M-LongDoc: A Benchmark For Multimodal Super-Long Document Understanding And A Retrieval-Aware Tuning Framework

Yew Ken Chia[*, 1, ♖] **Liying Cheng**[1] **Hou Pong Chan**[†, 1, 2] **Chaoqun Liu**[1, 3]
Maojia Song[♖] **Sharifah Mahani Aljunied**[1] **Soujanya Poria**[3, ♖] **Lidong Bing**[†, 1]
[♖] Singapore University of Technology and Design
[1] DAMO Academy, Alibaba Group, Singapore
[2] Hupan Lab, 310023, Hangzhou, China
[3] Nanyang Technological University, Singapore

Abstract

The ability to understand and answer questions over documents can be useful in many business and practical applications. However, documents often contain lengthy and diverse multimodal contents such as texts, figures, and tables, which are very time-consuming for humans to read thoroughly. Hence, there is an urgent need to develop effective and automated methods to aid humans in this task. In this work, we introduce M-LongDoc, a benchmark of 851 samples, and an automated framework to evaluate the performance of large multimodal models. We further propose a retrieval-aware tuning approach for efficient and effective multimodal document reading. Compared to existing works, our benchmark consists of more recent and lengthy documents with hundreds of pages, while also requiring open-ended explanations and not just extractive answers. To our knowledge, our training framework is the first to directly address the retrieval setting for multimodal long documents. To enhance open models, we construct a training corpus in a fully automatic manner. Experiments show that our tuning approach significantly improves the correctness of model responses by 4.6%. [1]

1 Introduction

The ability to comprehend long and complex multimodal documents is crucial in practical applications such as business intelligence, academic literature review, and legal research (Mathew et al., 2020). Recently, large multimodal models such as GPT-4V (OpenAI, 2023) have shown great potential in processing and analyzing diverse types of information, including text, images, and even structured data (Huang et al., 2024b). These models offer the promise of automating tasks that traditionally required extensive human effort, such as document analysis and and question-answering (Fujitake, 2024). However, real-world documents often present significant challenges due to their length and multimodal content, often involving text, figures, tables, and charts (Faysse et al., 2024). Full comprehension of long multimodal documents may require challenging aspects such as analytical reasoning, fine-grained visual interpretation, and domain knowledge. On the other hand, we observe that existing benchmarks often fall short in representing these challenges, typically focusing on short documents with less than 50 pages, and limited to simpler extraction-based questions (Ma et al., 2024).

To address these limitations, we introduce M-LongDoc, a comprehensive benchmark consisting of 851 samples specifically designed to evaluate the performance of large multimodal models on lengthy and diverse documents. Unlike previous works (Mathew et al., 2020; Liu et al., 2024; Ma et al., 2024) that mainly feature shorter documents, M-LongDoc involves recent documents *spanning hundreds of pages, encompassing diverse domains* as shown in Figures 1 and 2. In addition, as shown in Figure 3, our benchmark *goes beyond simpler extractive questions, requiring models to provide open-ended solutions* that demonstrate in-depth reasoning over the document content (Fan et al., 2019). M-LongDoc poses a question answering task where models have to analyze and reason over texts, figures, or tables in each multimodal long document.

Another challenge is the difficulty of evaluating open-ended model responses over complex multimodal documents. To assess such open-ended outputs in a scalable and standardized manner, we design an automated evaluation framework that does not require reference answers or human annotation. Inspired by previous works in model-based

*Yew Ken and Chaoqun were students under the Joint PhD Program between Alibaba and their corresponding university. Work done while Liying, Mahani, and Lidong were at Alibaba.

† Corresponding authors.

[1] Our multimodal benchmark, training corpus, and source code are publicly available at https://multimodal-documents.github.io/.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9233–9250
November 4-9, 2025 ©2025 Association for Computational Linguistics

Figure 1: Data distribution of document topics in our M-LongDoc benchmark.

	Pages	Tokens	Answer Length
DocVQA	1.0	151.5	2.4
ChartQA	1.0	236.9	1.1
InfoVQA	1.2	288.0	1.6
TAT-DQA	1.1	577.0	4.2
VisualWebBench	1.0	452.4	2.3
MP-DocVQA	8.3	2026.6	2.4
DUDE	5.7	1831.5	3.0
SlideVQA	20.0	151.5	2.1
MMLongBench	47.5	2030.5	2.6
Ours	210.8	120988.0	180.3

Figure 2: Comparison of document understanding benchmarks along three dimensions: the number of pages per document, the number of tokens per document, and the average length of responses. Specifically, we assess whether each benchmark emphasizes in-depth, comprehensive explanations or focuses on short or extractive answers. We include additional details and comparisons in Appendix A.8.

evaluation (Zheng et al., 2023; Zhao et al., 2024; Liu et al., 2023b), our evaluation framework leverages a detailed evaluation guide and multiple judge models to score the correctness of each response, achieving high agreement with human evaluators.

With the proposed M-LongDoc and evaluation framework, our preliminary study on existing models show that they struggle with figure and table-based questions compared to text-based questions, revealing their multimodal bias and weaknesses (Chen et al., 2024b). Furthermore, we observed a critical weakness that the models can be easily distracted by irrelevant content in the document pages (Shi et al., 2023), even with the aid of retrieval.

To enhance the robustness of multimodal models against potentially irrelevant retrieved content, we propose a retrieval-aware tuning approach for multimodal document reading. This framework unifies supervised fine-tuning and retrieval augmented generation by including distracting content from other modalities and pages in each document. Thus, we adapt models to effectively incorporate the domain knowledge in multimodal documents while ignoring the irrelevant contents or modalities. To our knowledge, this is the first tuning approach for retrieval-augmented multimodal document understanding. To support this training framework and the enhancement of open-source models, we construct a training corpus of 10,070 question-explanation pairs. Experiments show that our approach achieves a 4.6% relative improvement in the correctness of model responses.

The key contributions of this work are threefold: 1) We introduce M-LongDoc, a multimodal benchmark that more accurately represents the challenges of real-world document understanding tasks. Our automated evaluation framework enables scalable and standardized assessment of open-ended solutions. 2) Our evaluation of leading models indicates that most models struggle with figure and table-based questions compared to text-based questions, revealing their multimodal bias. 3) We propose a retrieval-aware tuning framework that together with our large-scale training corpus, significantly improves the efficiency and effectiveness of multimodal document reading.

Thus, we believe that this work contributes to the field of document understanding and paves the way for more capable and practical applications of large multimodal models in real-world scenarios.

2 M-LongDoc Benchmark

To evaluate the multimodal document understanding ability of existing models, we present M-LongDoc, a challenging and diverse benchmark. Notably, we focus on open-ended questions that require in-depth reasoning and analysis over very long documents with more than 200 pages on average. As shown in Appendix A.8, the questions cover the academic, financial, and product domains, involving aspects such as analytical reasoning, domain knowledge, and visual interpretation.

