STRUCTURED ATTENTION MATTERS TO MULTIMODAL LLMS IN DOCUMENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Document understanding remains a significant challenge for multimodal large language models (MLLMs). While previous research has primarily focused on locating evidence pages through precise multimodal queries, our work investigates a fundamental yet overlooked aspect: how input format influences document comprehension performance. Through systematic analysis, we discover that plain multi-element text extracted from PDFs often impairs rather than improves MLLMs' performance, a counterintuitive finding that we attribute to attention dispersion and loss of structure. To further substantiate our hypothesis, we propose using the LaTeX paradigm as a tool for encoding document elements, maintaining the hierarchical organization and spatial relationships critical for comprehension. Our attention analysis reveals that structured multi-element text induces structured attention patterns in both textual and visual content, directing models to focus on semantically meaningful regions while reducing attention waste. Specifically, we found that the structured text significantly enhance MLLMs' document question-answering performance across diverse document types without requiring architectural modifications or additional training.

1 INTRODUCTION

Document understanding is a common task in daily life. With the rapid development of multimodal large language models (MLLMs), capable general-purpose systems are increasingly expected to comprehend documents effectively (Zhu et al., 2023; Liu et al., 2024a; 2023; Wang et al., 2024a; Bai et al., 2023; Team, 2025; Anthropic, 2023; Dong et al., 2024; Chen et al., 2024b; Li et al., 2023; Wang et al., 2024b). The difficulty in document question answering arises from the diverse nature of questions and the variety of information required, which evidence elements include text blocks, charts, diagrams, and figures, often requiring integration of multiple information sources. Document understanding through MLLMs presents three key challenges: (1) the information diversity challenge of processing heterogeneous elements, (2) the context integration challenge of synthesizing scattered information, and (3) the structural relationship challenge of understanding how elements relate spatially and logically—relationships intuitive for humans but difficult for machines (Han et al., 2025).

Previous research has addressed these challenges primarily by extending context windows to accommodate more content or by developing specialized architectures for extracting multi-granularity information (Ding et al., 2022; Hu et al., 2024; Ye et al., 2023; Liu et al., 2024b; Park et al., 2024; Tito et al., 2023). With general-purpose MLLMs, the trend has shifted toward retrieval-augmented generation (RAG), which locates relevant evidence and feeds it into models as images, text, or both (Mishra et al., 2019; Suri et al., 2024; Park et al., 2024). Despite these advances, a critical question remains unexplored: how does the format of input information, rather than merely its content—influence document understanding? Current approaches typically extract text but discard critical structural information (Khattab & Zaharia, 2020; Faysse et al., 2024). This leads to a counterintuitive phenomenon we observed across multiple datasets: unstructured plain multi-element text extracted from PDFs often degrades rather than enhances MLLM performance compared to using images alone (Cho et al., 2024; Deng et al., 2024; Han et al., 2025; Zhang et al., 2024a). This observation motivated us to investigate the relationship between input structure and model comprehension. We discovered that information structure fundamentally shapes how MLLMs allocate

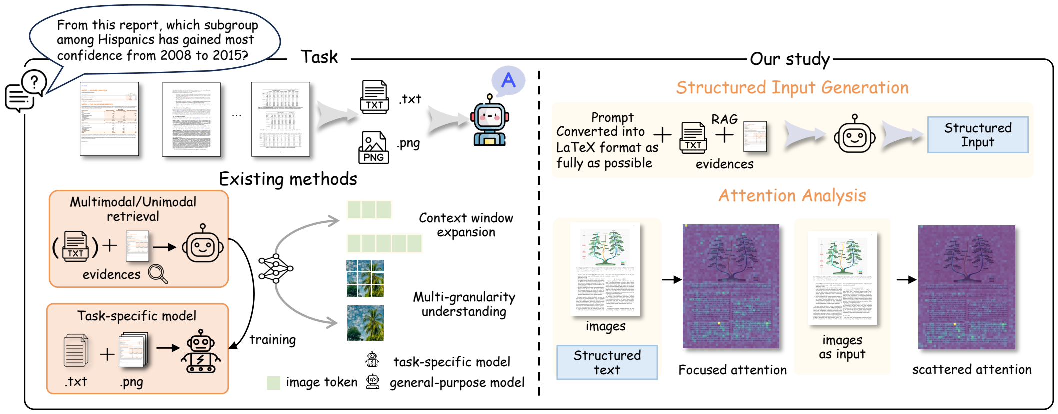

Figure 1: Comparison of Different Approaches for DocQA: Previous research methods focused on using RAG to precisely locate the evidence and then directly input the evidence into general-purpose MLLMs, or on designing task-specific models that focus on multi-granularity extraction of image information and expanding the context window. We propose a novel structure-preserving approach based on the LaTeX paradigm to explore the impact of input format on general models' responses to DocQA and investigate potential causes through attention analysis.

attention across document elements. With unstructured text, models exhibit scattered attention patterns, wasting computational resources and struggling to identify relationships between elements.

