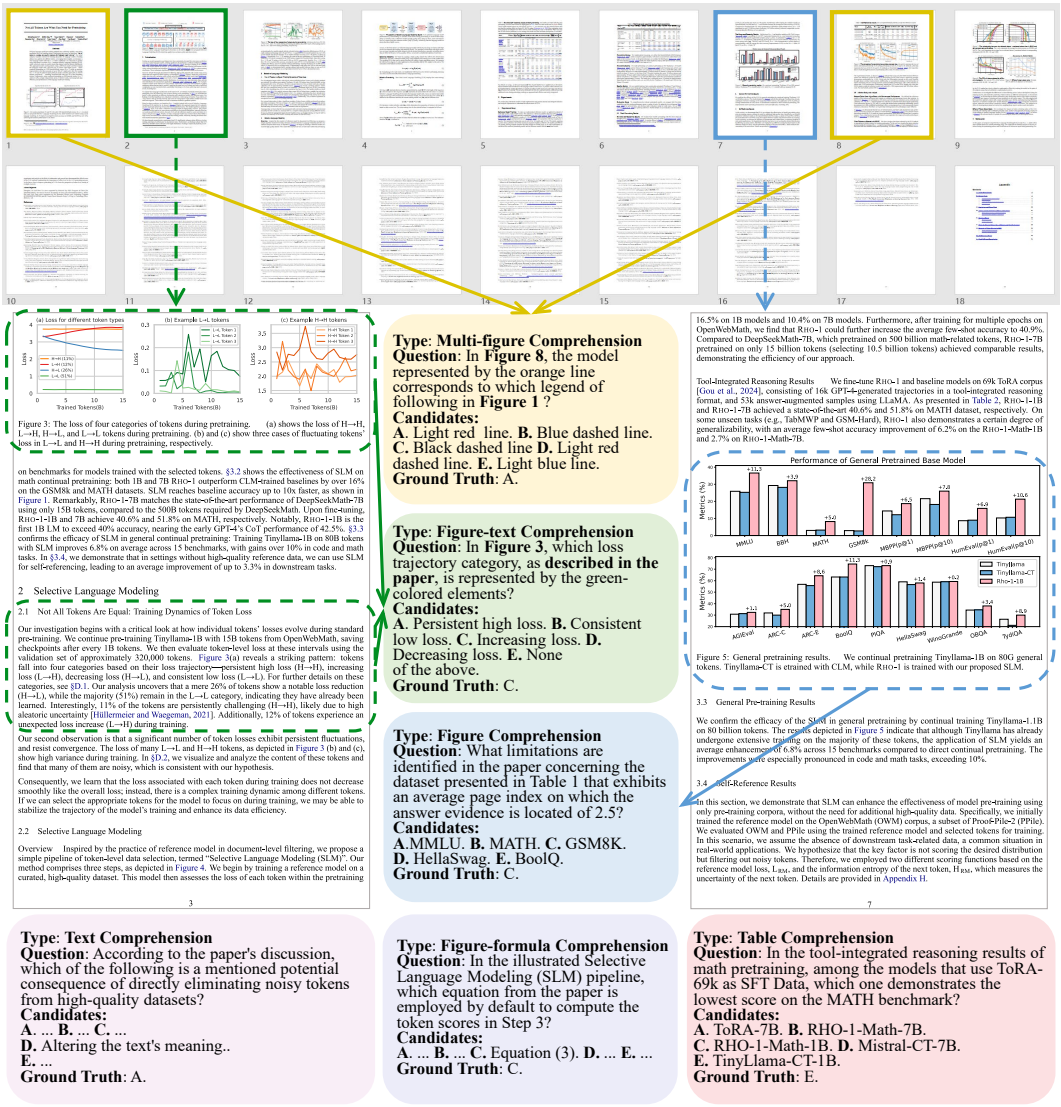

Figure 2. Examples of MMCR. We meticulously curated 10 distinct question types, each accompanied by five answer options, to assess the ability of visual-language models to perform reasoning across multiple information sources.