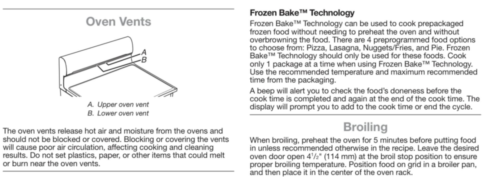
Dataset	Example Question	Example Answer
DocVQA	What is the underlined heading just above the table?	Indications for implantation
	What is the Extension Number as per the voucher?	(910) 741-0673
MMLongBench	What is the number of red logos in page 10?	0
	I'm at the location "J" shown in the campus map. Tell me the name of the nearest coffee shop.	Ten Years After Cafe
Ours	 Oven Vents A. Upper oven vent B. Lower oven vent The oven vents release hot air and moisture from the ovens and should not be blocked or covered. Blocking or covering the vents will cause poor air circulation, affecting cooking and cleaning results. Do not set plastics, paper, or other items that could melt or burn near the oven vents. **Frozen Bake™ Technology** Frozen Bake™ Technology can be used to cook prepackaged frozen food without needing to preheat the oven and without overbrowning the food. There are 4 preprogrammed food options to choose from: Pizza, Lasagna, Nuggets/Fries, and Pie. Frozen Bake™ Technology should only be used for these foods. Cook only 1 package at a time when using Frozen Bake™ Technology. Use the recommended temperature and maximum recommended time from the packaging. A beep will alert you to check the food's doneness before the cook time is completed and again at the end of the cook time. The display will prompt you to add to the cook time or end the cycle. Broiling When broiling, preheat the oven for 5 minutes before putting food in unless recommended otherwise in the recipe. Leave the desired oven door open 4 1/2" (114 mm) at the broil stop position to ensure proper broiling temperature. Position food on grid in a broiler pan, and then place it in the center of the oven rack. Where are the oven vents located on this range model, and why is their positioning important for proper oven function?	The oven vents are located at the top front of the oven, with one vent on the upper front and another on the lower front. Their positioning is important for proper oven function because they release hot air and moisture from the oven during cooking and cleaning. Blocking or covering the vents can cause poor air circulation, affecting cooking and cleaning results. The vents also help to maintain a consistent temperature in the oven by releasing excess heat and preventing the oven from overheating.

Figure 3: Example questions in different multimodal document benchmarks, showing the more complex explanations required for our benchmark. For illustration, we include a screenshot from the original document pages.

2.1 Data Collection

To support our evaluation benchmark, we manually source high-quality multimodal documents from publicly accessible sources. Concretely, we source research papers[2], company reports[3] and product instruction manuals[4] for the academic, financial, and product domains respectively. Thus, the dataset covers a range of document formats and domains. As research papers often require domain expertise, we constrain the academic domain to computer science topics. To reduce the risk of data contamination or memorization when evaluating existing models (Dong et al., 2024), we constrain the documents to be published in January 2024 or later. As most existing models are unable to process raw PDF files, we conduct a simple data processing to extract the texts and relevant images from each document. Specifically, we use the PyMuPDF[5] tool to automatically extract the text from each page. To extract the figures and tables from each page, we leverage an existing object detection model (Pfitzmann et al., 2022). Thus, the processed data consists of interleaved texts and images, where the figures and tables are extracted as images.

2.2 Question Generation

To construct diverse and challenging open-ended questions, we leverage a semi-automated pipeline. Concretely, as shown in Figure 4, given a specified content category, we first randomly select a page from the document that contains the specific content category, such as texts, tables, or figures. Consequently, we randomly select a model from a pool of leading multimodal models and instruct it to generate a challenging question based on the document page. To ensure that the question generator has sufficient context, we also provide the previous page and subsequent page as additional inputs during the question generation process.

To improve the quality of the generated questions, we conduct an automated verification process as a preliminary filter. Concretely, the question generator is also instructed to validate the generated question by following a multi-step checklist. For example, the checklist includes checking if the question is relevant to the content, if the content is extracted correctly and required to answer the question, and whether the question is answerable. The question is rejected if it does not satisfy any condition in the checklist. Lastly, we employ a team of annotators to conduct final validation for each question. We employ expert annotators who are Ph.D. level and above in computer science for the academic domain, and professional annotators for the finance and product domains. To be consistent, we provide a similar checklist and instruction as our automated verification stage, as shown in Appendix A.2. We found that 80.1% of the generated questions satisfied the automated verification. Of these questions that passed automated verification, 80.9% also satisfied the human verification. Thus, we only retain questions that satisfied both

[2] https://arxiv.org
[3] https://www.annualreports.com
[4] https://www.manualslib.com
[5] https://pymupdf.readthedocs.io

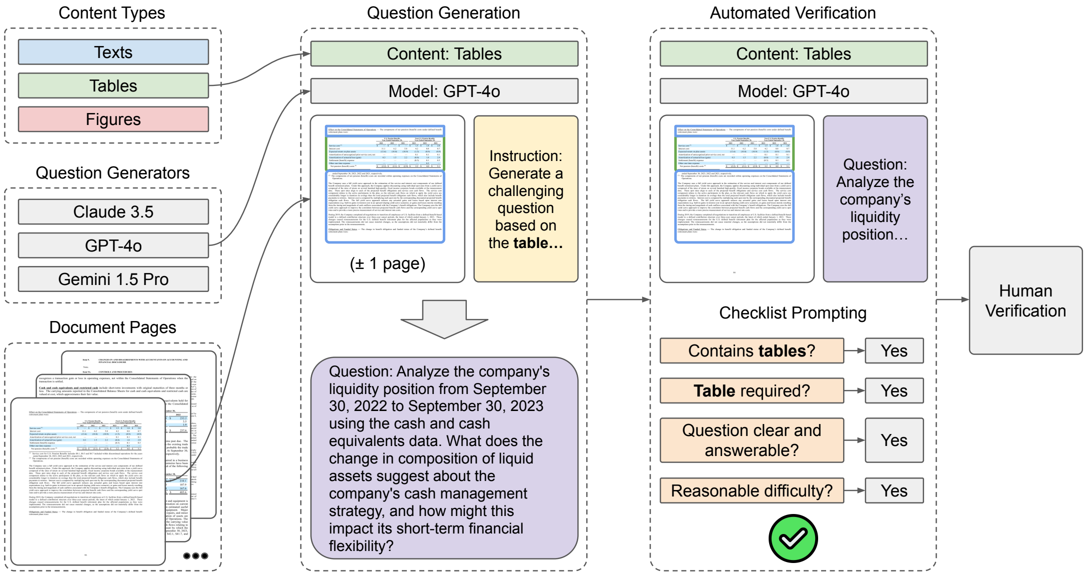


Figure 4: Overview of our data construction process with question verification stages. For brevity, we shorten the checklist prompts and include the full details in Appendix A.2.

the automated and human verification.

