VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Understanding information from a collection of multiple documents, particularly those with visually rich elements, is important for document-grounded question answering. This paper introduces VisDoMBench, the first comprehensive benchmark designed to evaluate QA systems in multi-document settings with rich multimodal content, including tables, charts, and presentation slides. We propose VisDoMRAG, a novel multimodal Retrieval Augmented Generation (RAG) approach that simultaneously utilizes visual and textual RAG, thereby combining robust visual retrieval capabilities with sophisticated linguistic reasoning. VisDoMRAG employs a multi-step reasoning process encompassing evidence curation and chain-of-thought reasoning for concurrent textual and visual RAG pipelines. A key novelty of VisDoMRAG is its consistency-constrained modality fusion mechanism, which aligns the reasoning processes across modalities at inference time to produce a coherent final answer. This leads to enhanced accuracy in scenarios where critical information is distributed across modalities and improved answer verifiability through implicit context attribution. Through extensive experiments involving open-source and proprietary large language models, we benchmark state-of-the-art document QA methods on VisDoMBench. Extensive results show that VisDoMRAG outperforms unimodal and long-context LLM baselines for end-to-end multimodal document QA by 12-20%.

1 Introduction

In today's information-rich landscape, PDF documents play a crucial role in storing and disseminating information across various domains, including finance, legal, scientific research, and more. These documents often contain a rich blend of textual, visual, and tabular data, making them a unique challenge for information retrieval systems. Unlike structured formats like databases, PDFs are inherently unstructured, with diverse layouts combining paragraphs, images, charts, and tables. This complexity demands sophisticated multimodal processing techniques capable of interpreting both the textual and visual content. Effective handling of multimodal content from PDFs is essential for downstream tasks such as question-answering (Ding et al., 2022; Mathew et al., 2021), summarization (Pang et al., 2023), and knowledge extraction (Pal et al., 2023), where accurate and context-aware data extraction can significantly enhance decision-making processes. As a result, developing advanced methods that can fully leverage the multimodal nature of PDF documents has become a critical research challenge.

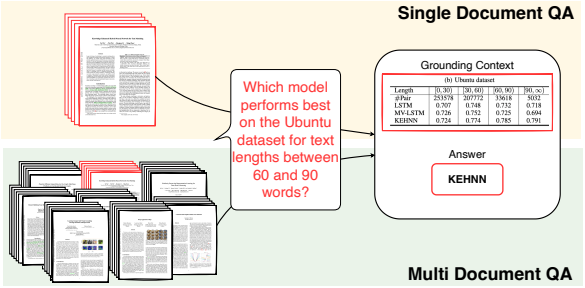


Figure 1: Multi-document QA systems require inferring relevant context from a large volume of unstructured data, inherently making it a more challenging task than single-document QA.

In real-world document QA systems, queries are often directed over a collection of source documents rather than a single source, requiring the system to identify the document that contains the relevant answer. This reflects common scenarios in domains such as finance, science, and policy analysis, where users interact with large, varied document sets to find specific information. In these cases, the challenge lies in effectively localizing context relevant to the query, from a large volume of information distributed across multiple documents

(akin to finding a "needle in a haystack (Wang et al., 2024b)).

Multi-document QA datasets are scarce, with existing multi-document benchmarks (Bai et al., 2023; Wang et al., 2024c), predominantly focused on textual information, often overlooking the diverse content forms found in real-world documents, such as tables, charts, and visual elements. Visually rich elements, such as tables, charts, and slides, provide structured data and visual summaries that are critical for answering certain types of questions. Tables often present dense, organized information that cannot be captured through plain text. At the same time, charts and slides can visually depict trends, relationships, or distributions that require interpretation beyond textual descriptions. The absence of datasets that include these modalities limits the ability of current QA models to address complex, multimodal questions. For instance, answering a financial or scientific question may require interpreting both numerical data in tables and trends in graphs alongside the surrounding text.

In the context of visually rich content-based documents, existing RAG systems face a critical limitation due to their reliance on a singular modality (either text or vision) for retrieval. Text-based systems are proficient in linguistic reasoning but often overlook vital visual elements, such as tables and figures, that may contain key information. Conversely, multimodal RAG (Chen et al., 2022) systems that leverage vision-based retrieval can effectively extract visual data but are often constrained in end-to-end performance by the LLM's visual reasoning abilities, as text often performs better than visual input when given the same context (Deng et al., 2024), which can be attributed to language bias in visual LLMs (Niu et al., 2021; Wang et al., 2024a), and visual hallucination (Ghosh et al., 2024).

Main Results: We introduce `VisDoMBench`, the first multi-document, multi-modal QA dataset specifically designed to address rich visual content, including tables, charts, and slides. `VisDoMBench` encompasses a diverse range of complex content and question types, along with annotated evidence, allowing for a comprehensive evaluation of multimodal QA systems. In this work, we benchmark the performance of various visual and textual retrieval methods on `VisDoMBench`, providing insights into their effectiveness in handling visually rich, multi-document queries.