Our study includes two main steps. First, we propose an approach to preserve the structure of plain multi-element text extracted from PDFs. We input the evidence plain text and images into an MLLM and prompt it to generate structured text. This step aims to maintain the structure of diagrams, tables, and texts. The layout information can be preserved in the form of text, which is helpful during the answer generation process. We input the images along with the LaTeX paradigm to generate answers, exploring the importance of document layout and the structure of sub-elements on the evidence page. We find that the accuracy of MLLMs in answering questions is significantly improved. Second, we analyze the attention distribution and transformation with different inputs, comparing the cases of using images alone and images combined with structured text as input. We find that attention is more focused when constrained by the structured input.

Overall, our findings show that the ability of general-purpose models in document understanding can be improved simply by changing the input format. Furthermore, the attention transformation between different inputs demonstrates that the structured attention brought by structured input is key to improving the ability of MLLMs to answer DocQAs. This indicates that there are other aspects worth exploring beyond focusing solely on enhancing the effectiveness of Retrieval Augmented Generation (RAG).

2 Related Work

We discuss two lines of related work: MLLMs in DocQA tasks and methods used in DocQA tasks.

MLLMs in DocQA Tasks. Recent Document Visual Question Answering (DocVQA) models focus on expanding context window sizes, improving fine-grained visual comprehension, and enhancing layout analysis (Hu et al., 2024; Liu et al., 2024b; Tworkowski et al., 2023). These systems often boost performance through architectural enhancements (e.g., added modules) or multi-stage training pipelines with diverse data inputs (Park et al., 2024; Tito et al., 2023; Chen et al., 2023). While general multimodal models like Gemini (Team et al., 2024) and Qwen-VL (Bai et al., 2023) show improved visual processing and context handling, they remain constrained by input length limitations and multi-page processing capabilities. Their reliance on image-derived representations also hinders precise localization of detailed elements, while layout extraction from visual data risks diverting the model's focus, reducing answer accuracy (Han et al., 2025).

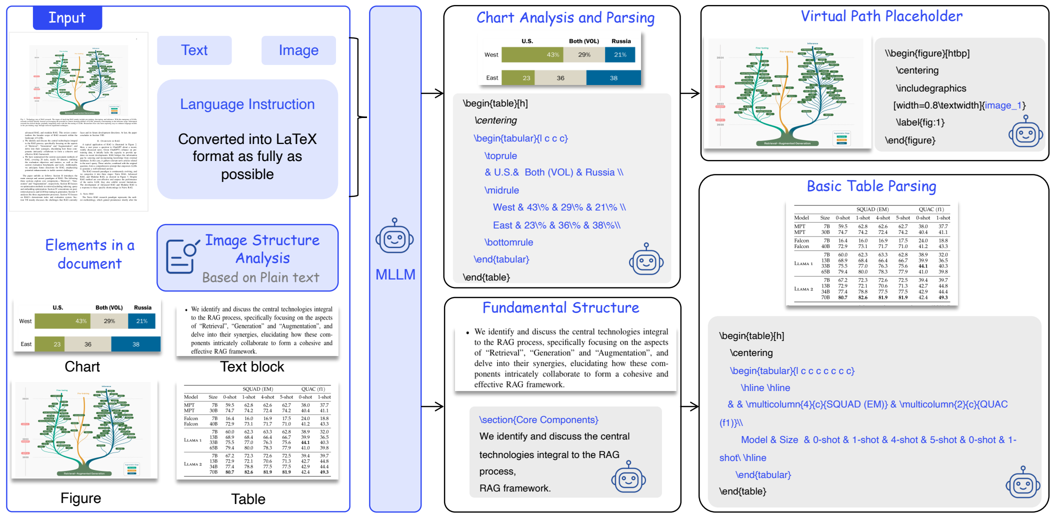

Figure 2: Structured text is generated using LATEX. We prompt the MLLM to capture the layout of the given images as accurately as possible, producing blocks that include text, charts, and tables. Figures that cannot be parsed into concrete content are represented using virtual paths and the LATEX paradigm. This approach is simple and only requires API-level access with instruction-level control.