remain underdeveloped and insufficiently explored. In scientific paper understanding, data generation processes are mainly semi-automated. For example, Charxiv [32] collects charts from Arxiv papers and employs GPT-4v to generate QA pairs, which are then manually reviewed to ensure data quality. ArxivQA [16] also utilizes GPT-4v for generating QA pairs specific to figures in Arxiv papers. MMSCI [18], despite relying on manual annotation and selecting high-quality figures from `Nature Communications`, does not generate content-related QA pairs. Instead, all tasks in the MMSCI datasets are based solely on captions, with minimal connection to the actual content of the papers. Moreover, none of these datasets address cross-page understanding of the entire paper, and the information sources required for answering the questions are singular. More detailed descriptions and comparisons are presented in Table 1.

2.2. Benchmarks for Document Understanding

Compared to scientific paper understanding, document understanding is more advanced in its development. While the research of document understanding has shifted focus from single-page [22–24] to multi-page understanding [13, 21, 29, 30], existing benchmarks for multi-page document understanding remain limited in critical aspects. MP-DocVQA [29], an extension of the DocVQA [23], fails to adequately assess cross-page reasoning by omitting inter-page dependency questions. Although the DUDE [30] dataset demonstrates rigorous annotation protocols with substantial manual effort, its average sample length of 5.7 pages per document provides insufficient scope for eval-

uating models on authentic multi-page challenges. SlideVQA [27] features cross-page questioning through 20-page slides, but the inherent low information density of slides makes its task less challenging. MMLongBench-Doc [21] represents progress with comprehensive 47.5-page document samples across seven document types. However, it neglects to evaluate cross-source information synthesis, which is a crucial capability for real-world applications requiring integration of disparate information in a document. These limitations collectively highlight the need for benchmarks that simultaneously address extended document length, cross-page relationships, heterogeneous source integration, and domain-specific information density.

2.3. Models for Document Understanding

Document understanding models can be broadly categorized into two main approaches. The first relies on optical character recognition (OCR) tools, exemplified by models like LLaVA-Read [38], which use dual-stream vision encoder in conjunction with OCR tools to extract text from images, making it highly effective for text-rich image understanding tasks. The second approach is OCR-free, where models [3–5, 20, 31] process documents end-to-end without the dependency of OCR tools. Currently, there is a clear trend in document understanding models evolving from OCR-dependent architectures to OCR-free approaches.

3. MMCR Benchmark

MMCR is a comprehensive and challenging benchmark, which is designed to evaluate the cross-source reasoning capability of VLMs using multiple intertwined information sources in scientific papers. In this section, we detail the processes of data collection, question and answer annotation, quality control, and evaluation method. Finally, we present the statistical data of MMCR.

3.1. Data Collection

We collected scientific papers from both Arxiv and high-quality publications on the OpenReview platform. Due to the highly specialized nature of scientific papers, annotating them presents significant challenges for non-experts. To address this, we engaged expert annotators who are not only proficient in English reading and writing but also have extensive experience in their respective research fields. The collected papers were distributed in batches to the annotators, who were given the flexibility to select papers within their areas of expertise for annotation.

3.2. Question and Answer Annotation

Before beginning the formal annotation process, we first pre-define the annotation categories and conduct annotation exercises. These categories are based on the sources of information needed to answer the questions. Specifically, the annotation categories are as follows:

Statistic	Number
Scientific Paper	31
- Avg./Max. Pages	19 / 35
- Avg./Max. Number of Options	5 / 5
- Average Size (px)	1573×1210
- Maximum Size (px)	1584×1224
Subject	7
- 2D Object Recognition	4 (12.90%)
- Efficient AI	2 (6.45%)
- Multimodal Learning	15 (48.39%)
- Datasets and Evaluation	4 (12.90%)
- 3D from Multi-View and Sensors	1 (3.23%)
- Large Language Model	2 (6.45%)
- Robotics	3 (9.68%)
Total Questions	276
- Figure Comprehension	43 (15.58%)
- Multi-Figure Comprehension	20 (7.25%)
- Figure-Table Comprehension	26 (9.42%)
- Figure-Text Comprehension	53 (19.20%)
- Figure-Formula Comprehension	13 (4.71%)
- Table Comprehension	62 (22.46%)
- Multi-Table Comprehension	10 (3.62%)
- Text Comprehension	35 (12.68%)
- Formula Comprehension	9 (3.26%)
- Pseudocode Comprehension	5 (1.81%)

Table 2. Benchmark Statistics.

1) **Figure Comprehension:** Figures in scientific papers serve as visual representations of models, theories, data, and other key elements. Questions based on these figures assess the model's ability to interpret and understand the visual information they convey. What's more, the annotated questions do not specify which figure is being referenced, thereby evaluating the model's ability to comprehend figures in the context of a multi-page scientific paper.

2) **Multi-Figure Comprehension:** In scientific papers, figures are often not standalone elements but are interconnected with other figures. By reasoning across multiple figures, it avoids being limited to the interpretation of individual figures, leading to more comprehensive conclusions. These questions evaluate the model's ability to utilize information across figures to answer effectively. Through meticulous verification, we ensure that all the necessary evidence for reasoning is confined to the relevant set of figures, with no information sourced from elsewhere.

3) **Figure-Table Comprehension:** The figure provides rich graphical information, while the table offers detailed structured data. These questions require VLMs to grasp the underlying relationships between graphical and structured information and perform reasoning across both sources.

Model	Date	#Param	Overall	FIC	MF	FTA	FTE	FF	TAC	MT	TEC	FOC	PC
Open-Source Multimodal Large Language Models													
4-9B Models													
Phi-3.5-Vision [2]	2024-09	4B	4.71	6.98	5.00	0.00	5.66	0.00	3.23	0.00	11.43	0.00	0.00
YI-VL [36]*	2024-06	6B	4.35	2.33	15.00	3.85	5.66	7.69	1.61	0.00	5.71	0.00	0.00
Molmo-7B-O [6]*	2024-09	7B	0.36	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
XComposer2.5 [37]*	2024-07	7B	3.99	4.65	20.00	0.00	3.77	0.00	3.22	0.00	2.86	0.00	0.00
mPLUG-Owl3 [35]	2024-07	7B	7.61	18.60	20.00	3.85	3.77	7.69	4.84	0.00	5.71	0.00	0.00
Qwen2.5-VL [3]	2025-01	7B	21.74	16.28	25.00	7.69	26.42	30.77	16.13	10.00	45.71	11.11	0.00
Bunny [8]*	2024-06	8B	4.71	4.65	10.00	3.85	1.89	0.00	4.84	0.00	5.71	11.11	20.00
Idefics3 [14]	2024-08	8B	9.78	11.63	10.00	7.69	7.55	15.38	6.45	10.00	20.00	0.00	0.00
MiniCPM-o 2.6 [11]	2025-01	8B	10.51	13.53	5.00	11.54	5.66	0.00	6.45	0.00	31.43	0.00	20.00
InternVL2.5 [4]	2024-12	8B	12.32	11.63	15.00	7.69	13.21	7.69	4.84	0.00	31.43	22.22	0.00
GLM-4v [7]*	2024-06	9B	2.54	2.33	5.00	0.00	5.66	0.00	1.61	0.00	0.00	11.11	0.00
11-78B Models													
Llama3.2-Vision [25]*	2024-09	11B	1.81	0.00	0.00	0.00	0.00	7.69	3.23	0.00	5.71	0.00	0.00
DeepSeek-VL2 [33]	2024-05	27B	1.09	0.00	0.00	0.00	5.66	0.00	0.00	0.00	0.00	0.00	0.00
InternVL2.5 [4]	2024-12	78B	14.49	18.60	20.00	7.69	11.32	7.69	6.45	0.00	34.29	22.22	20.00
Qwen2.5-VL [3]	2025-01	72B	39.86	30.23	35.00	26.92	50.94	38.46	35.48	0.00	**68.57**	33.33	40.00
Proprietary Multimodal Large Language Models													
GPT-4o mini [12]	2024-07	-	32.61	25.58	35.00	26.92	39.62	46.15	30.65	10.00	37.14	33.33	40.00
GeminiFlash2.0 [28]	2024-12	-	34.42	23.26	30.00	30.77	39.62	30.77	35.48	**30.00**	40.00	44.44	**60.00**
GPT-4o [12]	2024-11	-	**48.55**	**44.19**	**40.00**	**46.15**	**54.72**	**84.61**	**37.10**	20.00	65.71	**55.56**	40.00