Further, we propose `VisDoMRAG`, a novel multimodal RAG approach that effectively performs modality fusion over textual and visual RAG pipelines, benefiting from the inherent strengths of both these approaches, unlike contemporary approaches, which perform only-text or only-vision-based retrieval. `VisDoMRAG` employs parallel RAG pipelines for text and visual elements, each with a multi-step reasoning process involving evidence curation, chain-of-thought reasoning, and answer generation. The system then integrates the outputs from both pipelines using modality fusion, which imposes a consistency constraint on the reasoning chains, ensuring inference-time alignment across the modalities' reasoning processes to produce the final answer. `VisDoMRAG` offers several significant advantages over traditional unimodal or simpler multimodal systems. Firstly, it ensures comprehensive information utilization by fully leveraging both textual and visual cues, leading to more accurate and complete answers, particularly in scenarios where critical information is distributed across different modalities. Moreover, the evidence curation step provides an additional advantage of answer verifiability, since context attribution is built into our approach. We conduct experiments utilizing various open-source and closed-source LLMs, comparing multiple strategies such as long-context processing, textual RAG, and visual RAG, with our proposed system. We find that our `VisDoMRAG` improves end-to-end QA performance on our benchmarks, with performance gains in the range of 12%-20%. Our **main contributions** are:

- **`VisDoMBench`, a novel multi-document, multimodal QA benchmark** designed to address QA tasks across visually rich document content such as tables, charts, and slides, allowing for a comprehensive evaluation of multimodal document QA systems.

- **`VisDoMRAG`**, a **novel multimodal RAG** approach that effectively parallelly performs textual and visual RAG via Evidence Curation and Chain-of-Thought reasoning. The output reasoning chains from both the modalities are aligned using consistency analysis and resultant answers are ensembled together via LLM-based modality fusion to enhance visually-rich document QA.

- **`VisDoMRAG` significantly outperforms strong baselines** such as long-context processing, textual RAG, and visual RAG on the

VisDoMBench corpus by 12-20% across various open and closed-source LLM settings.

2 Related Work

Retrieval Augmented Generation While Large Language Models (LLMs) have achieved significant advancements, they still encounter challenges in integrating external knowledge and adapting to new, unseen data. Retrieval Augmented Generation (RAG) addresses these gaps by incorporating external information, enhancing the precision and reliability of LLM responses (Lewis et al., 2020). RAG is utilized across various downstream unimodal NLP tasks, including machine translation (Gu et al., 2018; He et al., 2021), dialogue generation (Cai et al., 2018), abstractive summarization (Peng et al., 2019), and knowledge-intensive generation (Izacard and Grave, 2020; Lewis et al., 2020). In visual question answering (VQA), (Lin and Byrne, 2022) addresses open-domain challenges by using object detection, image captioning, and optical character recognition (OCR) to transform target images into textual data. Moving beyond text-only contexts, MuRAG retrieves both text and image data, incorporating images as visual tokens (Chen et al., 2022). RAMM enhances performance by retrieving and encoding similar biomedical images and their captions through distinct networks (Yuan et al., 2023).

Long Context Document Benchmarks The comparison of long context document question-answer benchmarks (Table 1), highlights the diversity in content types, multi-document capabilities, and domains. Existing benchmarks such as L-Eval (An et al., 2023), Marathon (Zhang et al., 2023), and LooGLE (Li et al., 2023) primarily focus on text-based content from multi-domain sources but do not support multi-document inputs. LongBench (Bai et al., 2023) and Loong (Wang et al., 2024c) extend their evaluations to include multi-document settings, although they remain text-centric.

Comparison with existing datasets: Certain benchmarks like MPDocVQA (Tito et al., 2023), UDA (Hui et al., 2024), and MMLONGBENCH-DOC (Ma et al., 2024) expand the content spectrum by incorporating tables, charts, and slides, but they are limited to single-document question answering. In contrast, VisDoMBench supports multi-document question answering across various content types, including text, tables, charts, and slides, offering a more comprehensive multi-domain evaluation framework.

3 Problem Formulation

Given a query q, we have a collection of M documents $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$, wherein each document d_i may consist of a set of N_i pages represented by $P^i = \{p^i_1, p^i_2, \ldots, p^i_{N_i}\}$. We aim to generate text $\hat{a}$ for each query q that accurately answers the user query. The answer generation relies on retrieving relevant evidence context from one or more documents. Each query q may require information spread across different pages from one or more of the associated documents in D.

We aim to propose a framework that can accurately answer questions over a collection of multi-page documents where the system first retrieves relevant evidence at the level of individual pages, paragraphs or text chunks, followed by using the retrieved context to generate answer text.