The statistics of our benchmark dataset are shown in Appendix A.1, where we ensure a balanced distribution of questions and documents from each domain and question category. In this work, we focus on questions that require a single page of content to answer, and only retain answerable questions. Please also note that while each question focuses on a specific category in a document page, the page may contain content from other categories as context. For instance, a table-based question may also require comparisons to other tables or texts from the same page.

Compared to the existing benchmarks in Figure 2, M-LongDoc poses a greater challenge in two main aspects. Firstly, the significantly greater number of pages and tokens in each multimodal document poses extreme computational costs and opportunities to be distracted by irrelevant content (Shi et al., 2023). While this challenge may be mitigated by retrieval-augmented generation (Chen et al., 2022), our preliminary study in Section 2.4 shows that existing models are still hindered by their multimodal bias (Chen et al., 2024b). In addition, our benchmark poses challenging open-ended questions as shown in Figure 3, requiring models to produce in-depth analyses in their outputs. Thus, we believe M-LongDoc is a more realistic and challenging benchmark compared with existing datasets focusing on short answers that can often be extracted directly from the source document.

2.3 Automated Evaluation

Given the challenging nature of our multimodal long document benchmark, it is crucial to have a scalable and standardized evaluation method. However, it is less feasible to conduct comprehensive human evaluation due to high labour costs and lack of reproducibility (Clark et al., 2021). Thus, inspired by previous works in automatic evaluation (Zheng et al., 2023; Zhao et al., 2024; Liu et al., 2023b), we propose an evaluation framework based on a committee of multimodal judges. Concretely, we leverage multiple leading multimodal models to score each answer to a question based on the criteria of correctness. To provide a clear guideline for evaluation, we define the task introduction, criteria, and evaluation steps as shown in Figure 5.

To provide a more reliable evaluation and reduce intra-model bias (Verga et al., 2024), we leverage multiple judges to evaluate each candidate answer. Specifically, each judge model M_j is provided with the evaluation guide g, ground-truth evidence page as context c, question q, and candidate answer $\hat{a}$, and instructed to assign a correctness score from 1 to 5. To reduce variance, we sample multiple scores from each judge model M_j and aggregate the scores to obtain a fine-grained, continuous score that better reflects the quality of the answer:

$$\text{Score} = \frac{1}{J \cdot K} \sum_{j=1}^{J} \sum_{k=1}^{K} s_{j,k} \sim M_j(g, c, q, \hat{a}) \quad (1)$$

where $J = 3$ is the number of judge models and

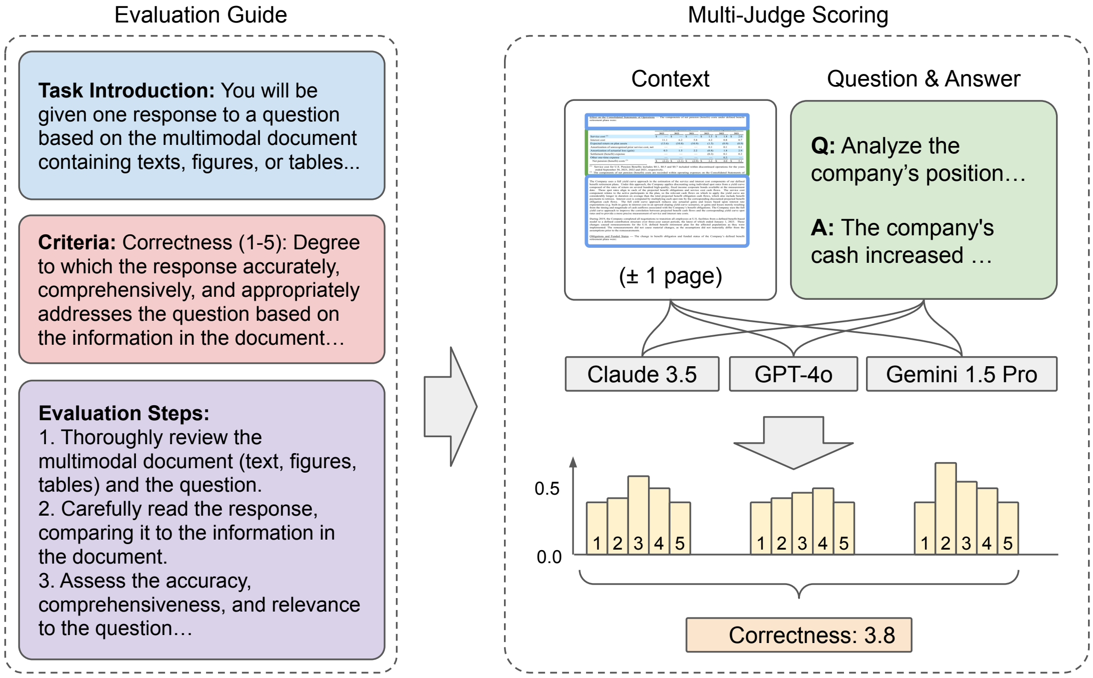


Figure 5: Our automated evaluation framework to assess the correctness of open-ended explanations for multimodal question answering. The full evaluation guide is included in Appendix A.4.

	Text	Figure	Table
Gemini-1.5-pro-002			
w/ top $k = 1$ pages	4.38	3.73	4.16
w/ top $k = 5$ pages	4.60	4.31	4.54
w/ top $k = 10$ pages	4.61	4.29	4.62
w/ top $k = 20$ pages	4.63	4.33	4.38
Qwen2-VL-7B-Instruct			
w/ top $k = 1$ pages	4.05	3.25	3.36
w/ top $k = 5$ pages	4.17	3.67	3.46
w/ top $k = 10$ pages	4.08	3.62	3.19
w/ top $k = 20$ pages	OOM	OOM	OOM

Table 1: Preliminary study on model correctness for text, figure, and table-based questions respectively.

$K = 5$ is the number of sampled scores per judge.

2.4 Preliminary Study

To investigate the limitations of existing models, we conduct a preliminary study on a subset of 100 random samples from our M-LongDoc benchmark. Concretely, we select Gemini (Google, 2024) and Qwen2-VL (Wang et al., 2024a) to represent highly capable models for the close and open-source settings respectively. Despite recent advances, current models often struggle with very long multimodal documents and may incur great computational costs (Dingjie et al., 2024). Thus, we focus our study on the retrieval-augmented generation paradigm (Lewis et al., 2020), which practically retrieves the most relevant content for the multimodal question answering model. Concretely, we use ColPali (Faysse et al., 2024) as a state-of-the-art multimodal retriever and leverage the top k pages of multimodal content as context. The retriever details are discussed in Appendix A.3.

Notably, as shown in Table 1, we observe that current models are weaker in processing image-based contents such as figures and tables in multimodal documents, and may be biased towards the textual content, even when they are trained on interleaved multimodal data (Chen et al., 2024b). Furthermore, increasing the amount of retrieved content may not improve overall performance, and may even lead to worse performance or out-of-memory (OOM) issues. This indicates a key bottleneck that the models may be easily distracted by irrelevant content (Shi et al., 2023).

