Deconstructing Documents, Reconstructing Understanding: Modality-Split Retrieval-Augmented Question Answering

William Albert Branch
College of Charleston
branchwa@g.cofc.edu

Navid Hashemi Tonekaboni
College of Charleston
hashemin@cofc.edu

Abstract

Modern retrieval-augmented generation (RAG) systems often struggle with reasoning over complex documents that contain both text and images, leading to hallucinations and unreliable answers. This poses a challenge for knowledge-intensive applications that demand trustworthy, grounded AI responses. To address this, we propose Modality-Split Retrieval-Augmented Generation (MS-RAG), a framework that explicitly separates reasoning over visual and textual modalities to improve answer accuracy and interpretability. We evaluate MS-RAG on a curated dataset of scientific papers paired with 120 manually constructed, visually grounded question–answer examples. The framework retrieves relevant content, decomposes it into modality-specific components, applies tailored prompts to a vision-language model (VLM), and aggregates the results into a coherent final response. Our experiments show that MS-RAG consistently outperforms strong text-only and unified VLM-RAG baselines in factual accuracy, semantic relevance, and alignment with human judgment. These results highlight the value of modality-specific reasoning for improving transparency and reliability in document-based question answering, offering a practical foundation for knowledge management systems that integrate multimodal scientific and organizational content.

Keywords: Knowledge Management Systems, Retrieval-Augmented Generation (RAG), Vision Language Models (VLMs), Question Answering, Multimodal Question Answering, Hallucination Reduction

1. Introduction

Large language models (LLMs) have demonstrated impressive performance across a wide range of natural language processing tasks, including summarization, generation, and question answering. With advances in vision-language models (VLMs), these capabilities now extend to multimodal documents, enabling new forms of machine understanding for image-text pairs (Zhang et al., 2024). VLMs have shown promise for applications such as image captioning, visual question answering, and multimodal retrieval—tools that could enrich knowledge management (KM) in domains ranging from healthcare to enterprise compliance.

A popular framework commonly used to implement document question and answering systems is Retrieval-Augmented Generation (RAG). Instead of relying solely on a model's memory, RAG systems first encode a set of information using an embedding model (Lewis et al., 2020). When the user presents a query, it is encoded using the same embedding model. The most relevant information segments are retrieved using a similarity search between the encoded query and information. The retrieved content is passed along with the original query into a language model for query answer generation. This approach is particularly valuable in long-context or knowledge-intensive tasks, as it allows the system to dynamically pull information from a source (Salemi and Zamani, 2024). In combination with VLMs, RAG systems show the potential to answer questions regarding both images and text.

However, due to their architecture, the use of VLMs for AI-powered knowledge management or RAG systems is not easily implemented. Recent VLMs typically utilize a Vision Transformer (ViT) (Dosovitskiy et al., 2021), an extension of the standard transformer architecture (Vaswani et al., 2017) for images. A ViT breaks down an image into smaller patches-rather than text into tokens as done by a standard transformer. The patches are serialized into embedding vectors and augmented with positional data to maintain location within the original image. The sequence is fed into a transformer encoder, which learns the relationship between the patches on the basis of a self-attention mechanism. Despite the similar architecture to standard text transformers, ViTs require more data to train a greater number of parameters since image data is higher dimensional than text data.

Due to the larger training set required for VLMs and ViTs to accurately understand complex details, these models still struggle with fine-grained visual reasoning—especially when interpreting complex tables, graphs and diagrams often found in scientific or enterprise documents. These limitations can lead to shallow reasoning, hallucinated outputs, or contextual misunderstandings, which compromise the integrity and trustworthiness of AI-generated knowledge (Y. Li et al., 2025).

This shortcoming with VLMs can be highlighted in multimodal document RAG pipelines. These implementations often rely on VLMs to treat entire pages of documents as singular visual inputs-putting a large amount of high dimensional image data into a VLM. This approach entangles modalities and requires the ViT to distinguish between figures, diagrams, tables and texts all on the same page. This modality tangling often leads to incorrect reasoning when visual and textual elements are interwoven.

To mitigate these limitations, some recent approaches adopt multi-hop reasoning strategies, where a model processes one modality (e.g., text) before conditioning a follow-up query over another modality such as an image (Talmor et al., 2021). While more effective than singular queries, these methods still rely on a shared model to integrate both modalities, limiting specialization and interpretability.