Methods in DocQA Tasks. Retrieval augmentation has become pivotal for adapting general multimodal models to document question answering, addressing input constraints and information overload by retrieving fine-grained evidence through multimodal or unimodal queries (Lewis et al., 2020; Gao et al., 2023; Chen et al., 2024a; Cho et al., 2024). While current methods enhance accuracy through optimized retrieval strategies (Deng et al., 2024; Lu et al., 2024), input formatting remains challenging: image-only inputs hinder textual comprehension due to incomplete integration, whereas plain multi-element text extracted from PDFs neglect visual elements critical for image-dependent questions. Multimodal inputs combining text and images risk overwhelming models with redundant or conflicting data, leading to attention dispersion. To maximize performance, refining input strategies—such as modality prioritization, structured evidence fusion, or dynamic filtering—is essential for balancing information richness and focused reasoning in retrieved evidence.

3 METHODOLOGY

Understanding documents through multimodal large language models (MLLMs) represents a complex challenge at the intersection of vision and language processing. Other research mainly investigates the importance of locating precise evidence pages that are directly input into MLLMs, while often ignoring how the input representation influences the performance in document question answering.

In this section, we propose using the LATEX paradigm as a tool for retaining document structure. This is a novel structure-preserving approach that encodes multiple elements in documents, which are then transferred into structured multi-element text. Images combined with structured text significantly enhance the MLLMs' document understanding capabilities without requiring architectural modifications or additional training. We analyze the attention transformation and question answering ability under different inputs, revealing that structured multi-element text induces structured attention patterns on both textual and visual content, collectively enabling the comprehension of MLLMs.

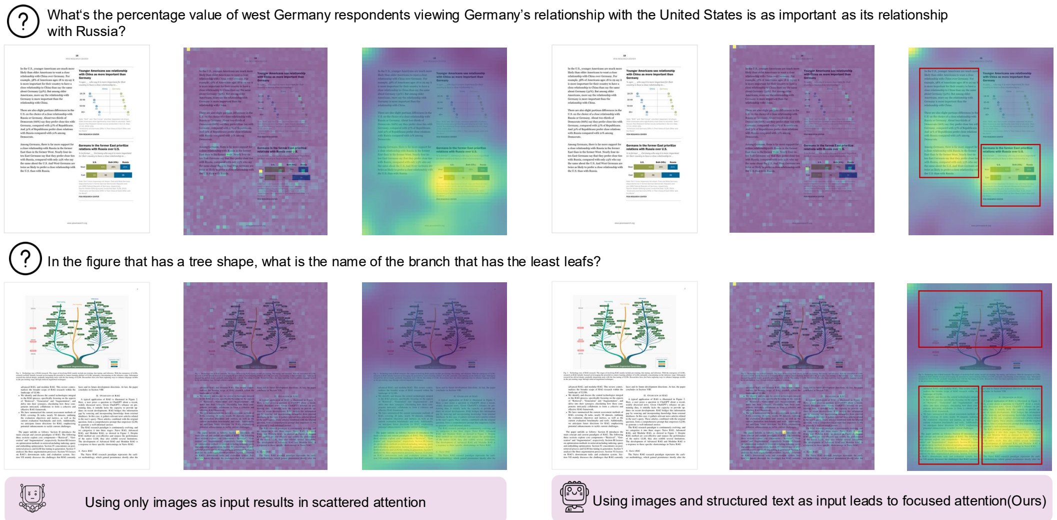

Figure 3: An example of attention transformation under two conditions: images alone versus images combined with structured text. The MLLMs tend to focus on text blocks, figures, and charts, rather than the overall layout.

3.1 Input Representation in Document Understanding

Document understanding requires models to interpret diverse information types—text blocks, tables, figures, charts, and their interrelationships. Through systematic analysis of MLLM performance in document question answering (DocQA) tasks, we identified a critical paradox.

Observation 1. *Providing unstructured multi-element text alongside document images often degrades MLLM performance across benchmark datasets, despite increasing the total information available to the model.*

This counterintuitive finding contradicts the common assumption in retrieval-augmented generation that more textual context improves performance. To investigate this phenomenon, we conducted a series of controlled experiments comparing MLLMs' attention patterns when processing: (1) document images alone, (2) images with unstructured plain multi-element text, and (3) images with our proposed structured text representation.