4 VisDoMBench

Every data point in VisDoMBench can be expressed as triple $(q, D, \hat{a})$, where a question q is posed to a set of documents D, with ground-truth answer $\hat{a}$. We re-purpose five existing document-QA datasets to form our benchmark. Table 2 summarises different data splits present in VisDoMBench, including summary statistics, QA type, and content type.

4.1 VisDoMBench

Data Sourcing: In the curation of document question-answering datasets, we adhered to the following criteria: (1) the inclusion of visually rich content, encompassing tables, charts, and presentation slides; (2) the utilization of publicly accessible source documents; and (3) the presence of grounded evidence. These parameters were established to ensure the datasets' relevance to multi-modal information retrieval and their applicability to real-world question-answering tasks. Our corpus comprises test/eval sets sourced from several established datasets. We incorporated the PaperTab and FeTaTab splits from the UDA Benchmark(Hui et al., 2024), which in turn sourced these datasets from QASPER(Dasigi et al., 2021) and FeTaQA(Nan et al., 2022), respectively. For chart-based question-answering samples, we drew from SciGraphQA (Li and Tajbakhsh, 2023), which is multi-turn QA dataset on charts from scientific papers, and SPIQA(Pramanick et al., 2024), a chart and table QA dataset system sourced from (Dasigi et al., 2021). Additionally, we included SlideVQA

Benchmark	Content Type	Multi Document	Domain
L-Eval (An et al., 2023)	Text	✗	Multi-domain
LongBench (Bai et al., 2023)	Text	✓	Wikipedia
Marathon (Zhang et al., 2023)	Text	✗	Multi-domain
LooGLE (Li et al., 2023)	Text	✗	Multi-domain
MPDocVQA (Tito et al., 2023)	Text, Tables, Charts	✗	Multi-domain
∞Bench (Zhang et al., 2024)	Text	✗	Multi-domain
Ruler (Hsieh et al., 2024)	Text	✗	Wikipedia
Loong (Wang et al., 2024c)	Text	✓	Multi-domain
UDA (Hui et al., 2024)	Text, Tables	✗	Multi-domain
NarrativeQA (Kočiskỳ et al., 2018)	Text	✗	Movies and Shows
MMLONGBENCH-DOC (Ma et al., 2024)	Text, Tables, Charts, Slides	✗	Multi-domain
VisDoMBench (Ours)	**Text, Tables, Charts, Slides**	✓	**Multi-domain**

Table 1: Comparison of long context document QA benchmarks with VisDoMBench.

Dataset	Domain	Content Type	Queries	Docs	Avg. Question Length	Avg. Doc Length (Pages)	Avg. Docs per Query	Avg. Pages per Query
PaperTab	Wikipedia	Tables, Text	377	297	29.44 ±6.3	10.55 ±6.3	10.82 ±4.4	113.10 ±50.4
FetaTab	Scientific Papers	Tables	350	300	12.96 ±4.1	15.77 ±23.9	7.77 ±3.1	124.33 ±83.0
SciGraphQA	Scientific Papers	Charts	407	319	18.05 ±1.9	22.75 ±29.1	5.91 ±2.0	129.71 ±81.7
SPIQA	Scientific Papers	Tables, Charts	586	117	16.06 ±6.6	14.03 ±7.9	9.51 ±3.5	135.58 ±55.2
SlideVQA	Presentation Decks	Slides	551	244	22.39 ±7.8	20.00 ±0.0	6.99 ±2.0	139.71 ±40.6

Table 2: Summary of data splits included in VisDoMBench.

(Tanaka et al., 2023), a multi-image, multi-hop QA dataset centered on presentation slide decks.

Data Sampling: Sourced QA pairs need to be sampled to retain high quality samples. To maintain the integrity and uniqueness of our benchmark, we meticulously removed overlapping samples between PaperTab and SPIQA and implemented rigorous de-duplication of QA pairs across all included datasets. Further, we also perform question-level de-duplication to ensure similar questions are not repeated across different document collections. This ensures that QA systems are not rewarded disproportionately for better handling particular question types. For SciGraphQA, we filter out trivial questions related to layout and document metadata. From the remaining questions, we randomly sample 500 questions from the top 50%-ile of questions by length. The rationale for filtering on answer length filter is based on the heuristic that longer questions tend to be more specific, making them better suited for multi-document QA tasks, where specificity is crucial. For SlideVQA, we exclude single-hop questions, as they are generally non-specific and may have more than one correct answer from the document collection. We heuristically observe that multi-hop questions in this dataset are more likely to reference content from specific documents, thus making them a better fit for multi-document setups. SciGraphQA and SPIQA contain questions specific to charts or tables extracted from scientific papers. We use the ArXiv API[1] to extract full document PDFs.