To address these challenges, we propose Modality Split Retrieval-Augmented Generation (MS-RAG)—a modular framework designed to enhance multimodal document understanding by explicitly separating image and text reasoning. After retrieving relevant content, MS-RAG decomposes the retrieved document into modality specific components-text and images in our experiments. Each is passed separately to an instance of the Qwen2.5-VL-7B-Instruct model (hereafter referred to as Qwen) (Wang et al., 2024), using specially engineered, modality specific prompts. A final aggregation step, again using the same Qwen model with an engineered prompt, combines the outputs into a unified response.

This architecture reduces cross-modal interference by ensuring each modality is analyzed independently, while also maintaining a lightweight solution to limit computing constraints. We evaluate MS-RAG on a benchmark of visually grounded questions from scientific PDFs and show that it significantly outperforms a generic VLM-RAG baseline. Our results demonstrate the value of modality-specific reasoning for AI-generated knowledge, offering a pathway toward more trustworthy KM systems.

2. Related Work

Recent advances in multimodal question answering and RAG have focused on unified architectures that process visual and textual inputs jointly. However, less attention has been given to modular strategies that explicitly separate reasoning across modalities and sub-modalities. We review prior work in RAG, visual question answering, and vision-language models that inform the design of MS-RAG.

2.1. Vision-Language Models for Document Understanding

Recent advances in VLMs have enabled AI systems to jointly reason over image and text data. Pioneering VLMs such as CLIP (Radford et al., 2021) and BLIP (J. Li et al., 2022) showed promise for future VLMs. They fuse modalities via shared embeddings or cross-attention mechanisms, powering applications like visual question answering and image captioning. However, when applied to document-level knowledge extraction, VLMs often fail to capture symbolic and structural nuances—particularly in settings like scientific or legal documents where complex diagrams and charts are common (Mukhopadhyay et al., 2024). These limitations pose serious risks for AI-driven knowledge management, where hallucinated or misinterpreted content can distort organizational decision-making. We aim to minimize hallucinations caused by fused modalities by analyzing each modality independently to reduce the amount of context provided to a model in a single query.

2.2. Multi-hop QA for Multimodal Reasoning

MultimodalQA (Talmor et al., 2021) is a benchmark introduced to test reasoning across text, tables, and images using multi-hop logic. The benchmark includes 29,918 questions, of which 35.7% of questions require multimodal reasoning. The benchmark was created to test ImplicitDecomp, a model that processes queries

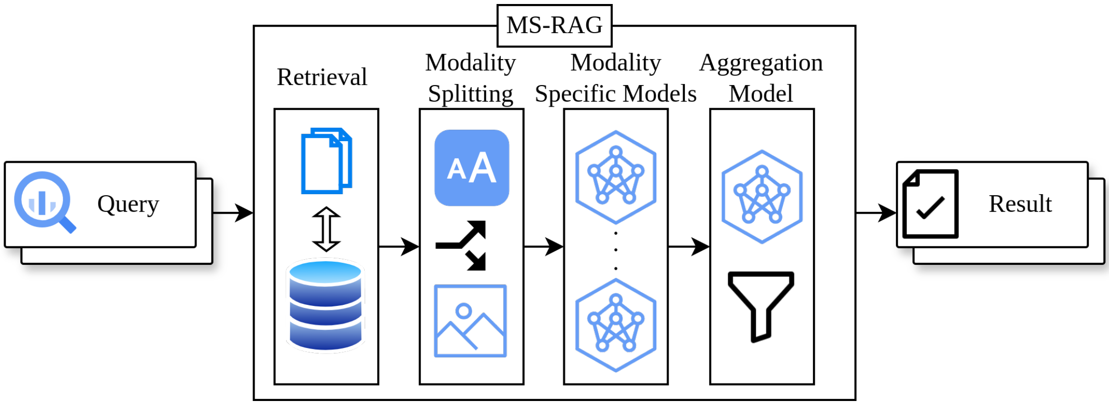


Figure 1. MS-RAG System Architecture

in two hops, first analyzing one modality (e.g., text) before conditioning the second hop on another (e.g., image). While this strategy supports compositional reasoning, it relies on a single processing pipeline that limits modality specialization. In contrast, MS-RAG separates visual and textual reasoning into independent pathways-not hops-enabling independent interpretation of modality specific content. The benchmark also does not require any KM system such as RAG.