3.2 Plain Multi-Element Text with Structure

Based on observation 1, we propose an approach to preserve the structure of the plain multi-element text extracted from PDFs. As shown in Figure 2, we use LaTeX to encapsulate the plain text. It is easy for MLLMs to understand LaTeX code because it usually follows a fixed and easy-to-understand paradigm. The LaTeX code constrains the content in titles, tables, and figures, providing efficient and additional references when MLLMs answer DocQAs with images. We input the image and the corresponding plain text into the MLLM to obtain the structured text. We prompt the MLLM to preserve the structure and text related to charts in the image as much as possible. If a figure in the image cannot be converted into LaTeX content, we instruct the MLLM to use a virtual path to represent the figure while keeping the structure intact. An example is provided in Figure 2.

We use the proposed method to obtain structured text and compare the performance of MLLMs under three cases: images only, images with plain text, and images with structured text. The results confirm the hypothesis in Section 3.1, leading to the following observation:

Observation 2. *Structured text and images work together to improve MLLMs' performance in answering DocQAs.*

Observation 2 shows that the performance of MLLMs are improved significantly simply by preserving the structure of the input text, without the need to provide additional fine-grained positional information to help answer the questions. This observation proves the importance of structure, which means that merely informing the MLLMs of the structure of document helps them gain a better overview and answer questions correctly.

3.3 ATTENTION TRANSFORMATION WITH STRUCTURED TEXT

Given observation 2, this subsection seeks to uncover the underlying reasons why structured text matters by analyzing the attention distribution when MLLMs answer questions using only images versus using structured text as references. Figure 4 shows that in the structured case, the attention scores are less sensitive to the boundaries of the image and more concentrated on the main body, indicating that the MLLMs know where to focus under the constraints of structured text. Based on these findings and experimental results, we have the following observation:

Observation 3. *Structured text brings structured attention to both texts and images, which directly improves the abilities of MLLMs. This shows that structured attention is the key to helping MLLMs answer DocQAs.*

Observation 3 demonstrates the necessity of structure in document understanding. As shown in the example in Figure 3, structured text constrains MLLMs and reduces irrelevant attention when focusing on images. This helps MLLMs recover distracted attention and focus on images and relevant text to answer the given question.

4 EXPERIMENTS

We conduct evaluations on four document understanding benchmarks covering multiple scenarios to provide solid evidence supporting the observations presented in this paper.

4.1 DATASETS AND MODELS

Implementation Details. We apply the document preprocessing method from Han et al. (2025) to obtain plain multi-element text using a combination of Optical Character Recognition (OCR) and PDF parsing techniques, together with visual content obtained by transforming pages in long documents into images. Specifically, OCR is employed to recognize text within image-based PDFs, while PDF parsing extracts text directly from digitally encoded text within the PDF. This dual approach ensures robust text extraction across various document formats and structures. We then locate the evidence pages as images, from which we extract the corresponding plain text. We subsequently apply the approach presented in Section 3.2 to obtain the structured text. All experiments are conducted on 4 NVIDIA A100 GPUs.

Models. Our evaluation uses the following four multimodal LLMs: QWEN2-VL-7B-INSTRUCT, QWEN2.5-VL-7B-INSTRUCT, LLAVA-V1.6-MISTRAL-7B and PHI-3.5-VISION-INSTRUCT. These models accept both images and text as input and generate answers and analyses for the given questions.

Datasets. The benchmarks include MMLongBench (Ma et al., 2024), LongDocUrl (Deng et al., 2024), PaperTab (Hui et al., 2024), and FetaTab (Hui et al., 2024). These evaluation datasets cover a variety of scenarios, including open- and closed-domain, textual and visual, as well as long and short documents, ensuring fairness and completeness in the evaluation.

Metrics. For all benchmarks, we leverage LLAMA-3.1-8B-INSTRUCT as the evaluation model to assess the consistency between the model's output and the reference answer producing a binary decision (correct / incorrect). We report the average accuracy for each benchmark.

4.2 IMPACT OF PLAIN MULTI-ELEMENT TEXT

We first compare the performance of MLLMs when using images alone versus using images with the help of plain multi-element text. According to Table 1, the results show that plain text impairs MLLMs' performance on MMLongBench, which contains multi-element evidence images including charts, tables, blocks of text, and figures. The additional text information decreases the accuracy of

Table 1: Our proposed structure-preserving method effectively retains the document structure and significantly improves the accuracy of DocQA responses. For the open-domain dataset, we followed the retrieval method used in Han et al. (2025), using the top-1 retrieval result as input. On the closed-domain dataset, we used the evidence pages of the dataset as input to test the importance of document structure.