[1]https://info.arxiv.org/help/api/index.html

Document Augmentation To simulate realistic multi-document settings, we augment each question across all data splits with varying number of distracting documents, ($|\mathcal{D}_i = M|$). We intend to keep the expected number of total pages per query between 50 to 200 to ensure that there is sufficient distracting content while maintaining the practical feasibility of contemporary long-context models. Hence, based on the average number of pages per document P_{avg}, we randomly sample the number of distracting documents l to lie between the range $[\lfloor \frac{50}{P_{avg}} \rfloor, \lfloor \frac{200}{P_{avg}} \rfloor]$. Randomly sampling l ensures that each benchmark instance contains a diverse degree of multi-document evidence, allowing for a more thorough evaluation of the QA model's retrieval and reasoning capabilities.

Query Augmentation: To address the challenge of ambiguous questions in datasets such as SciGraphQA, and PaperTab, we implement a query augmentation procedure to create a one-to-one mapping between a given question and the document(s) that exclusively answer it. Given an original question and the document containing answer, we utilize GPT-4o to generate more specific variations of the question, ensuring that the generated question can only be answered by the corresponding document. To maintain consistency, we constrain the LLM such that the answer to the generated question must match the provided answer. Once the augmented queries are generated, a human annotator reviews them using a predefined rubric. The rubric guides the annotator to either select one of the five generated questions, retain the

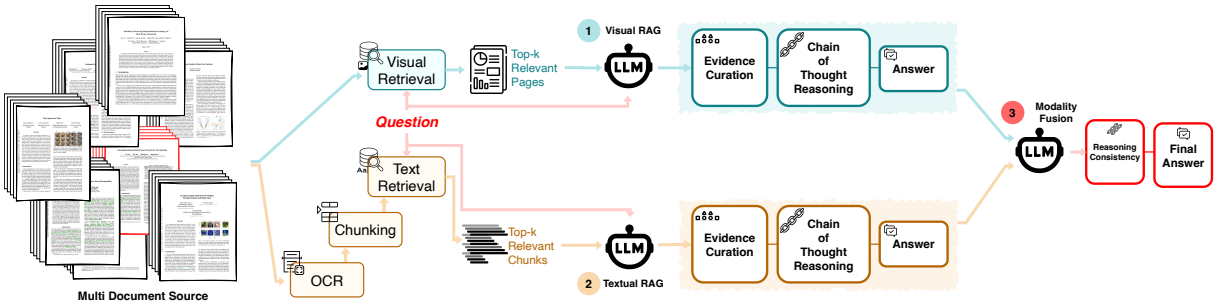


Figure 2: VisDoMRAG: Given a set of documents, VisDoMRAG parallelly performs evidence-driven ❶ Visual RAG and ❷ Textual RAG, prompting the LLMs to answer a query based on the respective retrieved context via Evidence Curation and Chain-of-Thought reasoning. The reasoning chains, and answers from the text and visual pipeline are ensembled together via ❸ Modality Fusion, where the outputs of both the modalities are aligned using consistency analysis on their reasoning chain to arrive at the final answer.

original question, or mark all questions (synthetic and actual) as ambiguous, in which case, the data point is discarded. The annotator is tasked with ensuring that the question is sufficiently specific by cross-referencing the localized evidence. Additionally, the annotator performs a simple search across the entire document collection to verify that the question cannot be ambiguously answered by any other document.

5 VisDoMRAG

VisDoMRAG (Fig 2) is a multimodal RAG approach for visually rich document QA consisting of two steps: (i) parallel evidence-driven unimodal (vision and textual) RAG pipelines, and (ii) Modality Fusion, which imposes consistency constraints to combine unimodal reasoning chains and arrive at a final answer.

5.1 Evidence-driven Parallel Unimodal RAG

Textual Retrieval Pipeline The textual RAG pipeline commences with the extraction of text from the set of documents utilizing Optical Character Recognition (OCR), followed by the segmentation of the extracted text into smaller, indexable chunks. Metadata indicating the source document and page number is preserved to facilitate traceability. These chunks are then indexed using a text embedding model, enabling efficient retrieval. Relevant chunks are subsequently retrieved in relation to the specified query by a text retrieval model and provided as contextual input to the LLM along with the query to generate textual answer response.

Visual Retrieval Pipeline Simultaneously, the visual RAG pipeline is dedicated to the extraction and analysis of graphical elements, including images, charts, and diagrams. For a given set of PDFs, a visual embedding model generates an index at the page-level granularity for all documents. Relevant pages are then retrieved by a visual retrieval model based on the specified query, and these pages are supplied to multimodal LLMs as visual context. This approach ensures that the model has access to critical visual information, employing its multimodal capability to utilize visual cues from document layout and graphical structures such as charts, diagrams and infographics.