2.3. Modular Reasoning with External Tools

MM-REACT (Yang et al., 2023) is a system that takes a modular, tool-based approach to multimodal reasoning. It prompts ChatGPT to invoke specialized vision tools (e.g., OCR, object detection), integrating their outputs into the response generation process. This approach enables zero-shot reasoning across modalities, but depends on prompt chaining and tool selection logic, and lacks a formal retrieval-augmented pipeline. MS-RAG differs in that it uses a structured RAG framework, with explicit visual-textual separation and controlled answer aggregation.

2.4. Multimodal RAG Architectures

Works such as MuRAG (Chen et al., 2022), VisRAG (Yu et al., 2025), and Visual-RAG (Wu et al., 2025) have extended RAG systems into the multimodal space. These systems retrieve image-text pairs or page-level image renderings and process them through joint encoders. However, they continue to struggle from hallucinations due to VLM constraints. MuRAG retrieves relevant content from a corpus that contains isolated images, text, or image-text pairs. However, it does not separate modalities or perform modality specific processing or response generations. VisRAG creates a pipeline that retrieves relevant document pages and feeds the pages as an image into a VLM for analysis, but does not separate the modalities. Visual-RAG introduces a benchmark to test image retrieval and question answering, but it does not test any sort of multimodal retrieval. MS-RAG addresses these issues by retrieving multimodal data, separating modalities into specific pipelines, and aggregating the outputs, therefore offering a more robust and interpretable solution for AI-enabled knowledge systems.

3. System Architecture

The Modality Split Retrieval-Augmented Generation (MS-RAG) framework is designed to enhance the reliability and accuracy of AI-generated responses in multimodal document question-answering settings. Traditional VLM-RAG pipelines often intertwine image and text content, leading to modality interference and hallucinations. MS-RAG mitigates these risks by explicitly decomposing the reasoning process into four distinct but coordinated stages: retrieval, modality-splitting, modality-specific inference, and aggregation steps as seen in Figure 2.

3.1. Retrieval with ColPali

MS-RAG begins with a retrieval stage. In this implementation we use the ColPali framework due to its high retrieval accuracy (Faysse et al., 2024). ColPali indexes each page of a document corpus as a dense grid of patch-level image embeddings, enabling visual similarity matching between user queries and PDF content. Given a natural language query, the top k most relevant document pages are retrieved using cosine similarity across the embedded representations, where k is determined by the user as the number of pages to return. ColPali successfully returns the correct page at an average rate of 81.3% across a variety of domains when $k = 1$. If k is too large, the potential for hallucinations increases by subsequently providing the VLM with too much context. We therefore chose to set $k = 2$ to reduce the chance that the desired page is not retrieved by ColPali while also mitigating the risk for hallucinations.

Unlike conventional RAG systems that retrieve raw text or perform OCR, this image-native retrieval supports scenarios where knowledge is encoded in diagrams, charts, and figures—formats commonly found in technical, medical, and academic documents.

3.2. Modality Splitting

After retrieval, each document page is decomposed into modality-specific components through a modality-aware parsing pipeline. In our implementation, text content is extracted using the PyMuPDF (McKie and Inc., 2023) library, which reliably isolates text structures. Visual content is segmented using a layout detection model based on the LayoutParser (Shen et al., 2021) framework, which preserves the spatial relation of compound image figures-which would otherwise be lost via a traditional PDF parser such as PyMuPDF.

The PyMuPDF library, provides access to the layout and content of PDF documents. Unlike traditional OCR approaches, PyMuPDF parses embedded text directly from the PDF file structure, preserving natural paragraph flow. We extract and concatenate text blocks in reading order while discarding headers, footers, and other layout noise when possible. This process produces clean, paragraph-level text segments that are well-suited for language model inference.

The LayoutParser framework is used to implement a layout-aware detection model. Rather than relying on standard PDF parsing libraries, which extract images without retaining their positional relationships, LayoutParser applies deep learning–based object detection to identify meaningful visual regions. This allows the system to preserve compound structures such as multi-panel figures, annotated diagrams, and grouped visual elements that would otherwise be separated. LayoutParser returns bounding box coordinates, which we use to crop and segment each figure before passing it to the model. This step is essential for maintaining the positional integrity of visual inputs.