Model	MMLongBench	LongDocUrl	PaperTab	FetaTab	Avg
Prior Multimodal LLMs					
QWEN2-VL-7B-INSTRUCT w/ image (Wang et al., 2024a)	0.292	0.195	0.160	0.441	0.272
QWEN2-VL-7B-INSTRUCT w/ image+text (Han et al., 2025)	0.287	0.219	0.188	0.507	0.300
QWEN2.5-VL-7B-INSTRUCT w/ image (Team, 2025)	0.389	0.197	0.163	0.508	0.314
QWEN2.5-VL-7B-INSTRUCT w/ image+text (Han et al., 2025)	0.375	0.399	0.229	0.557	0.427
LLAVA-V1.6-MISTRAL-7B w/ image (Liu et al., 2024a)	0.131	0.126	0.051	0.154	0.116
LLAVA-V1.6-MISTRAL-7B w/ image+text (Han et al., 2025)	0.183	0.159	0.127	0.406	0.219
PHI-3.5-VISION-INSTRUCT w/ image (Abdin et al., 2024)	0.189	0.131	0.077	0.245	0.161
PHI-3.5-VISION-INSTRUCT w/ image+text (Han et al., 2025)	0.226	0.213	0.160	0.443	0.261
Prior Multimodal LLMs + Structured text (Ours)					
QWEN2-VL-7B-INSTRUCT w/ image	**0.306**	**0.229**	**0.209**	**0.509**	**0.313**
+ **Structured Text**	+4.8 %	+17.4%	+30.6%	+15.4 %	
QWEN2.5-VL-7B-INSTRUCT w/ image	**0.435**	0.221	**0.252**	**0.575**	0.371
+ **Structured Text**	+11.8%	+12.2%	+54.6%	+13.2%	
LLAVA-V1.6-MISTRAL-7B w/ image	**0.224**	0.153	0.122	0.388	**0.222**
+ **Structured Text**	+71.0%	+21.4%	+139.2%	+151.9 %	
PHI-3.5-VISION-INSTRUCT w/ image	**0.284**	0.211	**0.224**	0.429	**0.287**
+ **Structured Text**	+50.3 %	+61.1%	+190.9%	+75.1%	

the large model on this dataset from 0.389 to 0.370. The accuracy of MLLMs on other datasets might increase, but this can also impair their ability in specific cases. Providing unstructured plain text alongside document images sometimes degrades MLLM performance across benchmark datasets, despite increasing the total information available to the model. These results and the case shown in Figure 5 clearly and convincingly support observation 1, showing that repeated and redundant text information does not help large models answer questions about long documents.

We hypothesize that the underlying cause lies in the loss of structural information during the conversion to plain text. When relevant elements in document images—such as text blocks, charts, and figures—are transformed into plain text, their original structural constraints are discarded. Consequently, MLLMs face difficulties in accurately interpreting both the individual elements and the relationships among them, which in turn results in information confusion and diminishes their effectiveness in answering document-centric questions.

4.3 PERFORMANCE IMPROVEMENTS WITH STRUCTURED TEXT

We consider the impact of structure on the process of MLLMs understanding documents. We compare using structured text as a reference with using only images for understanding. On MMLongBench, adding structured text information increases the accuracy of QWEN2.5-VL-7B-INSTRUCT from 0.389 to 0.435. This 10% performance improvement fully demonstrates the importance of structure in long document reading. The structure of different elements in the evidence images is preserved as much as possible due to the instructions given to the MLLM. The structured text surprisingly enhances the ability to answer questions. On MMlongBench, every models' performance are improved because almost all evidence pages in this dataset contain structured tables and graphs, which can be fully interpreted by the LATEX paradigm combined with plain text. Additionally, the accuracy of QWEN2.5-VL-7B-INSTRUCT is significantly higher compared to QWEN2-VL-7B-INSTRUCT, demonstrating that QWEN2.5-VL-7B-INSTRUCT has a stronger ability to understand LATEX and make better use of structure.

The corresponding data further observation 2. By employing our proposed novel structure-preserving method, which leverages an intuitive LaTeX paradigm, we are able to retain both the internal structure of individual elements within document images and the relationships among these elements. Without requiring additional training or modifications to the model architecture, this approach enhances the ability of MLLMs to comprehend documents, thereby demonstrating the critical importance of structural information.