Additionally, to measure the reliability of our automated evaluation, we conduct manual human scoring based on the same evaluation guide. For the samples in this preliminary study, we find a high Pearson correlation of 88.9% with $p < 0.001$ between the aggregated judge score and human scoring. Thus, our framework can reliably judge open-ended explanations for multimodal documents.

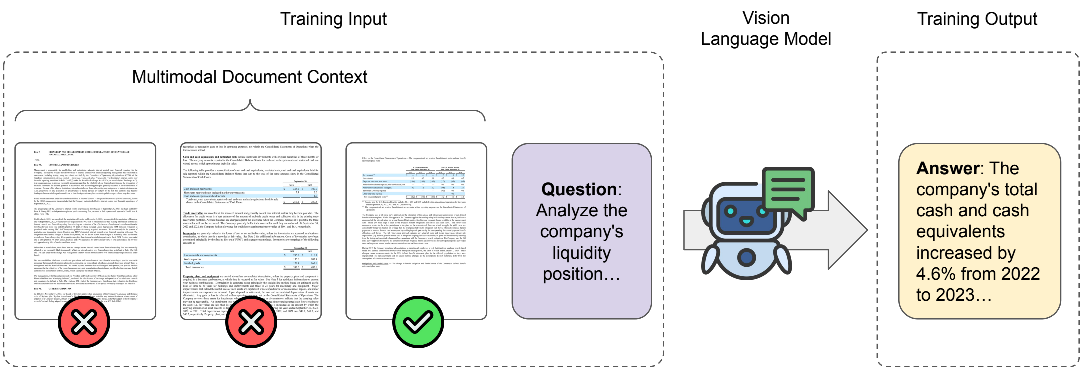


Figure 6: Our retrieval-aware multimodal tuning framework enhances the ability of models to identify and utilize pertinent content in multimodal documents. At training time, the model is provided with more relevant pages retrieved from the document, which may contain both the gold evidence page and multiple 'distractor' pages.

3 Retrieval-Aware Multimodal Tuning

To fully understand multimodal documents, models depend on skills such as fine-grained visual interpretation, technical analysis and reasoning as shown in Appendix A.8. Furthermore, the long documents require models to identify and leverage only the pertinent content. However our preliminary study in Section 2.4 have shown that despite retrieval augmentation, models can still be easily misled by some irrelevant information in the multimodal documents. To our knowledge, there is no open-source model that addresses these challenges.

To this end, we propose a simple and effective retrieval-aware multimodal document tuning approach, building upon previous works in retrieval augmentation (Chen et al., 2022; Zhang et al., 2024). Specifically, we integrate multimodal context from the ground-truth evidence page as well as potentially irrelevant pages during training. As shown in Figure 6, this presents a training paradigm that is more realistic and similar to the challenges faced during test-time retrieval of multimodal content. Thus, the model is tuned to identify and leverage the most pertinent multimodal information.

To construct the training data, we leverage the same process as shown in Figure 4 to construct a training corpus of 10,070 samples across 300 documents, and leverage the respective question generator models to also produce a high-quality answer based on each ground-truth evidence page. Our analysis in Appendix A.5, shows that the training data answers are of very high quality.

4 Experiments

4.1 Task Setting

To ensure a practical task setting, we focus on the retrieval-based paradigm, which avoids the exorbitant cost to process the full document. Based on our preliminary study in Section 2.4, we use the top $k = 5$ pages ranked by the retriever as a reasonable amount of context for each question. Thus, each model is provided with the retrieved context and question as input, to generate an open-ended explanation as output. As discussed in Section 2.3, we leverage our automated framework with multiple judge models to score the correctness of each output response, on a scale of 1 to 5.

4.2 Models

To provide a more comprehensive investigation of current models, we cover both open-source and close-source models in this work. Concretely, we select GPT-4o (gpt-4o-2024-05-13)[6], Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)[7] and Gemini 1.5 Pro (gemini-1.5-pro-002) (Google, 2024) due to their leading performance on multimodal benchmarks (Yue et al., 2023). Regarding open-source models, we specifically select models which support interleaved multimodal inputs with multiple images, and fine-grained visual perception of structured content. Thus, we mainly focus on Pixtral-12B (Agrawal et al., 2024), LLaVA-OneVision-7B (Li et al., 2024) and Qwen2-VL-7B-Instruct (Wang et al., 2024a). Due to training instabilities with other models, we mainly focus the training experiments on the Qwen2-VL-7B-

[6] https://openai.com/index/gpt-4o-system-card/
[7] https://www.anthropic.com/claude/sonnet

Model	Size	Domain			Question Category			
		Academic	Product	Finance	Text	Figure	Table	All
Proprietary Models								
GPT-4o	-	4.56	4.38	4.51	4.55	4.38	4.53	4.49
Claude 3.5 Sonnet	-	4.59	4.43	4.51	4.57	4.42	4.54	4.51
Gemini 1.5 Pro	-	4.66	4.43	4.43	4.59	4.43	4.52	4.51
Open-Source Models								
Pixtral	12B	4.42	4.14	4.06	4.38	4.20	4.09	4.22
LLaVA OneVision	7B	3.71	3.74	3.39	4.03	3.57	3.30	3.62
Qwen2-VL	7B	4.03	3.88	3.56	4.08	3.83	3.62	3.84
w/ Retrieval Tuning	7B	**4.17**	**4.01**	**3.86**	**4.31**	**4.00**	**3.77**	**4.02**

Table 2: Evaluation of model performance for proprietary and open-source multimodal models. We report the correctness on our benchmark across different document domains and question categories. We bold the highest scores obtained by open-source models among 7B models.

Instruct model, which demonstrates leading performance compared to similar-sized models. Detailed hyperparameters are included in Appendix A.6.

5 Results

To assess the effectiveness of our approach and the performance of existing models, we report the main evaluation results in Table 2. First, we find that our retrieval-aware multimodal tuning significantly and consistently enhances the performance of Qwen2-VL, representing a relative improvement of 4.6% in answer correctness. Thus, we view the proposed training approach as a promising strategy to enhance multimodal long document understanding. Second, open-source models show notably worse performance in image-based questions involving figures and tables. This discrepancy highlights the need for more efforts to enhance the visual interpretation and reasoning ability of open models.

Additional analysis in Appendix 5.1 shows the importance of leveraging all modalities for document understanding, instead of conventional text-only settings. Furthermore, Appendix A.7 demonstrate the robustness of our method under varying numbers of retrieved pages, and Appendix A.8 provides case studies of model outputs.

5.1 Importance of Multimodal Processing

To investigate the importance of multimodal content for long documents, we explore the text-only setting without images, and full-image setting without extracted texts. As discussed in Section 2.1 our main setting first extracts the texts, figures, and tables separately from each page, with the figures and tables represented as individual images. However, results in Table 3 shows significant degradation of 12.0% and 6.6% respectively when the input images for figures and tables are omitted. While the model may answer questions to a limited extent due to partial information of figures and tables in the text, this setting is clearly suboptimal. Furthermore, the slight benefit for text-based questions in the absence of figure and table images, may indicate that models still cannot leverage the full multimodal content effectively.