3.3. Modality-Specific Reasoning via Prompt Specialization

Each modality is then processed independently in its isolated pipeline using a single instance of Qwen. Instead of deploying multiple, different models, MS-RAG employs prompt specialization to direct the model's reasoning behavior. Crucially, each modality is kept strictly isolated at both the input and prompt level to minimize modality interference. In our implementation, the two pipelines include:

- **Textual Reasoning:** Extracted text blocks along with the user query are presented to the model using a prompt tailored for textual reasoning. These prompts contain no references to figures, images, or visual elements.

- **Visual Reasoning:** Cropped figure regions along with the user query are shown to the model using a prompt tailored for image reasoning. These prompts assume no textual content and make no use of surrounding document text.

Each pipeline sends Qwen a custom system and user prompt, each playing a critical role in controlling model behavior. The system prompt acts as an instruction layer that establishes model constraints while the user prompt contains the actual query alongside the modality-specific retrieved context. To enforce strict adherence to the retrieved information, the system prompt explicitly instructs the model to generate answers only using the provided context, and to avoid relying on prior knowledge or making speculative inferences. If the answer cannot be determined from the supplied content, the model is directed to say so explicitly. This strict modality isolation ensures that reasoning remains grounded in the retrieved, modality specific context, and avoids hallucinations or unsupported claims. It also removes opportunities for cross-modal interference, which can bias outputs.

3.4. Aggregation and Conflict Resolution

The final stage merges the modality specific responses into a unified answer. Using Qwen, an aggregation prompt is issued that presents the original query alongside both intermediate answers. The model is tasked with synthesizing the two candidate responses into a singular final output. Information conflicts are resolved by constructing a system prompt that requires the model to include each side of the contradicting information-therefore allowing the user to determine the most appropriate answer.

This aggregation phase plays a critical role in knowledge validation, acting as an arbitration mechanism that reduces hallucinations by highlighting discrepancies and promotes consistency. Importantly, it enables human-AI collaboration by exposing the

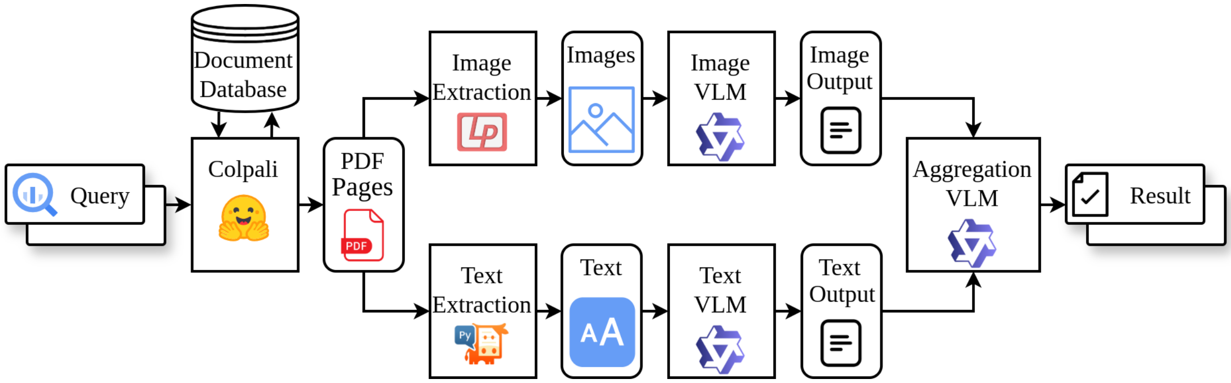


Figure 2. MS-RAG Implementation Flowchart

distinct reasoning paths and making final responses more interpretable to downstream users.

4. Dataset

To evaluate the performance of MS-RAG in multimodal knowledge reasoning, we curated a dataset of 100 scientific papers from the NeurIPS 2024 conference. NeurIPS is one of the most prestigious venues in machine learning, known for publishing cutting-edge research that often includes detailed visual elements such as diagrams, plots, and charts. These visual components tend to encode complex analytical findings, making NeurIPS papers a strong benchmark for evaluating systems that aim to reason across both text and image modalities.

The curation process emphasized diversity in visual content types, including architectural diagrams, performance plots, and annotated examples, to reflect the breadth of visual reasoning tasks encountered in technical literature. The papers varied in length, with most ranging from 20 to 60 pages. We prioritized papers that presented technical results visually, such as model comparisons, data distributions, or architectural schematics, to ensure a diverse and representative multimodal corpus.

From this curated collection, we constructed a benchmark of 120 visually grounded questions that explicitly require interpretation of visual content. Each question is associated with a specific page in one of the selected papers that contains both textual and visual information. These questions were written to probe a model's ability to retrieve, interpret, and reason over image-based content, such as identifying trends in plots or extracting conclusions from diagrams.