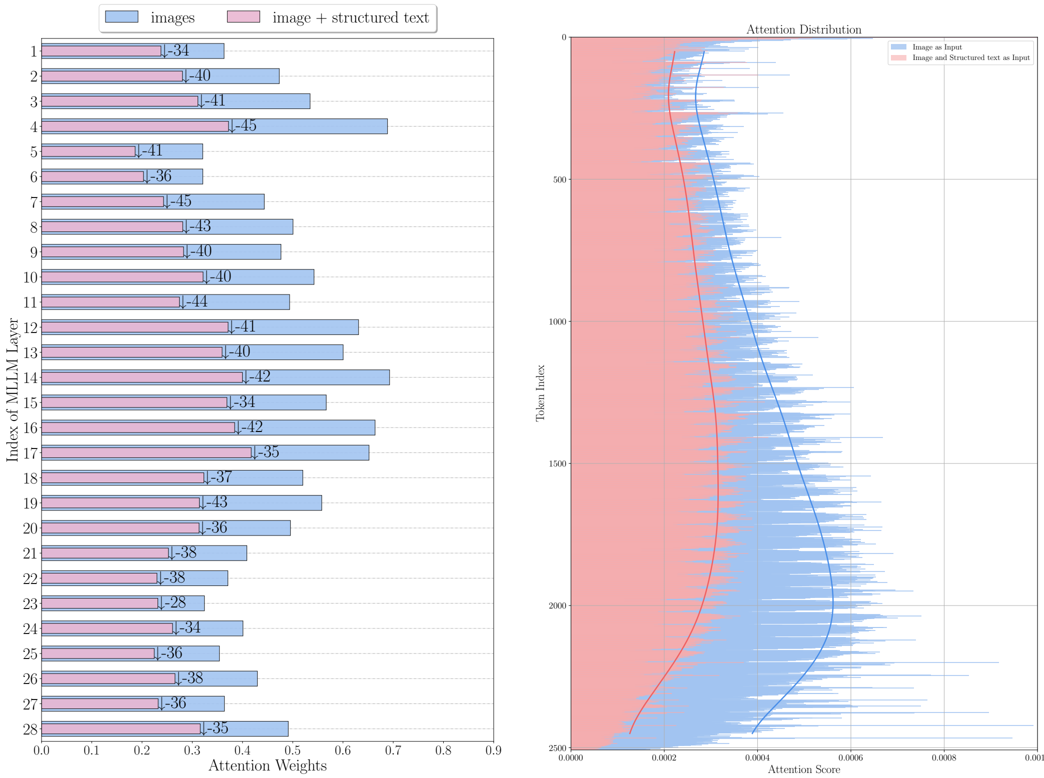

Figure 4: (a) Comparison of attention weights across different layers of MLLMs for images alone versus images combined with structured text. (b) MLLMs are less sensitive to image border tokens when constrained by structured text. The attention distribution shows that attention scores at the borders are lower in the presence of structured text.

4.4 Attention Analysis

We follow Zhang et al. (2024b) to conduct attention analysis in order to understand why structured text matters to MLLMs. We analyze attention from two perspectives: attention distribution and attention transformation across different cases. We present the following findings to support Observation 3.

Attention Distribution. We consider 370 samples with the same number of image tokens to demonstrate the distribution of attention transformation with or without structured text as references when MLLMs answer questions based on evidence images. We use Qwen2.5-VL-7B-Instruct as the main model to generate answers on MMLongBench. In the case where MLLMs rely only on images, the attention distribution shows that the model is sensitive to boundary tokens of images and exhibits an uneven distribution of attention. These results indicate that MLLMs treat almost every image token equally and have no clear focus on specific regions of the original images. In contrast, MLLMs are constrained when using structured text as references. MLLMs learn to focus on useful image tokens located in figures, blocks of text, or tables. This transformation of attention distribution highlights the importance of structured text.

Figure 4 illustrate the attention distribution of Qwen2.5-VL-7B-Instruct under different inputs, where the former shows the overall variation across 28 layers and the latter presents the attention changes within a single layer. We observe that the input of structured text induces more focused attention and reduces attention loss at the page boundaries. This indicates that structured text contributes to a dual form of structured attention—spanning both images and text—which becomes a primary factor in enhancing the document comprehension ability of MLLMs.