Table 3. Evaluation of various models on MMCR. We evaluate model performance across ten categories of scientific comprehension tasks: Figure Comprehension (**FIC**), Multi-figure Comprehension (**MF**), Figure-Table Comprehension (**FTA**), Figure-Text Comprehension (**FTE**), Figure-Formula Comprehension (**FF**), Table Comprehension (**TAC**), Multi-table Comprehension (**MT**), Text Comprehension (**TEC**), Formula Comprehension (**FOC**), Pseudocode Comprehension (**PC**). For each category, we report both individual accuracy rates and overall performance metrics. The best results are marked **bold** and the second results are underlined. $*$ indicate that the input images are concatenated into one image.

4) **Figure-Text Comprehension:** Figures in scientific papers enormously necessitate specialized contextual knowledge for full comprehension. Questions of this task evaluate a model's ability to integrate information from both figures and text content to formulate accurate answers. If the model lacks a deep understanding of the text or fails to interpret the figures properly, these questions can pose considerable challenges.

5) **Figure-Formula Comprehension:** Formulas provide essential supplementary details that are crucial for understanding the content of figures. These questions evaluate a model's ability to answer queries that require interpreting both figures and formulas. To generate accurate responses, the model must effectively integrate the relevant formula with the corresponding figure.

6) **Table Comprehension:** Tables in scientific papers serve as a foundational component for empirical analysis. These questions are designed to evaluate the model's ability to extract information from individual tables and reason without explicitly referencing a specific table in the queries. This approach assesses the model's capacity to interpret and navigate tables within a scientific paper.

7) **Multi-table Comprehension:** Scientific papers frequently include multiple tables, and drawing conclusions through cross-table comparisons is a common practice. This task assesses a model's capability to extract and integrate information from several tables, enabling it to synthesize insights from diverse data sources effectively.

8) **Text Comprehension:** Given that scientific papers are text-dense, these questions evaluate a model's ability to extract and understand meaningful information from scientific papers.

9) **Formula Comprehension:** Formulas present condensed, symbolic information, which poses a greater challenge for the model compared to plain text. This task tests the model's ability to understand and extract information from the formulas in scientific papers.

10) **Pseudocode Comprehension:** Although pseudocode appears less frequently than figures or tables in scientific papers, it often provides crucial additional context. These questions test the model's ability to interpret the symbolic and graphical components of pseudocode.

Figure 2 presents an annotated example. The questions are designed to assess the model's ability not only to derive

answers from cross-source information but also to comprehend content across pages, as relevant evidence may be distributed throughout the document. This presents heightened challenges for the model's performance. Detailed examples are provided in the supplementary material.

In our answer annotation process, we employed a multiple-choice format that required annotators to generate a minimum of eight questions per scientific paper. To ensure diversity in question types, annotators were instructed to avoid repetitive question patterns. Each question was mandated to have five distinct answer options; questions that could not meet this requirement were excluded. Papers yielding fewer than eight valid questions were omitted from the annotation.