Prompting Strategy Both the textual and visual pipelines employ a sophisticated three-step prompting strategy. Given a set of context artifacts (page images or textual chunks), and a query, the LLM is prompted with the following steps:

1. Evidence Curation: As a first step, we prompt the LLM to extract relevant evidence from the retrieved context. The LLM must isolate key sections, such as paragraphs, tables, and figure details, that are most likely to address the query and verbalize them in a structured form. This curation is crucial in a multi-document setup, where non-uniform sources introduce irrelevant, distracting, or adversarial content. Accurately identifying relevant information enhances the model's reasoning abilities by filtering out noise and helps mitigate LLM hallucinations.

2. Chain of Thought Reasoning: Extracting reasoning chains from multi-document artefacts can help contextualize curated evidence for final answer generation. We utilize Chain-of-Thought (CoT) (Wei et al., 2022) reasoning to link individual pieces of evidence that form a coherent step-by-step narrative, ensuring that the answer is not

only accurate but also logically derived from the evidence, leading to more robust and reliable responses.

3. Answer Generation: By leveraging insights from curated, contextually relevant evidence and applying CoT reasoning processes, the answer generation step produces responses that are both precise and well-justified. Additionally, we use targeted prompts to guide the LLM about the appropriate format for answer generation as per the question type.

5.2 Modality Fusion

The modality fusion stage is a key contribution in VisDoMRAG which differentiates it from simpler multimodal approaches. This stage takes as input the outputs from both the textual and visual pipelines, including the curated evidence, reasoning chains, and generated answers. The fusion process is orchestrated by prompting an LLM to evaluate the consistency between the reasoning chains produced by the textual and visual pipelines. This idea is inspired by self-consistency in CoT (Wang et al., 2023), which leveraged multiple thought-chains and derives an answer based on the consistency of the individual chains' results. Consistency constraint prompting is crucial for identifying and resolving any discrepancies, contradictions and filling in reasoning gaps that may arise from the separate processing of different modalities. When inconsistencies are detected, the LLM is tasked with reconciling the differences, potentially by re-evaluating the evidence or adjusting the reasoning steps. This process ensures that the final answer integrates information from both modalities in a coherent and logically consistent manner.

6 Experiments

In our experiments, we first evaluate different retrieval and indexing models on our benchmark, followed by end-to-end QA evaluation using the identified optimal retrieval models with different LLMs. The experiments, baselines and evaluation are discussed below:

6.1 Retrieval

Baselines: We use popular text based retrieval models: BM25 (Robertson et al., 1995) a statistical baseline, and , MPNet (Song et al., 2020), MiniLM (Wang et al., 2020), and BGE-1.5 (Xiao et al., 2023), which represent SoTA dense retrieval baselines. Text extraction from PDF documents is performed using PyTesseract. The extracted text is then segmented into 3000-character chunks using the recursive-split method (Sarmah et al., 2023), with a 10% overlap to mitigate information loss.

For visual retrieval, we utilize recent advances late interaction based multi-vector retrieval models built on top of LLMs (Faysse et al., 2024), namely ColPali and ColQwen2, which have PaliGemma (Beyer et al., 2024) and Qwen2 (Yang et al., 2024) as their base LLMs. Readers are encouraged to refer to the appendix for further details of these models.

Evaluation: Evidence extraction is assessed using ANLCS between ground truth evidence and retrieved chunks/pages. Document identification evaluates the retrievers' ability to select the correct source document in a multi-document setup. We report the rate of instances where the ground truth document is the source of the majority of the retrieved context.

6.2 End-to-End QA

We use the best text and visual retrieval models from the retrieval experiments for End-to-End QA evaluation.

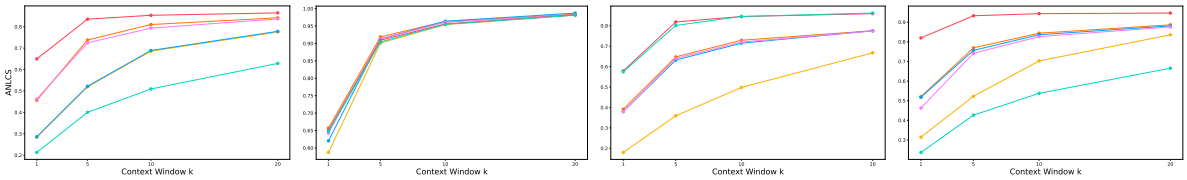
Baselines: We benchmark our method using LLMs capable of handling multi-image inputs and long context. To this extent, we include two off-the-shelf models Gemini-1.5-Flash (Reid et al., 2024), and ChatGPT-4o (OpenAI, 2024), as well as Qwen2-VL-7B-Instruct (Yang et al., 2024), an open-source LLM with visual and long context capabilities. We evaluate these LLMs in four approaches: 1. Long Context: where text content of all documents queries for a sample is passed as context, and 2. TextualRAG, 3. VisualRAG, and, 4.VisDoMRAG as described in Section 5.