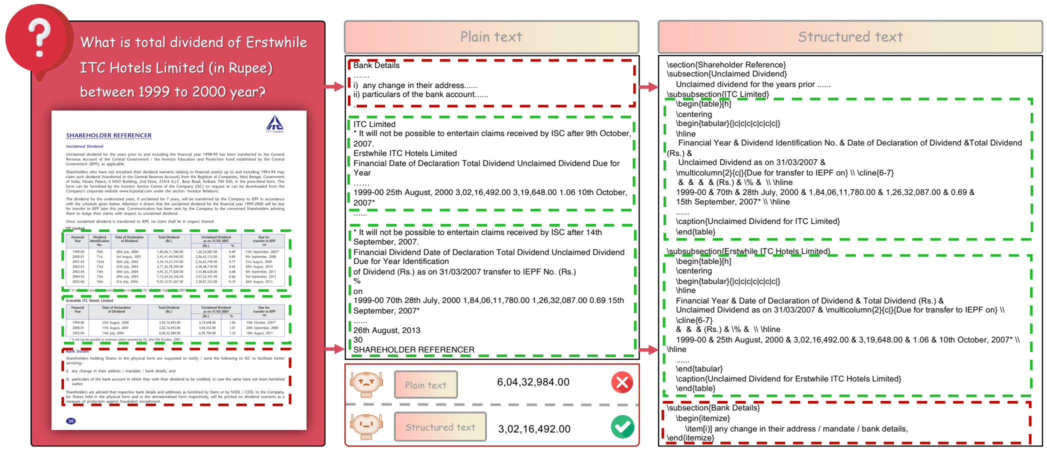

Figure 5: A comparison of generated answers from MLLMs using plain multi-element text versus structured text shows that structured text improves MLLM performance on DocQA tasks. The LaTeX paradigm helps preserve the image's structure, aiding the model in locating evidence relevant to the question.

Case Study. We perform a case study to better understand attention transformation with the help of structured text. Figure 5 illustrates an example. The question requires MLLMs to find the percentage of West Germany respondents who view Germany's relationship with the United States as equally important as its relationship with Russia, based on the chart and corresponding text. The attention map in Figure 3 shows that attention is distributed everywhere without the control of structured attention, even in the blank areas of the given image. With the constraint of structured attention, attention focuses on blocks of text, charts, and so on. We conclude that structured text helps MLLMs reduce attention loss, guides them where to look, and increases the probability of finding fine-grained evidence regions. All these constraints enable MLLMs to better understand documents and answer DocQAs more effectively.

4.5 Ablation Studies

We almost obtain the best results on almost every dataset and model by using structured text and images as input. We further conduct experiments on the following cases: structured text as input, plain text as input, and the LaTeX format acting as a placeholder without specific text information. Based on these cases, we aim to demonstrate the necessity of combining structured text and

Table 2: Performance comparison on different datasets. The ablation results show the importance of combining structured text and image. When structured text is combined with an image, it results in structured attention, which eventually helps MLLMs answer questions.

Model	**MMLongBench**	**LongDocUrl**	**Avg**
Qwen2-VL-7B-Instruct w/ text	**0.291**	**0.175**	**0.233**
Qwen2-VL-7B-Instruct w/ structured text	0.212	0.163	0.188
Qwen2.5-VL-7B-Instruct w/ text	**0.318**	**0.188**	**0.253**
Qwen2.5-VL-7B-Instruct w/ structured text	0.267	0.155	0.211
LLaVA-v1.6-Mistral-7B w/ text	**0.287**	**0.162**	**0.225**
LLaVA-v1.6-Mistral-7B w/ structured text	0.246	0.151	0.199
Phi-3.5-Vision-Instruct w/ text	**0.299**	**0.185**	**0.242**
Phi-3.5-Vision-Instruct w/ structured text	0.244	0.167	0.206

Table 3: Performance comparison across different evidence sources on MMLongBench reveals how various document elements demonstrate the differing effects of structured input on document understanding.