To better simulate real-world information-seeking behavior, the questions are phrased in natural language and intentionally avoid explicit references to figure numbers or titles. This design ensures that the retrieval component of MS-RAG must rely on semantic understanding to identify and associate the appropriate visual content, rather than depending on superficial cues. As a result, the benchmark provides a more realistic and challenging test of multimodal retrieval and reasoning capabilities.

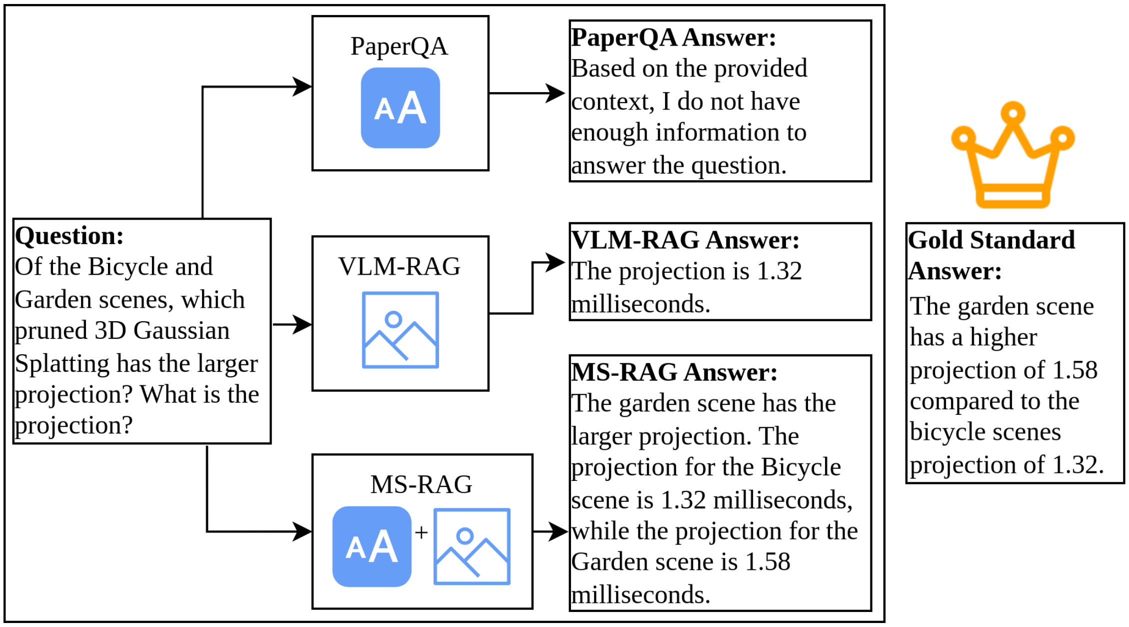
5. Experiments

To assess the effectiveness of MS-RAG in multimodal document understanding, we evaluated its performance by creating a curated 120-question benchmark where each question targets a specific image within the document set. For a more detailed analysis, each question was categorized into one of three sub-modalities based on what type of image the question was targeting:

- **Tables** – Standard tabular representations.
- **Graphs and Charts** – Includes plots, charts, histograms, and line graphs.
- **Diagrams and Visual Schemas** – Includes flowcharts, illustrations, and visual workflows.

This categorical breakdown allows for a more fine-grained evaluation of each model's strengths and weaknesses across visual modalities. It also mirrors real-world use cases in knowledge management systems, where structured visuals play a critical role in encoding domain expertise.

To generate each question, we scanned the documents for images. When an image was identified, it was first classified into one of the three categories. A question was written to directly reference the visual content, deliberately avoiding mention of figure numbers and titles to simulate a more realistic document-level question answering scenario. After



Question: Of the Bicycle and Garden scenes, which pruned 3D Gaussian Splatting has the larger projection? What is the projection?

PaperQA Answer: Based on the provided context, I do not have enough information to answer the question.

VLM-RAG Answer: The projection is 1.32 milliseconds.

MS-RAG Answer: The garden scene has the larger projection. The projection for the Bicycle scene is 1.32 milliseconds, while the projection for the Garden scene is 1.58 milliseconds.

Gold Standard Answer: The garden scene has a higher projection of 1.58 compared to the bicycle scenes projection of 1.32.