3.3. Quality Control

To enhance the annotation quality of the benchmark, we implemented a two-round semi-automatic quality control procedure that combines the merits of manual annotations and LLMs.

LLMs-Assisted Data Filtering. Despite the meticulous efforts of our annotators, some questions addressed general knowledge that large language models could accurately answer based solely on their internal knowledge or by reasoning directly from the question and options. To evaluate this, we employed two language models, Llama-3.3-70B [25] and Qwen2.5-72B [34], to perform reasoning using only the questions and options as input, without incorporating any associated scientific papers, images, or supplementary text. Questions that the models answered correctly were flagged as potential low-quality questions. However, before discarding these questions entirely, we carefully reviewed the reasoning processes of the models to ensure their correct answers were not the result of random guessing.

Cross-Checking. Annotators conducted parallel cross-checking of each other's work to enhance the quality of the benchmark. Errors identified during this process were either corrected or eliminated. Furthermore, a voting mechanism was implemented to filter out question-and-answer pairs that were overly similar in structure or content, ensuring greater diversity and reliability in the final benchmark.

3.4. Evaluation Method

With only five answer options, randomly selecting an option would result in an accuracy of approximately 20%, which potentially reduces the discernible performance differences among VLMs. To address this, and inspired by MMBench [19], we implemented a circular evaluation to further minimize the likelihood of the model guessing the correct answer. In this approach, the order of the answer options is rotated five times and presented to the VLMs. A model is considered to have successfully solved the problem only if it answers correctly in all five tests.

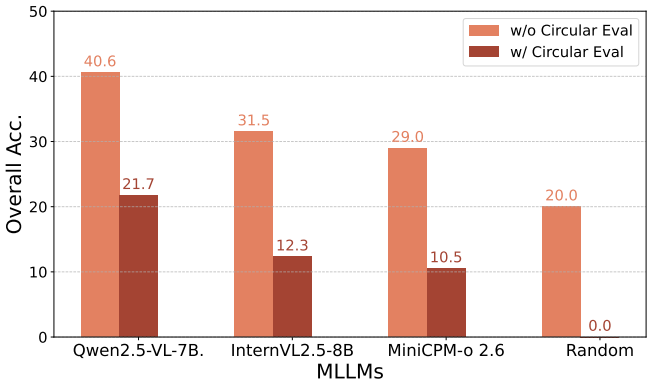

Figure 3. Comparison of the overall accuracy of VLMs with and without circular eval.

3.5. Statistics of Benchmark

The detailed statistics of our benchmark are presented in Table 2. This benchmark is distinguished by its focus on cross-source reasoning, encompassing a total of 31 scientific papers with an average length of 19 pages per paper, spanning 7 diverse subjects. A total of 276 questions have been annotated, distributed across 10 distinct categories, with an average of 8.9 questions per paper. This comprehensive benchmark is designed to rigorously evaluate the capability of VLMs to perform reasoning by integrating information from multiple sources within scientific papers.

4. Experiments and Analysis

4.1. Evaluation Protocol

We conduct a three-step evaluation protocol: response generation, answer extraction, and score calculation. Specifically, in the answer generation stage, we deliberately refrain from imposing task-specific constraints on the model, allowing it to generate responses freely. Subsequently, we apply a heuristic rule-based extraction method to extract the model's output. The specific heuristic rule-based extraction method is described in detail in the supplementary material. Finally, we compute the exact match score to evaluate the model's performance. We report both the overall accuracy as well as the accuracy for each individual category.