Evaluation: For PaperTab, we borrow the modified implementation of Word Overlap F1 from (Hui et al., 2024), which takes into account different answer types (binary, short text). For all other datasets, we report the Word Overlap F1, which serves as a flexible metric to evaluate different answer types.

7 Results

7.1 Retrieval Evaluation on VisDoMBench

Fig. 3 presents the performance of various retrieval models in extracting evidence from documents, evaluated using the Averaged Normal-



(a) PaperTab (b) FetaTab (c) SciGraphQA (d) SPIQA

Figure 3: Comparison of retrieval performance across datasets, for benchmarked retrievers (BM25, MiniLM, MPNet, BGE1.5, ColPali, ColQwen), at different context window lengths, varying $k \in [1, 5, 10, 20]$.

Baseline	LLM	PaperTab	FetaTab	SciGraphQA	SPIQA	SlideVQA	Average
Long Context	Qwen2-VL	8.23	23.1	16.74	9.93	2.46	12.09
	Gemini	27.62	62.02	22.1	38.82	13.47	34.89
	GPT4o	28.37	60.03	24.12	36.3	15.06	34.53
Text RAG	Qwen2-VL	25.33	57.56	26.75	39.77	8.82	32.87
	Gemini	33.6	63.86	26.48	42.33	10.3	37.52
	ChatGPT4o	37.34	60.82	29.74	42.8	15.97	39.23
Visual RAG	Qwen2-VL	27.37	58.57	28.13	42.81	38.42	41.79
	Gemini	29.23	52.82	23.56	41.43	51.96	43.86
	ChatGPT4o	42.01	61.89	31.12	43.28	66.82	53.50
VisDoMRAG	Qwen2-VL	29.89	59.24	27.98	42.8	39.77	42.93
	Gemini	39.66	60.89	25.82	41.03	52.74	48.58
	ChatGPT4o	**44.11**	**63.28**	**31.36**	**44.09**	**67.22**	**54.68**

Table 3: Performance of our approach, `VisDoMRAG`, compared to baseline approaches on `VisDoMBench`. `VisDoMRAG` outperforms long-context LLM, visual and text-only RAG baselines.

Retriever	PaperTab	FetaTab	SciGraphQA	SPIQA	SlideVQA	Average
BM25	65.51	84.00	72.73	88.23	**98.55**	81.80
MiniLM	65.51	88.85	91.65	61.06	0.73	61.56
MPNet	90.18	89.71	91.40	95.84	0.73	73.57
BGE1.5	96.81	94.00	90.91	**98.43**	81.85	92.40
ColPali	9.81	**97.71**	95.28	11.43	97.64	54.15
ColQwen2	**97.61**	96.86	**95.58**	96.85	97.82	**96.94**

Table 4: Comparison of performance in source document identification, at $k = 5$.

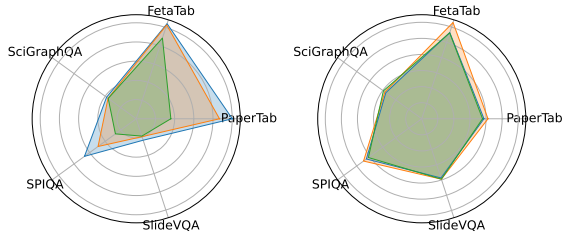
ized Longest Common Subsequence (ANLCS) between retrieved evidence and ground truth evidence, for different context window lengths ($k = [1, 5, 10, 20]$). Based on a threshold of ANLCS = 0.7, we use a context window of $k = 5$, $k = 7$ for Visual RAG and Textual RAG, respectively, with ColQwen2 and BGE-1.5 as the visual and textual retrievers. ColQwen2 outperforms other retrieval baselines across different datasets due to the presence of a strong LLM backbone (Qwen2).

Table 4 evaluates the retriever performance in identifying the correct source document, presenting the proportion of queries with accurate document retrieval for $k = 5$. A document is considered correctly retrieved if at least $\lceil k/2 \rceil$ of the retrieved documents correspond to the ground truth source documents. We observe that ColQwen2 is better than the next closest BGE1.5 model by 4.5%. Notably, we observe a substantial performance gap in this metric for SlideVQA, with visual models significantly outperforming text-only models. BM25 exhibits better performance than text-only models in this case, as slides typically contain sparse text, often comprising keywords that directly match between the query and context. Conversely, neural models struggle to capture semantic information effectively, as the textual content lacks complete sentences, limiting their ability to exploit contextual meaning.