Model	Question Category		
	Text	Figure	Table
Qwen2-VL	4.08	3.83	3.62
w/o Image Inputs	4.22	3.37	3.38
w/ Render Page as Inputs	3.99	3.70	3.39

Table 3: Analysis on alternative settings for our benchmark, including removing images from model inputs, and using only the render image of each page as document context, without text extraction.

On the other hand, while rendering each page as a single unified image instead of separate extraction may benefit retrieval (Faysse et al., 2024), we do not observe the same benefits for document understanding. While the rendered page image does contain the original information and layout of the document, including texts, tables, and figures, models may be less capable of distinguishing the content between texts and tables.

6 Related Work

Large Multimodal Models Recently, large multimodal models have demonstrated their capability to process and comprehend data across various formats. Leading models such as GPT-4o (AI, 2024), Claude 3.5 Sonnet (Anthropic, 2024), and Gemini 1.5 Pro (Google, 2024) show marked improvements in visual understanding and reasoning tasks. Open-source models such as Llava (Liu et al., 2023a), Idefics (LaurenÃğon et al., 2023), Otter (Li et al., 2023), and InternVL (Chen et al., 2024c) have also shown promise in processing diverse multimodal content including document images (Mathew et al., 2020), slides (Tanaka et al., 2023), and charts (Huang et al., 2024a). However, the benchmark performance of open-source models tends to lag behind that of close-source ones (Yue et al., 2023), highlighting the need to bridge the gap. Our framework offers to significantly improve multimodal long document understanding.

Document Understanding Given the practical and business applications of document understanding, researchers have devoted significant effort to this area by introducing new datasets and methods. SearchQA (Dunn et al., 2017), NarrativeQA (Kočiskỳ et al., 2018), QuALITY (Zhu et al., 2020) are reading comprehension datasets over purely textual data, while FinQA (Chen et al., 2021), DocFinQA (Reddy et al., 2024) are introduced in the financial domain. DocVQA(Mathew et al., 2020) presents a visual question answering dataset on document images. VisualWebBench (Liu et al., 2024) is a multimodal benchmark over single-page documents focusing on various QA-style tasks. Methods such as PDFTriage (Saad-Falcon et al., 2023) enables models to retrieve the context from long and structured documents. TAT-LLM (Zhu et al., 2024) addresses QA over a hybrid of tabular and textual data. ChartQA (Masry et al., 2022) poses questions over a chart image, while Chocolate (Huang et al., 2024b) annotates factual errors in machine-generated chart captions.

Recently, MMLongBench (Ma et al., 2024) and DocBench (Zou et al., 2024) evaluate multimodal document understanding, but mainly feature questions with short or extractive answers. In contrast, we consider longer, open-ended answers which require more thorough understanding and analysis of the documents. Furthermore, DocBench showed that multimodal models such as GPT-4o perform worse than text-only GPT-4, indicating the benchmark may be less suitable for multimodal evaluation. Compared to the datasets above, M-LongDoc focuses on longer documents and open-ended questions which require in-depth explanations. We further develop an automated and reliable evaluation framework which demonstrates very high agreement with human preferences.

Retrieval-Augmented Generation Processing multimodal documents with large models often poses high computational costs due to long documents and multiple high-resolution images. Therefore, retrieval augmented generation (Lewis et al., 2020; Chen et al., 2022) is a promising and practical paradigm. In this work, we have investigated multiple retrieval methods optimized for document page retrieval (Xiao et al., 2024; Robertson and Zaragoza, 2009; Chen et al., 2024a; Faysse et al., 2024). However, existing multimodal models are still constrained by multimodal biases (Chen et al., 2024b) and distraction by irrelevant content (Shi et al., 2023). To this end, our retrieval-aware tuning framework enables models to more effectively identify and leverage relevant content.

In addition, concurrent works in multimodal retrieval and question answering include VisRAG (Yu et al., 2025) and SV-RAG (Chen et al., 2025). Compared to our work which focuses on enhancing the multimodal understanding stage based on our preliminary study, VisRAG focuses on the retrieval stage. On the other hand, SV-RAG focuses on short-form question answering over multimodal documents, while we focus on more complex and open-ended questions which require longer explanations. Thus, we believe that both works are orthogonal and complementary to this work.

Lastly, MM-NIAH (Wang et al., 2024b) similarly evaluates models on long multimodal documents. However, their data construction relies on synthetic samples which are concatenated from unrelated image-text sequences. In contrast, our benchmark is more realistic as it uses only real-world long documents such as financial reports, product manuals, and academic papers that naturally reach hundreds of pages.

7 Conclusion

In this work, we present M-LongDoc, a benchmark dataset and automated framework for evaluating long-form multimodal document understanding. Designed to handle diverse types of document, including text, figures, and tables, M-LongDoc requires open-ended explanations with reasoning rather than short or extractive answers.

To address the challenges posed by lengthy and complex documents, we propose a retrieval-aware tuning approach that guides models to focus on the most relevant content. Our experiments show a substantial 4.6% improvement in model correctness over the baseline, highlighting the effectiveness of our method. These findings demonstrate the potential of our approach to support real-world applications that require deep comprehension and reasoning across multimodal documents.

Limitations

While our benchmark advances the evaluation of long multimodal document understanding, it is important to acknowledge its limitations. These fall into two main categories: data construction and scope considerations.

Regarding data construction, our approach relies on model-generated questions to achieve scalability. To address potential quality concerns, we have implemented rigorous verification processes and human quality filtering. Our detailed analysis in Appendix A.8 demonstrates that the questions effectively cover realistic scenarios and diverse aspects, including analytical reasoning, technical analysis, and domain knowledge. Additionally, the dataset's scale aligns with established benchmarks like MM-LongBench (Ma et al., 2024) and DocBench (Zou et al., 2024), ensuring sufficient depth for meaningful evaluation.

In terms of scope, we have deliberately focused on questions requiring single-page evidence rather than multi-page evidence. This design choice enables more reliable data annotation while still providing a challenging and controlled setting for model evaluation. For instance, we are able to pinpoint the multimodal biases and weaknesses of current models in our preliminary study. While we recognize that real-world scenarios often involve multi-page reasoning, this constraint allows us to provide a more controlled and high-quality benchmark. We plan to expand the evaluation to cover more scenarios in future extensions of this work.

Ethical Considerations

We will release the benchmark and training dataset publicly to facilitate further research in this area. To observe copyright rules, we do not release the documents directly, but instead the links to download each document. All annotators in this work were volunteers. While we focus on how models may answer questions based on multimodal documents, it is still possible for them to hallucinate information that is false or not verifiable. In this work, we have included the details of our training framework and hyperparameters in Section 3 and 4. As discussed above, our benchmark dataset and questions will be released under a public licence. For reproducibility, our code will be found at `https://anonymous.4open.science/r/private-multimodal-documents-B2CF/`.