Figure 3. An example of each models output in comparison to the gold standard answer.

formulating a question, a gold-standard answer was manually composed. To align the gold-standard answers stylistically with model generations—and to reduce lexical bias in automatic evaluation metrics such as cosine similarity and ROUGE—we used Qwen to paraphrase each human-written answer. The result is a set of gold-standard answers that better represents the generative tendencies of Qwen.

Each question was submitted to three different systems, each implemented using the Qwen model for answer generation:

- **PaperQA (Text Only)** – A popular open-source document QA system that uses a RAG system over text and document structure (Lála et al., 2023). We use the PaperQA pipeline with Qwen as the underlying model.

- **VLM-RAG (Baseline)** – A conventional vision-language approach in which each page retrieved via ColPali is passed as an image input to Qwen along with the query. No modality separation or aggregation is performed.

- **MS-RAG (Proposed)** – Our modality-split framework in which retrieved content is decomposed into separate text and image components. Each is processed independently via prompt specialization using a single instance of Qwen, and responses are merged via an aggregation step.

To evaluate model performance, each generated answer was compared to a manually curated gold-standard reference. An example of each model's generated output compared to the gold-standard output can be seen in Figure 3. We employed four complementary evaluation metrics. Cosine similarity was used to compute the semantic distance between model outputs and gold-standard answers via sentence embeddings. ROUGE-1 and ROUGE-2 measure the overlap of unigrams and bigrams respectively between a generated response and a reference answer (Lin, 2004). ROUGE-1 is better suited to show the amount of lexical overlap while ROUGE-2 captures short phrases. Finally, we conducted a manual review in which each response was scored on a 4-point scale: 1 = irrelevant, 2 = incorrect, 3 = partially correct, and 4 = correct. This combination of quantitative and human-judged metrics allows for a robust comparison of model accuracy, relevance, and knowledge fidelity—key dimensions for trustworthy AI systems.

6. Results

Table 1 reports the average cosine similarity between model outputs and gold-standard answers across three question categories. MS-RAG consistently outperforms the VLM-RAG and PaperQA across all categories. The largest improvement appears in the *Graphs & Charts* category, where MS-RAG yields a similarity of 0.5731 compared to 0.5193 and 0.1766 for the VLM-RAG and PaperQA models respectively. Gains

Category	PaperQA	VLM-RAG	MS-RAG	# Questions
Tables	0.1892	0.3657	**0.4325**	40
Graphs & Charts	0.1766	0.5193	**0.5731**	40
Diagrams & Visual Schemas	0.0569	0.4661	**0.4823**	40

Table 1. Average cosine similarity between model responses and gold-standard answers.

Category	PaperQA		VLM-RAG		MS-RAG	
	R-1	R-2	R-1	R-2	R-1	R-2
Tables	0.1628	0.0796	0.2001	0.0788	**0.2986**	**0.1555**
Graphs & Charts	0.1664	0.0689	0.3386	0.1588	**0.3825**	**0.1922**
Diagrams & Visual Schemas	0.1715	0.0702	**0.3779**	0.1845	0.3464	**0.2148**

Table 2. ROUGE-1 and ROUGE-2 F1 scores across question categories and models.

Category	PaperQA	VLM-RAG	MS-RAG	# Questions
Tables	2.25	3.00	**3.30**	40
Graphs & Charts	1.28	2.75	**3.12**	40
Diagrams & Visual Schemas	1.44	2.20	**3.00**	40

Table 3. Average manual evaluation scores (1 = irrelevant, 4 = fully correct) across question categories.

are also evident in *Tables* (0.4325 vs. 0.3657 vs. 0.1892) and *Diagrams & Visual Schemas* (0.4823 vs. 0.4661 vs 0.0569), supporting the hypothesis that modality-specific reasoning improves semantic alignment.

Table 2 shows ROUGE-1 and ROUGE-2 F scores. MS-RAG demonstrates stronger performance on most metrics, particularly in the *Tables* category. MS-RAG also improves ROUGE-1 and ROUGE-2 on *Graphs & Charts*, though it slightly underperformed on ROUGE-1 in the *Diagrams* category in comparison to the VLM-RAG model.

Table 3 presents the results of human evaluation. We manually scored each answer from 1 (irrelevant) to 4 (fully correct). MS-RAG again performs better overall, achieving an average score of 3.30 in *Tables* in comparison to 3.00 and 2.25, 3.12 in *Graphs & Charts* in comparison to 2.75 and 1.28, and 3.00 in *Diagrams and Visual Schemas* in comparison to 2.20 and 1.44.