7.2 End-to-End Evaluation

Table 3 presents the comparative performance of VisDoMRAG against Visual RAG, Textual RAG, and Long Context methods across multiple LLMs, including Qwen2VL (7B), Gemini Flash, and GPT-4. The results indicate that VisDoMRAG consistently achieves superior performance over the baseline methods across datasets, with performance gains ranging from 2.1-21.6% (PaperTab), 0.67-36.14% (FetaTab), 0.24-11.24% (SciGraphQA), 0.81-32.87% (SPIQA), 0.40-52.16% (SlideVQA). Additionally, within each baseline method for most datasets, we observe a positive correlation between model size and performance, which aligns with established expectations in LLM scaling behavior (Hestness et al., 2017).

Textual vs Visual RAG: In comparing the performance of textual RAG vis-à-vis visual RAG, we observe that visual RAG consistently outperforms textual RAG. This behaviour can be explained on



(a) Long Context (b) VisDoMRAG

Figure 4: Comparative performance between Long Context and VisDoMRAG (averaged across LLMs) evaluated on different ranges of number of pages $\bar{p} = \sum_{d \in \mathcal{D}} |d|$, with Low ($\bar{p} \leq 100$), Medium ($100 < \bar{p} \leq 150$), and High ($150 \leq \bar{p}$) volumes.

the basis of our dataset composition which predominantly consists of visually-rich content, and visual RAG is able to leverage visual information directly. However, the performance difference is less pronounced in scientific figure datasets such as SciGraphQA and SPIQA due to the text-rich nature of scientific papers, where figures are often accompanied by detailed descriptions within the text and captions, particularly emphasizing key results and structural details. In contrast, we see a substantial performance gap between textual and visual RAG for SlideVQA, as slides typically lack extensive textual descriptions of visualizations, forcing the visual modality to be the primary source for answering questions. Additionally, we find that Gemini often performs better in the textual modality compared to the visual modality across most datasets. This disparity could be attributed to factors such as linguistic bias (Niu et al., 2021; Wang et al., 2024a) or visual hallucination (Ghosh et al., 2024), where the model's visual perception may be less reliable than its linguistic capabilities.

Effect of Long-Context LLMs: We observe that VisDoMRAG has the ability to significantly enhance the performance of smaller models, as seen from Qwen2VL. This improvement can be attributed to its ability to integrate visual and textual reasoning, compensating for the weaker long-context understanding and visual perception. The long-context LLM baselines prove to be less effective in our setup due to the high token count and the nature of the task, which requires retrieval of specific, localized evidence—essentially a needle-in-the-haystack problem. The combination of modalities in VisDoMRAG mitigates these challenges, resulting in more robust answer generation, as reflected in the results.

Effect of Increasing Page Count: Figure 4 evaluate the performance of different approaches averaged across LLMs, segmented by the volume of pages associated with each query. As anticipated, long-context models exhibit significant performance drop with increasing number of pages in the collection. Contrastively, our multimodal RAG approach shows consistent QA performance even at high page counts as it is able to constrain the amount of context the LLM needs to process to answer the question effectively.

Qualitative Examples: Fig 5 represents a qualitative example from the PaperTab dataset, where VisDoMRAG effectively uses reasoning chains and answers from unimodal RAG outputs to synthesize the correct answer. More qualitative results are presented in the Appendix.

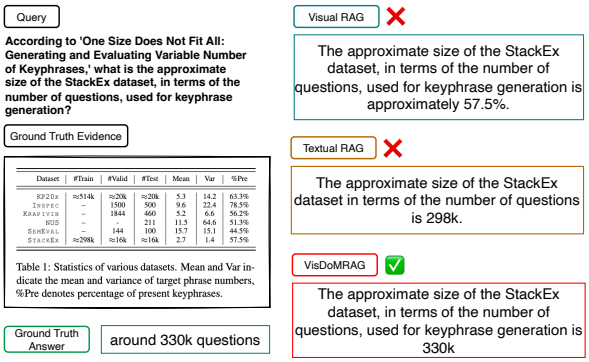


Figure 5: Qualitative example from the PaperTab dataset, comparing VisDoMRAG with Unimodal RAG strategies.

8 Conclusion and Future Work

In this work, we introduced VisDoMBench, the first QA dataset designed to evaluate multi-document systems incorporating visually rich elements such as tables, charts, and slides. By targeting documents that require both textual and visual comprehension, VisDoMBench offers a novel benchmark to assess the capability of multimodal retrieval systems. We also presented VisDoMRAG, a multimodal Retrieval-Augmented Generation approach that fuses visual and textual pipelines using consistency-constrained modality fusion. This method demonstrated a significant improvement over traditional long context, textual, and visual RAG by 12-20%. While the current work focuses on RAG in multimodal multi-doc settings, future work will extend this approach to include reasoning through end-to-end trained models, especially in low-resource settings.