Acknowledgment

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. (AISG Award No: AISG-NMLP-2024-005).

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, AmÃl'lie HÃl'liou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, TimothÃl'e Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12B. *arXiv preprint*. ArXiv:2410.07073 [cs].

Open AI. 2024. Hello gpt-4o.

Anthropic. 2024. Claude 3.5 sonnet.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, and Tong Sun. 2025. SV-RAG: LoRA-contextualizing adaptation of MLLMs for long document understanding. In *The Thirteenth International Conference on Learning Representations*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Meiqi Chen, Yixin Cao, Yan Zhang, and Chaochao Lu. 2024b. Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective. *Preprint*, arXiv:2403.18346.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema N Moussa, Matthew I. Beane, Ting-Hao 'Kenneth' Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking MLLMs in long context. In *First Conference on Language Modeling*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12039–12050, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, C'eline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *ArXiv*, abs/2407.01449.

Masato Fujitake. 2024. LayoutLLM: Large language model instruction tuning for visually rich document understanding. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10219–10224, Torino, Italia. ELRA and ICCL.

Gemini Team Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kung-Hsiang Huang, Hou Pong Chan, Yi Ren Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024a. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *CoRR*, abs/2403.12027.

Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. 2024b. Do LVLMs understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 730–749, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *TACL*, 6:317–328.

Hugo LaurenÃğon, Lucile Saulnier, LÃľo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Preprint*, arXiv:2306.16527.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023. Otter: A multi-modal model with in-context instruction tuning. *Preprint*, arXiv:2305.03726.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. 2024. Visualwebbench: How far have multimodal LLMs evolved in web page understanding and grounding? In *First Conference on Language Modeling*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiao wen Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *ArXiv*, abs/2407.01523.

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

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

OpenAI. 2023. Gpt-4v(ision) system card.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743âĂŞ3751, New York, NY, USA. Association for Computing Machinery.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, and Chris Tanner. 2024. Docfinqa: A long-context financial reasoning dataset. In *Proceedings of ACL*.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Ryan A Rossi, and Franck Dernoncourt. 2023. Pdftriage: question answering over long, structured documents. *arXiv preprint arXiv:2309.08872*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13636–13645.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, Xizhou Zhu, Ping Luo, Yu Qiao, Jifeng Dai, Wenqi Shao, and Wenhai Wang. 2024b. Needle in a multimodal haystack. In *Advances in Neural Information Processing Systems*, volume 37, pages 20540–20565. Curran Associates, Inc.

Han Xiao, Georgios Mastrapas, and Bo Wang. 2024. Jina CLIP: Your CLIP model is also your text retriever. In *Multi-modal Foundation Model meets Embodied AI Workshop @ ICML2024*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

Ruochen Zhao, Wenxuan Zhang, Yew Ken Chia, Deli Zhao, and Lidong Bing. 2024. Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions. *Preprint*, arXiv:2405.20267.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. 2024. Tat-llm: A specialized language model for discrete reasoning over tabular and textual data. *arXiv preprint arXiv:2401.13223*.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong

Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *Preprint*, arXiv:2407.10701.

A Appendix

A.1 Data Details

The detailed statistics of our dataset are shown in Figure 4. For question generation, we include the model prompt below:

Document context: {context}

Target content: {content}

Instruction: Based on the target content in this document, can you generate a test question? Ensure that the question is challenging and the answer cannot be simply copied from the content. Output the question only.

A.2 Data Verification

To verify each question in our data construction process, we use the following guide to prompt the question generator models for automated verification. Similarly, we use the same guide for human annotation in the human verification stage.

Based on the document content and question, answer yes or no only to the following questions:

1. Does the content contain any {category}? 2. Does the question require information from the {category}? 3. Is the question clear and answerable based on the {category}? 4. Is the question of reasonable difficulty and answer cannot be simply copied?

Where {category} refers to table or figure or text, which is denoted with the question.

Note: If questions require general knowledge or commonsense in addition to the content, it is still acceptable. In the document PDF file, each question is shown with the ID corresponding to excel sheet, and the document page as image In the excel sheet, indicate âĂIJyesâĂİ or âĂIJnoâĂİ for each check.

A.3 Retrieval Methods

To support our retrieval-based document question answering setting, we currently include four state-of-the art methods to retrieve relevant pages based on each question. They include text-based sparse methods such as BM25 (Robertson and Zaragoza, 2009) embedding-based methods such as BGE-M3 (Chen et al., 2024a), multimodal piece-wise embedding methods such as JINA-CLIP (Xiao et al., 2024), and multimodal page-wise embedding methods such as ColPali (Faysse et al., 2024). Note that piece-wise embedding methods separate encode each piece of text, table image, or figure image, whereas page-wise methods can encode the entire page content as a single image. Thus, we rank each page in the document based on the similarity score or relevance score of that page with respect to the given question. As each page may have multiple pieces of content, we consider the highest score of all pieces in a page to be the page-wise relevance score. To compare the effectiveness of each method, we implement a standardized MRR score which refers to the mean reciprocal rank of the gold evidence page for each question. Based on the results in Table 5, we find that ColPali which encodes each page as single image shows the best performance. Thus, we select ColPali as the preferred retrieval method in our main experiments.

A.4 Evaluation Guide

To evaluate each model answer, we use the following scoring guide. Similarly, we use the same guide for human annotation in our analysis.

You will be given one response to a question based on a multimodal document containing texts, figures, or tables. Your task is to rate the response on correctness using a 1-5 scale. Please read and understand these instructions carefully, and keep them open for reference while reviewing.

Correctness (1-5) refers to how accurately, comprehensively, and appropriately the response addresses the question based on the information in the document.

5 - Fully Correct: Completely accurate, comprehensive, fully integrates relevant information from all parts of the document, and provides a coherent answer.

4 - Mostly Correct: Largely accurate with only minor errors or omissions, addresses most main points, and integrates information well.

3 - Partially Correct: Contains a mix of accurate and inaccurate information, addresses some key points but misses others, and partially integrates information.

2 - Mostly Incorrect: Has multiple inaccuracies, addresses only a small portion correctly, and shows minimal integration of information.

1 - Completely Incorrect: Contains significant errors, is irrelevant, or fails to address the question based on the document.

Evaluation Steps: 1. Thoroughly review the multimodal document and question. 2. Carefully read the response, comparing it to the document

	Academic Paper	Product Manuals	Financial Report	All
Documents	60	60	60	180
Questions	311	279	261	851
Text-based questions	95	95	81	271
Figure-based questions	114	93	76	283
Table-based questions	102	91	104	297
Average pages per document	201.2	277.8	153.4	210.8
Average text tokens per document	114,129.8	109,745.0	139,089.3	120,988.0
Average figure images per document	90.8	368.3	24.1	161.13
Average table images per document	34.9	96.6	83.8	71.8

Table 4: Benchmark dataset statistics with respect to each domain.