4.2. Experimental Setup

We evaluated 18 VLMs, comprising 15 open-source models and 3 proprietary models. To rigorously assess the models' capability for reasoning across multiple sources of information in scientific papers, we re-rendered the PDF-format papers into high-resolution JPEG images at 144 DPI and fed them into the VLMs in an end-to-end approach. However, not all models support multi-image input. For those not trained on multi-image or video datasets, we concatenated the high-resolution JPEG images into a single large

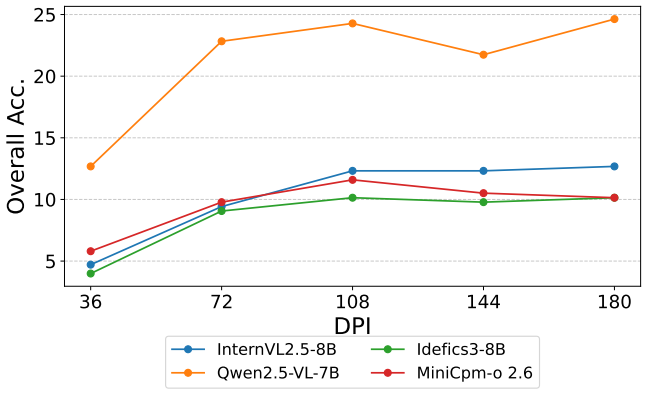

Figure 4. Model performance on increasing rendering DPI.

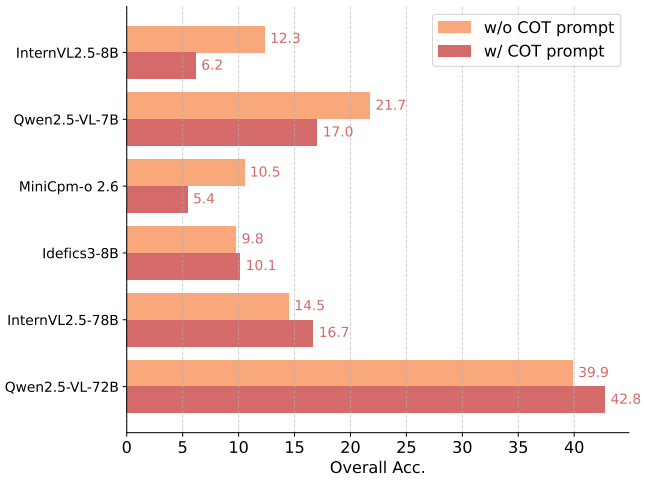

Figure 5. Comparison of the overall accuracy of VLMs with and without CoT prompt.

image before inputting it into the models. For models that do support multi-image input, we conducted evaluations using the native multi-image input method.

4.3. Main Results

The experimental results are comprehensively summarized in Table 3, which details both category-specific and overall accuracy metrics. From these findings, three principal conclusions emerge:

Proprietary models demonstrate superior performance. The proprietary GPT-4o model achieved state-of-the-art results on our benchmark, attaining overall accuracy of 48.55%. This represents an 8.69% performance advantage over the leading open-source model, Qwen2.5-VL-72B. Notably, GPT-4o matched or exceeded Qwen2.5-VL-72B's performance across nine distinct question categories, underscoring the persistent performance gap in multi-modal reasoning capabilities between proprietary and open-source architectures for scientific document analysis.

Current VLMs face significant challenges in reasoning with cross-source information. Even the highest-performing proprietary VLM (GPT-4o) failed to surpass 50% overall accuracy, while demonstrating sub-50% performance in five question categories. Open-source models exhibited more pronounced limitations, with most models achieving less than 10% overall accuracy. These results collectively highlight the inherent challenge of our benchmark.

Tabular comprehension presents unique difficulties. Comparative analysis revealed that 72.22% (13/18) of evaluated models exhibited equal or inferior performance on table comprehension questions compared to figure comprehension questions. When extended to multi-element comprehension tasks, all tested models demonstrated equivalent or reduced accuracy. This systematic performance discrepancy suggests inherent challenges in tabular data interpretation within complex scientific papers.