Together, the results demonstrate that MS-RAG's modality-split approach yields more factually accurate and contextually relevant answers. MS-RAG outperforms the baselines on almost all metrics, with notable improvements in semantic similarity and human evaluation scores—indicating fewer hallucinations and stronger grounding in both text and image content. The specialized prompting and aggregation steps help the model attend more effectively to the relevant modality, producing responses that align more closely with human expectations. These findings support our central claim that modality-specific reasoning enhances the reliability and trustworthiness of AI systems for document question answering.

7. Discussion

Our experimental results demonstrate that modality-split reasoning can significantly improve the quality and reliability of AI-generated answers in multimodal document settings. MS-RAG outperformed the VLM-RAG and PaperQA pipelines across most metrics and categories. These findings reinforce the central hypothesis that treating each modality separately—rather than passing full pages into a monolithic VLM—can lead to more accurate outputs and reduced hallucinations.

7.1. Model Selection

It is important, however, to contextualize these results within the capabilities of the underlying models. All evaluations in this paper were conducted using Qwen2.5-VL-7B-Instruct - a relatively small, publicly available, and non-finetuned model. The improvements observed are thus achieved without the benefit of scale, domain adaptation, or supervised fine-tuning. It is likely that a larger model, such as GPT-4o (OpenAI, 2024) or Gemini Pro (Google, 2025), would perform better overall, potentially resolving some of the failures observed in both the baseline and MS-RAG pipelines. Utilizing one of these large models would, however, come at the cost of efficiency. Still, the consistent gains achieved by MS-RAG suggest that architectural improvements in reasoning structure can yield meaningful benefits even at smaller scales.

An extension of this thought is the potential advantage of selecting specialized models for each modality, rather than relying on a single vision-language model (VLM) across all reasoning tasks. While VLMs like Qwen offer impressive multimodal capabilities, they are inherently designed to balance both visual and textual processing, which may compromise their effectiveness when dealing exclusively with one modality. For example, tasks involving purely textual reasoning or aggregation of multimodal outputs may be better served by a language model specifically optimized for instruction-following in text, such as LLaMA 3.2 Instruct (Meta AI, 2024). Similarly, for visual understanding, dedicated vision transformers pre-trained on high-resolution, structured image data may offer improved perception over more generalized VLMs. Future work could explore modular systems where each component—image processing, text comprehension, and answer aggregation—is handled by the model type best suited for that modality. This would most likely lead to gains in accuracy, but greatly sacrifice processing speed by requiring many different models to be loaded at once, something that we tried to avoid by utilizing a single instance of Qwen.

7.2. Inherent Randomness

Despite overall gains in favor of MS-RAG, the ROUGE-1 metric indicated that VLM-RAG performed better than MS-RAG in the Diagram and Visual Schemas Category. This may be due to the loss of holistic spatial context that occurs when segmenting figures from a page. However, it is more likely associated with how the answers are generated. Many steps were taken to align the semantics of the manually generated gold-standard outputs with the way Qwen generates its outputs, but it was not a perfect process. Aligning the semantics is important because that is how ROUGE and cosine similarity judge the responses. Language models are inherently random, and predicting how they generate an output can be difficult. The fact that all other metrics ruled in favor of MS-RAG within this category means that this discrepancy is most likely attributed to inherent randomness.

7.3. Modality-Split Necessity

It is also worth noting that MS-RAG's modality-split architecture may be unnecessary in scenarios where the question is clearly grounded in a single modality. For instance, when a document contains only textual data, or when a question pertains solely to an image without associated text, simpler models may suffice and offer efficiency benefits. In such cases, the overhead introduced by modality decomposition and aggregation may not provide meaningful gains. Modality splitting is best suited for contexts where accurate knowledge synthesis requires interpreting both text and image content—or at minimum, where the dominant modality is ambiguous or structurally interdependent with the others.

7.4. Failure Point

Incorrect answer generation issues are not typically caused by poor information retrieval; in nearly all cases, the relevant visual or textual context is correctly retrieved by ColPali. The breakdown occurs during interpretation, suggesting that the root cause is the underlying models rather than procedure. VLMs still struggle with symbolic visual reasoning because they encode image patches as token-like elements in transformer architectures originally designed for language. When images are treated like text without structural awareness, fine-grained visual understanding is lost.