Retriever	Text	Figure	Table	All
BM25	56.2	31.2	42.0	43.1
CLIP	57.1	37.9	50.4	48.5
BGE-M3	66.4	36.4	53.6	52.1
ColPali	68.7	67.5	65.9	67.4

Table 5: Retriever performance comparison.

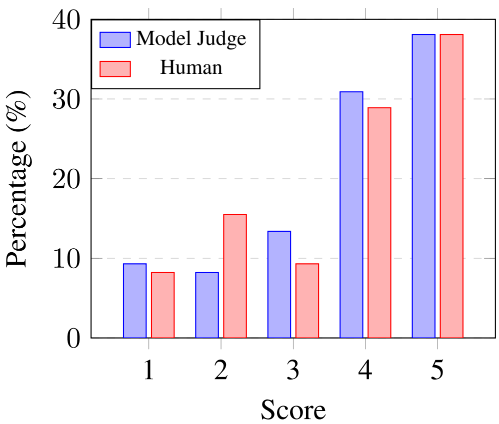


Figure 7: Distribution of evaluation scores by model judges and human annotators.

information. 3. Assess the response's accuracy, comprehensiveness, and relevance. 4. Assign a correctness score from 1 to 5 based on the criteria.

Question: {question} Response: {answer}

Evaluation Form (score only without explanation) Correctness:

A.4.1 Comparison of Model Judge and Human Scores

To further investigate the alignment between model judges and human annotators, we report the distribution of evaluation scores in Figure 7. The results show that while there are slight variations, the humans and models show very similar score distributions.

A.5 Training Data Analysis

For our generated training corpus, we omit the human verification stage for scalability and cost-efficiency, as majority of the automatically verified samples in Section 2.2 also satisfied human verification. To assess the quality of the generated responses, we evaluated using our automated scoring framework on a random subset of 100 samples. We observed an average correctness score of 4.82, indicating very high quality of answers in the training data. To avoid data leakage, we ensure that the documents used to construct the training corpus do not overlap with the evaluation set. For example, we ensure that the training documents are from different companies and products, and are published in an earlier time period.

A.6 Hyperparameters

We use a single A800 GPU for training experiments. Unfortunately, due to computational constraints, we are unable to investigate larger open-source models at this time. For all model evaluation, we use greedy decoding with temperature $T = 0$ to reduce variance. In our training framework, we set the number of training epochs to be 1, batch size as 16, and learning rate as 1e-4. To reduce the training cost due to limited computational resources, we leverage LoRA (Hu et al., 2022) training with rank as 64 and alpha as 32.

A.7 Robustness to Increasing Retrieval Context

Table 8 shows the performance of Qwen2-VL, both with and without retrieval-aware tuning, when using varying numbers of retrieved pages. A key advantage of our tuned model is its consistent performance even when processing a large number of retrieved pages. In contrast, the original model suffers a noticeable performance drop. This high-

Category	Description	Proportion	Example Question
Analytical Reasoning	Questions about trends, comparisons, and implications (e.g., engagement trends, performance trends)	49%	What is the total amount of financial liabilities at amortized cost for the year 2023, and how does it compare to the total amount for 2022? Consider the implications of any changes in these liabilities on the company's financial strategy.
Technical Analysis	Questions about specific technical details (e.g., UEFI BIOS, shutter speeds, X-sync speeds) and applications of technical concepts.	37%	What potential issue could arise if you fail to follow the instruction to tighten the screws twice when installing the top cover, and why might this step be particularly important for a laptop?
Commonsense and Domain Knowledge	Questions requiring general knowledge or background knowledge in fields such as finance, cybersecurity, photography.	46%	What are the key differences and potential advantages of using white-box analysis over machine learning for modeling the performance of configurable systems, as discussed by Velez et al. (2021)?
Visual Interpretation	Questions based on interpreting icons, diagrams, or charts.	60%	Explain the functionalities of the different sections (a, b, c, d) in the LaserFactory design toolbar and discuss how each section contributes to the overall design and fabrication process.
Mathematical Reasoning	Questions involving mathematical concepts or calculation from data.	17%	Calculate the percentage change in diluted net income per share attributable to common stockholders from fiscal year 2023 to fiscal year 2024. What factors likely contributed to this change?

Table 6: Categorization of question types with descriptions and examples.

Dataset	Question Length	Answer Length
DocVQA	8.5	2.4
MMLongBench	16.4	2.6
M-LongDoc (Ours)	31.6	180.3

Table 7: Comparison of average question length and answer length in tokens for different datasets.

lights the improved robustness of our approach in handling increasing amounts of retrieval context.

Table 8: The performance of our retrieval-aware tuning method under varying numbers of retrieved pages.

	Text	Figure	Table
Qwen2-VL			
w/ top k = 1 pages	4.05	3.25	3.36
w/ top k = 5 pages	4.17	3.67	3.46
w/ top k = 10 pages	4.08	3.62	3.19
w/ top k = 20 pages	OOM	OOM	OOM
Qwen2-VL w/ Retri. Tuning			
w/ top k = 1 pages	4.13	3.74	3.47
w/ top k = 5 pages	4.37	3.83	3.80
w/ top k = 10 pages	4.33	3.84	3.79
w/ top k = 20 pages	OOM	OOM	OOM

A.8 More examples

A.8.1 Example of M-LongDoc

To investigate the diversity and in-depth nature of questions in our dataset, we manually categorize 100 random examples based on five common types, as shown in Table 6. Furthermore, Table 9 illustrates an example of a challenging question in our M-LongDoc benchmark. This question tests the ability of the model to identify and analyze trends across different charts and draw meaningful comparisons. To further compare with existing datasets, we measure the average question lengths and answer lengths as shown in Table 7.

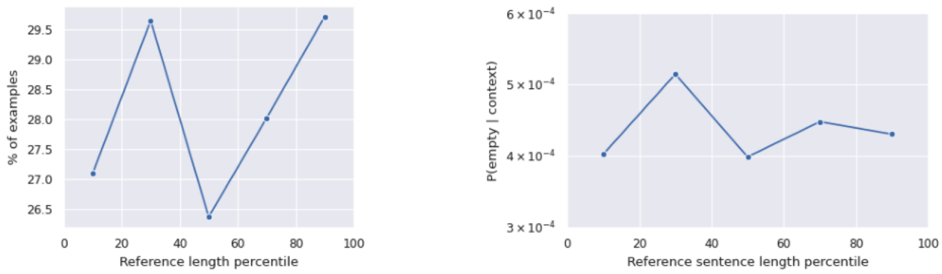
A.8.2 Case study of Retrieval-aware Tuning

Table 10 displays a sample question in M-LongDoc and the answers generated by Qwen2-VL and Qwen2-VL w/ Retrieval-aware Tuning. The answer generated by Qwen2-VL states that the Cosine method consistently shows the highest latent cosine similarity across all datasets, which is incorrect. In fact, the zero-shot stitching experiment does not involve the Cosine method. It appears that Qwen2-VL may have been misled by the keyword "cosine" appearing elsewhere in the retrieved context. In contrast, the answer generated by Qwen2-VL w/

Retrieval-aware Tuning correctly identifies that the affine method consistently obtains the highest latent cosine similarity (lcos) across all datasets. This demonstrates the effectiveness of our Retrieval-aware Tuning method in improving the model's capability to comprehend retrieved context.