4.4. Is Circular Evaluation more Robust?

In Figure 3, we evaluated three models and theoretical random selection with and without circular evaluation. As shown in the figure, all three models experienced significant performance degradation under circular evaluation, with MiniCPM-o 2.6's performance dropping to approximately 36% of its performance without circular evaluation. The accuracy of random selection decreased from 20% to $0.2^5 \times 100\%$. These results demonstrate that circular evaluation substantially reduces the probability of obtaining correct answers through random selection, validating the robustness of our benchmark.

4.5. Does High-Resolution Image Input Improve Model Performance?

In the case of single-image input, increasing the image resolution typically enhances model performance [5, 15, 17],. However, with multi-page image input, the large number of visual tokens present a significant challenge for the model in extracting useful information. Further increasing the image resolution also increases the number of visual tokens. How do the rich information and large number of visual tokens brought by high-resolution images affect the model's performance? To explore this question, we conducted experiments using four VLMs. The DPI of PDF-rendered JPEG images was gradually increased from 36 to 180. As shown in Figure 4, all four models exhibited steady performance improvements as the DPI increased to 108. However, when DPI exceeds 108, the model's performance fluctuates and even declines as DPI increases. This suggests that while moderately increasing resolution can benefit model performance for multi-image inputs, the model struggles to handle images with excessively high resolution.

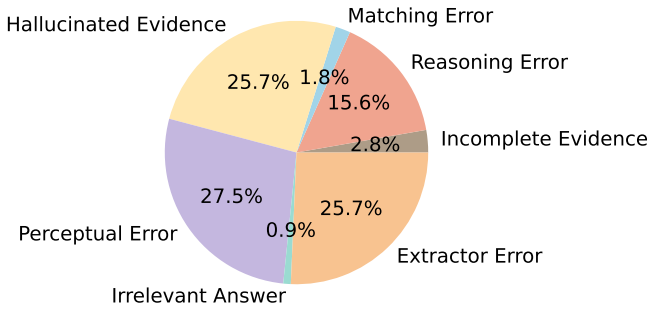

Figure 6. Error distribution over 109 annotated GPT-4o errors.

4.6. Does CoT Help in Answering Cross-Source Reasoning Questions?

Figure 5 presents the impact of CoT prompt on our benchmark. Specifically, we designed CoT prompt, as outlined in the supplementary material, and evaluated the performance of the VLMs with and without CoT prompt. Our experiments revealed that smaller models experienced a notable performance decline when using CoT prompt. To explore this further, we evaluated larger models, such as InternVL2.5-78B and Qwen2.5-VL-72B, and observed that their performance improved with the use of CoT prompt. We hypothesize that the smaller models may struggle with severe hallucinations when confronted with a high number of visual tokens, or they may fail to identify sufficient sources of evidence in tasks requiring cross-source reasoning, which leads to a significant drop in CoT performance. In contrast, the larger models appear better equipped to handle the complexity of visual tokens and more effectively retrieve relevant evidence for coherent chain-of-thought reasoning. Interestingly, for the Idefics3-8B model, CoT prompt led to a slight improvement in performance. A detailed review of the Idefics3-8B reasoning process revealed that it still delivered answers directly without engaging in reasoning under the CoT prompt. These results highlight the varying effects of CoT across models of different sizes, particularly in their capacity to leverage cross-source in multi-page scientific papers. They suggest that the ability to effectively extract and integrate cross-source information from numerous visual tokens is key to improving performance with CoT prompt.

4.7. Error Analysis

We further conducted an error analysis to identify the bottlenecks in current multimodal models when handling cross-source reasoning tasks. Specifically, we selected 109 instances where GPT-4o provided incorrect answers and performed a manual examination. The errors were categorized into the following types: 1) **Hallucinated Evidence:** The model failed to locate the necessary evidence to answer the question. 2) **Incomplete Evidence:** The evidence retrieved