From a knowledge management perspective, this failure mode is critical. Many organizations depend on highly structured documents—scientific reports, financial audits, medical records—where meaning is embedded in visuals such as tables, plots, and diagrams. MS-RAG offers a modular alternative: it introduces controlled separation of modalities and allows each to be processed under a more specialized, interpretable prompt. Its final aggregation step acts as a knowledge arbitration layer, enabling users and systems alike to audit how the answer was formed.

8. Conclusion

This paper introduced MS-RAG, a modality-split retrieval-augmented generation framework for answering questions over multimodal documents. Unlike conventional VLM-based or text-based RAG systems, MS-RAG processes each modality-visual and textual content-separately using distinct models, then aggregates the outputs to produce more accurate and interpretable answers. By disconnecting the reasoning pipelines for each modality, MS-RAG mitigates common failure modes such as hallucination and modality interference, especially in regard to image analysis misinterpretations.

We evaluated our approach on a curated benchmark of scientific PDF documents and showed that MS-RAG outperforms both text-only and VLM-RAG baselines. These results highlight the importance of modality-specifc frameworks in document question answering and suggest that more structured, decomposed reasoning approaches can better align with the strengths of current multimodal models.

MS-RAG provides a practical and scalable step toward more trustworthy multimodal systems. It reduces interference between modalities, makes reasoning pathways more inspectable, and shows gains even under modest compute constraints. As models become larger and more capable, we expect the principles behind MS-RAG—modular decomposition, prompt specialization, and aggregation-based arbitration—to remain relevant, particularly for domains where AI must justify its answers with clarity and accuracy.

References

Chen, W., Hu, H., Chen, X., Verga, P., & Cohen, W. W. (2022). Murag: Multimodal retrieval-augmented generator for open question answering over images and text. https://arxiv.org/abs/2210.02928

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., & Colombo, P. (2024). Colpali: Efficient document retrieval with vision language models. https://arxiv.org/abs/2407.01449

Google, D. /. (2025). Gemini 2.5 pro: Our most intelligent ai model [Model card published March 25, 2025].

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., & White, A. D. (2023). Paperqa: Retrieval-augmented generative agent for scientific research. https://arxiv.org/abs/2312.07559

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, *abs/2005.11401*. https://arxiv.org/abs/2005.11401

Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, *abs/2201.12086*. https://arxiv.org/abs/2201.12086

Li, Y., Lai, Z., Bao, W., Tan, Z., Dao, A., Sui, K., Shen, J., Liu, D., Liu, H., & Kong, Y. (2025). Visual large language models for generalized and specialized applications. *arXiv preprint arXiv:2501.02765*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text summarization branches out: Proceedings of the ACL-04 workshop*, 74–81.

McKie, J. X., & Inc., A. S. (2023). *PyMuPDF: Python bindings for mupdf* [Version 1.23.7. DOI: https://doi.org/10.5281/zenodo.8303351]. https://pymupdf.readthedocs.io

Meta AI. (2024). Llama 3.2 instruct: Instruction-tuned text generation models [Model card for LLaMA 3.2 Instruct series (1 B, 3 B, 11 B, 90 B)]. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Mukhopadhyay, S., Qidwai, A., Garimella, A., Ramu, P., Gupta, V., & Roth, D. (2024, November). Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 16696–16717). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.973

OpenAI. (2024). Hello GPT-4o: The omni-modal model [Model card published May 13, 2024].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *CoRR*, *abs/2103.00020*. https://arxiv.org/abs/2103.00020

Salemi, A., & Zamani, H. (2024). Evaluating retrieval quality in retrieval-augmented generation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2395–2400.

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*.

Talmor, A., Yoran, O., Catav, A., Lahav, D., Wang, Y., Asai, A., Ilharco, G., Hajishirzi, H., & Berant, J. (2021). Multimodalqa: Complex question answering over text, tables and images. *CoRR*, *abs/2104.06039*. https://arxiv.org/abs/2104.06039

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., &

Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, Y., Long, Q., Li, J., Yu, J., & Wang, W. (2025). Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. https://arxiv.org/abs/2502.16636

Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). Mm-react: Prompting chatgpt for multimodal reasoning and action. https://arxiv.org/abs/2303.11381

Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., & Sun, M. (2025). Visrag: Vision-based retrieval-augmented generation on multi-modality documents. https://arxiv.org/abs/2410.10594

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.