Question:
How does the relationship between reference length percentile and the percentage of empty modes differ from the relationship between reference sentence length percentile and the probability of empty context? Explain the key differences in the trends shown by these two graphs.

Relevant page (truncated):



(a) Percent of stories that have the empty sequence as their modal continuation.

(b) Geometric mean of the model's probability of the empty sequence given the first four sentences of the story.

Figure 6.2: Finetuned GPT-2-345M predictions of empty outputs on the ROC Stories validation set (1571 Stories). Stories are grouped into 5 equally sized bins by reference continuation length.

Table 6.2: Modal continuations of several lengths for prefix: "Sarah always had a fascination with the night sky. Noticing her passion, Sarah's father bought her a new telescope. She was ecstatic. She went outside every night to diligently view the night sky." The reference continuation is "Sarah loved her new telescope."

Length Constraint (tokens)	Log-probability	Text
Global mode	-7.79	`<|endoftext|>`
5	-9.14	Sarah loved astronomy!`<|endoftext|>`
6	-7.97	Sarah never looked back.`<|endoftext|>`
7	-8.59	Sarah loved her new telescope.`<|endoftext|>`
8	-9.38	Now, Sarah is an astronomer.`<|endoftext|>`
9	-8.68	Sarah was happy with her new telescope.`<|endoftext|>`
10	-8.77	Sarah was very happy with her new telescope.`<|endoftext|>`
12	-8.91	Sarah was amazed by the beauty of the night sky.`<|endoftext|>`

An interesting feature of these constrained modes is that the content can be correlated with the length in clear ways. Table 6.3 shows an example where the mode of length 5 is significantly different from all the other modes. It may be impossible to produce a 5 token output that has the right content, but the model "prefers" to output something grammatical, so we see different content. This is different from the short NMT modes, which were often truncated when the constraint was too short to express the content of the source sentence.

In order to show that these patterns aren't just cherry-picked, randomly sampled examples of modal outputs are shown in Table B.2. All 30 of the conditional modes are grammatical, relevant to the context, and don't show any evidence of degenerate behavior. This is further evidence that conditional MAP inference may be a promising direction of investigation.

Table 9: An example of a challenging question from M-LongDoc that requires the model to compare the trends of two charts in a document.

Question:
Which method consistently shows the highest latent cosine similarity (lcos) across all datasets (MNIST, F-MNIST, CIFAR-10, CIFAR-100) in the zero-shot stitching experiment, and what might this imply about the nature of transformations between latent spaces in autoencoders?

Answer generated by Qwen2-VL:
The method that consistently shows the highest latent cosine similarity (lcos) across all datasets (MNIST, F-MNIST, CIFAR-10, CIFAR-100) in the zero-shot stitching experiment is the "Cosine" method. This implies that the "Cosine" method is able to capture the most similar latent spaces across different architectures, modalities, and datasets. This suggests that the nature of transformations between latent spaces in autoencoders is not exclusively correlated by orthogonal transformations, but rather by a class of transformations larger than orthogonal transformations. This finding highlights the importance of understanding and modeling the particular class of transformations that control the correlation between NNs during image autoencoding.

Answer generated by Qwen2-VL w/ Retrieval-aware Tuning:
The method that consistently shows the highest latent cosine similarity (lcos) across all datasets (MNIST, F-MNIST, CIFAR-10, CIFAR-100) in the zero-shot stitching experiment is the "affine" method. This method achieves the highest lcos values for all datasets, with scores of 0.94 for MNIST, 0.94 for F-MNIST, 0.96 for CIFAR-10, and 0.96 for CIFAR-100. This consistency in high lcos values suggests that the affine transformation is particularly effective in aligning the latent spaces of different autoencoders. The high lcos indicates a strong alignment between the latent representations of the autoencoders, which implies that the affine transformation is capable of capturing the underlying structure and relationships between the latent spaces more accurately than other methods. The results imply that the nature of transformations between latent spaces in autoencoders is not exclusively controlled by orthogonal transformations. The affine transformation, which includes scaling, rotation, and translation, appears to be a more versatile and effective method for aligning these spaces. This finding suggests that the latent spaces of image autoencoders are related by a broader class of transformations, including those that are not strictly orthogonal, which is a significant insight into the nature of these transformations and their implications for model stitching and zero-shot learning.

Relevant page (truncated):

5.3. Latent Communication via Translation 39

TABLE 5.3: Zero-shot stitching for generation with various methods for estimating $\widehat{\mathcal{T}}$. The representation is normalized using Standard Scaling, and no additional normalization is applied to the stitched decoders. We report the latent cosine similarity (*lcos*) and MSE (*lmse*) between the target encoding and the translated one, but also the reconstruction MSE (*rmse*) between the input and the output. The absolute space dimension is 500, and we used 1000 anchors. Please refer to Table B.1 for results on decoders scale-invariant by design (with L2 normalization on the encodings).

	MNIST			F-MNIST			CIFAR-10			CIFAR-100		
	lcos	*lmse*	*rmse*	*lcos*	*lmse*	*rmse*	*lcos*	*lmse*	*rmse*	*lcos*	*lmse*	*rmse*
`absolute`	0.09	0.27	0.14	0.17	0.23	0.23	0.30	0.29	0.34	0.34	0.53	0.40
`affine`	0.94	0.08	0.02	0.94	0.06	0.03	0.96	0.03	0.05	0.96	0.04	0.05
`linear`	0.92	0.09	0.02	0.93	0.07	0.04	0.94	0.03	0.05	0.94	0.04	0.06
`l-ortho`	0.79	0.14	0.02	0.78	0.12	0.05	0.85	0.05	0.06	0.84	0.07	0.07
`ortho`	0.90	0.10	0.02	0.90	0.08	0.04	0.94	0.03	0.06	0.93	0.04	0.06

in structure, differing only in the random seed used for weight initialization and data shuffling. To perform Zero-Shot Stitching, we first translate each data point from the latent space of the first encoder to the latent space of the second using 1000 parallel anchors. We then apply the second decoder to the translated data, without any additional training or fine-tuning.

Result analysis. This experiment analyzes the alignment of latent spaces in different training regimens of the same AE. The performance evaluation, as shown in Table 5.3, demonstrates that all methods `affine`, `linear`, `l-ortho`, and `ortho` yield satisfactory results. Moreover, qualitative results depicted in Figure 5.6 reveals minimal visual differences in the stitching outcomes across various datasets using different methods. Please refer to Figures B.4 and B.5 for other qualitative results. In fact, these

Table 10: Sample answers generated by Qwen2-VL and Qwen2-VL w/ Retrieval-aware Tuning, respectively.