Model	**Chart**	**Table**	**Pure-Text**	**Generalized-text**	**Figure**	**Avg**
Prior Multimodal LLMs						
QWEN2-VL-7B-INSTRUCT w/ image	0.278	0.304	0.311	0.311	0.318	0.304
QWEN2-VL-7B-INSTRUCT w/ image+text	0.244	0.300	0.358	0.294	0.301	0.299
QWEN2.5-VL-7B-INSTRUCT w/ image	0.392	0.350	0.427	0.387	0.418	0.395
QWEN2.5-VL-7B-INSTRUCT w/ image+text	0.375	0.367	0.374	0.420	0.388	0.285
LLAVA-V1.6-MISTRAL-7B w/ image	0.120	0.060	0.116	0.126	0.134	0.111
LLAVA-V1.6-MISTRAL-7B w/ image+text	0.114	0.106	0.195	0.151	0.154	0.144
PHI-3.5-VISION-INSTRUCT w/ image	0.176	0.147	0.185	0.160	0.244	0.182
PHI-3.5-VISION-INSTRUCT w/ image+text	0.205	0.203	0.291	0.261	0.224	0.237
Prior Multimodal LLMs + Structured text (Ours)						
QWEN2-VL-7B-INSTRUCT w/ structured text	**0.307**	0.267	0.348	0.261	0.288	0.294
QWEN2.5-VL-7B-INSTRUCT w/ structured text	0.364	0.364	0.374	0.328	0.385	0.363
LLAVA-V1.6-MISTRAL-7B w/ structured text	**0.159**	**0.129**	**0.205**	**0.160**	**0.161**	**0.163**
PHI-3.5-VISION-INSTRUCT w/ structured text	**0.261**	**0.263**	**0.315**	0.227	0.231	**0.259**

images. Table 2 shows that cases relying only on plain text or only on structured text result in poor performance of MLLMs. Structured text is ineffective when not input alongside images.

We conducted ablation experiments categorized by different evidence elements from MMLongBench. The results in Table 3 indicate that for questions in documents that necessitate answers from charts or tables, using structured input provides a more pronounced performance gain than for other types of document elements. This observation further underscores the critical role of structured input.In the cases needed images to answer questions, the accuracy is lower. This is understandable because the lack of images causes information loss in cases where MLLMs need figures from the evidence images to answer questions. The accuracy with structured text ans images is better than with plain text and images, further highlighting the importance of structure. This experimental result convincingly demonstrates the importance of structure. We conclude that the combination of structured text and images is key to improving the ability of MLLMs in long document understanding, supporting Observation 3 from an ablation perspective.

5 CONCLUSION

Preserving the structure of the input to improve the ability of general-purpose multimodal MLLMs is essential to comprehend the underlying patterns and key factors in document understanding. This work makes a step in this line. We propose to use LATEX paradigm to keep the structure of plain text, the images combined with structured text efficiently improve the accuracy of answering questions in document understanding. Furthermore, we analyze the attention transformation of different kinds of input. This study find the structured attention is the key to make MLLMs understand better in answering DocQAs after comparing other inputs. Future work includes proposing novel and efficient structure extraction or attention control method to effectively unlock the ability of general-purpose MLLMs in document understanding.

6 Ethics statement

This research adhered to ethical guidelines, ensuring proper consent, data privacy, and confidentiality. No conflicts of interest are reported, and the findings are presented transparently.

7 Reproducibility statement

All relevant code, data, and materials are publicly available at `https://anonymous.4open.science/r/structured-attention_anonymous-8994/`, with documentation provided for reproducibility. Any limitations are clearly stated.

References

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.

Anthropic. Introducing claude, 2023. URL `https://www.anthropic.com/index/introducing-claude`.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024a.

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, 2024b.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*, 2024.

Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. V-doc: Visual questions answers with documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21492–21498, 2022.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Advances in Neural Information Processing Systems*, 37:42566–42592, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

Yulong Hui, Yao Lu, and Huanchen Zhang. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *arXiv preprint arXiv:2406.15187*, 2024.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

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*, 2024b.

Yujie Lu, Xiujun Li, Tsu-Jui Fu, Miguel Eckstein, and William Yang Wang. From text to pixel: Advancing long-context understanding in mllms. *arXiv preprint arXiv:2405.14213*, 2024.

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.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. Hierarchical visual feature aggregation for ocr-free document understanding. *Advances in Neural Information Processing Systems*, 37:105972–105996, 2024.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. *arXiv preprint arXiv:2412.10704*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. Focused transformer: Contrastive training for context scaling. *Advances in neural information processing systems*, 36:42661–42688, 2023.

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*, 2024a.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024b.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. *arXiv preprint arXiv:2412.02592*, 2024a.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024b.

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

A Appendix

A.1 Use of Large Language Models (LLMs)

In the preparation of this manuscript, Large Language Models (LLMs, e.g., ChatGPT) were used solely as a language editing tool to improve readability, grammar, and style. They were not involved in research ideation, experimental design, data analysis, or the formulation of scientific conclusions. The authors take full responsibility for all scientific content and claims presented in this paper.