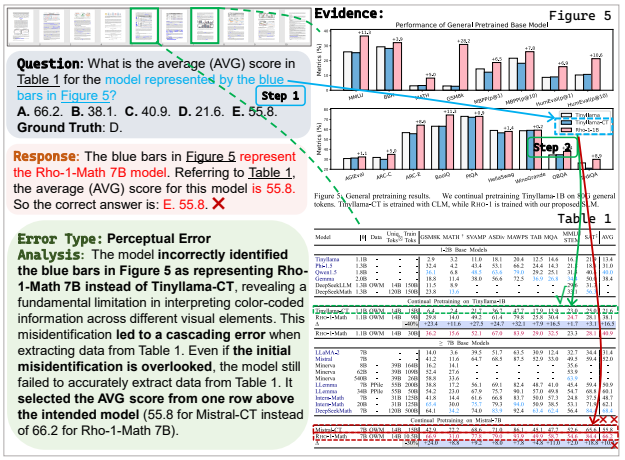

Figure 7. An example of a perceptual error is observed in GPT-4o, where the model fails to accurately interpret both the figure and the table, resulting in unsuccessful cross-source reasoning. The correct reasoning process is outlined in Step 1 and Step 2.

was insufficient. 3) **Perceptual Error:** The model misinterpreted graphs, symbols, colors, etc. 4) **Extractor Error:** The model extracted values that did not exist in the table or extracted values from misaligned tables. 5) **Reasoning Error:** The model made incorrect inferences based on the retrieved information. 6) **Irrelevant Answer:** The model's answer was unrelated to the question. 7) **Matching Error.** Rule-based methods failed to extract the correct response from the model. The distribution of these errors is shown in Figure 6. Our findings reveal that the most prevalent error is perceptual error, followed by extractor error, highlighting the model's difficulty in accurately retrieving relevant information from multi-page images. Hallucinated and incomplete evidence together account for 28.5% of the errors, further underscoring the substantial room for improvement in current VLMs in leveraging cross-source information for reasoning. Figure 7 presents our analysis of a perceptual error, offering insights into its underlying causes.

5. Conclusion

In this work, we present MMCR, a benchmark for evaluating VLMs' ability to perform cross-source reasoning in scientific papers. Our extensive experiments reveal that current VLMs struggle to identify complete and accurate evidence required for reasoning from cross-source information in scientific papers. We hope that the development of this benchmark will promote advancement in VLMs' capabilities to perform reasoning utilizing cross-source information.

Acknowledgement. This work was funded by National Natural Science Foundation of China under No. 62306046. We thank MolarData for their technical and resource support.

References

[1] The claude 3 model family: Opus, sonnet, haiku. 1

[2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 5

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 4, 5

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 5

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 4, 7

[6] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 5

[7] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 5

[8] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024. 5

[9] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 1

[10] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024. 1

[11] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. 5

[12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 5

[13] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. 3

[14] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024. 5

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 7

[16] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 1, 2, 3

[17] 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. 7

[18] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. *arXiv preprint arXiv:2407.04903*, 2024. 2, 3

[19] 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?(2023). *arXiv preprint arXiv:2307.06281*, 2023. 6

[20] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 1, 4

[21] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024. 2, 3, 4

[22] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1, 2, 3

[23] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceed-*

ings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021. 1, 2, 3

[24] 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. 1, 3

[25] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20:2024, 2024. 5, 6

[26] Lingdong Shen, Kun Ding, Gaofeng Meng, Shiming Xiang, et al. Rethinking comprehensive benchmark for chart understanding: A perspective from scientific literature. *arXiv preprint arXiv:2412.12150*, 2024. 1, 2

[27] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13636–13645, 2023. 2, 4

[28] 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. 1, 5

[29] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023. 1, 2, 3

[30] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 1, 2, 3

[31] 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. 4

[32] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. 2, 3

[33] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 5

[34] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 6

[35] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 5

[36] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 5

[37] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 5

[38] Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tong Sun. Llava-read: Enhancing reading ability of multimodal language models. *arXiv preprint arXiv:2407.19185*, 2024. 4