9 Ethics Statement

We use publicly available datasets in this research. The identities of human evaluators remain confidential, and no personally identifiable information (PII) is used at any stage of our experiments. Our work is solely intended for document QA applications. For a deeper understanding of potential risks and mitigation strategies in LLM safety, we direct users to relevant works by (Kumar et al., 2024; Cui et al., 2024; Luu et al., 2024).

10 Limitations

Despite the advancements presented in this study, several limitations warrant consideration:

(1) Text Extraction and Document Parsing: A key argument for the efficacy of visual retrieval methods is the elimination of text extraction and document parsing pipelines (Faysse et al., 2024). However, our approach retains this overhead, which may introduce additional complexity and processing time.

(2) Multiple LLM calls: Our methodology necessitates multiple LLM calls; specifically, we make three LLM calls per query. While this approach may not be optimal, it is still more cost-effective than utilizing long-context models.

(3) Hallucinations: As with all works involving large language models (LLMs), our approach is subject to inherent limitations related to AI safety and the risk of hallucination. These issues can affect the reliability and accuracy of the generated outputs and underscore safety risks, highlighting the need for ongoing research and refinement in the field of AI to mitigate these challenges.

Additionally, unlike previous visual QA research, which typically required models to answer questions based solely on visual data, our framework incorporates document context. This inclusion allows for relevant textual information from other sections of the paper to contribute to the query response. However, this reliance on document context represents a limitation common to all visually rich document QA datasets, as it challenges the isolation of visual performance testing. Nonetheless, this characteristic may not be entirely detrimental; in fact, it more accurately reflects the complexity of real-world systems where multimodal information is often interdependent.

References

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Deng Cai, Yan Wang, Victoria Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2018. Skeleton-to-response: Dialogue generation guided by retrieval memory. *arXiv preprint arXiv:1809.05296*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *Preprint*, arXiv:2401.05778.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as images? exploring the strengths and limitations of llms on multimodal representations of tabular data. *arXiv preprint arXiv:2402.12424*.

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

Manuel Faysse, Hugues Sibille, Tony Wu, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Sreyan Ghosh, Chandra Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2024. Vdgd: Mitigating lvlm hallucinations in cognitive prompts by bridging the visual perception gap.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. 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 3170–3180.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Frederick Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *ArXiv*, abs/1712.00409.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

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

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms. *Preprint*, arXiv:2401.12273.

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

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Quan Khanh Luu, Xiyu Deng, Anh Van Ho, and Yorie Nakahira. 2024. Context-aware llm-based safe control against latent risks. *Preprint*, arXiv:2403.11863.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

OpenAI. 2024. Hello, gpt-4o! `https://openai.com/index/hello-gpt-4o/`.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. MultiTabQA: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6322–6334, Toronto, Canada. Association for Computational Linguistics.

Bo Pang, Erik Nijkamp, Wojciech Kryscinski, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. 2023. Long document summarization with top-down and bottom-up inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1267–1284, Dubrovnik, Croatia. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071.

Hao Peng, Ankur P Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. *arXiv preprint arXiv:1904.04428*.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Bhaskarjit Sarmah, Tianjie Zhu, Dhagash Mehta, and Stefano Pasquali. 2023. Towards reducing hallucination in extracting information from financial reports using large language models. *Preprint*, arXiv:2310.10760.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

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

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. 2024b. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. *arXiv preprint arXiv:2406.11230*.

Minzheng Wang, Longze Chen, Cheng Fu, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024c. Leave no document behind: Benchmarking long-context llms with extended multi-doc qa. *Preprint*, arXiv:2406.17419.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556.

Lei Zhang, Yunshui Li, Ziqiang Liu, Junhao Liu, Min Yang, et al. 2023. Marathon: A race through the realm of long context with large language models. *arXiv preprint arXiv:2312.09542*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. 2024.
infty bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*.

A Appendix

A.1 Baselines

A.1.1 Retrieval Models

BM25 BM25 (Robertson et al., 1995) is a widely adopted term-based ranking function based on the probabilistic information retrieval model. It calculates the relevance of a document to a given query by considering term frequency, inverse document

frequency, and document length normalization. BM25 is effective for sparse text retrieval tasks, making it a standard baseline in information retrieval evaluations. We use the Python `rank_bm25` implementation for our experiments.

MiniLM MiniLM (Wang et al., 2020) is a lightweight, transformer-based model designed for efficient knowledge distillation. It compresses the knowledge of larger pre-trained models into a smaller architecture while maintaining competitive performance in natural language understanding tasks. MiniLM is used in retrieval tasks due to its ability to balance computational efficiency and accuracy. We use the `sentence-transformers/all-MiniLM-L6-v2` implementation in our experiments.

MPNet MPNet (Song et al., 2020) is a transformer-based model that leverages permuted language modeling for pre-training, which helps it capture contextual information more effectively than traditional masked language models. It excels in a variety of natural language processing tasks, including text retrieval, due to its robust contextual embeddings and representation learning capabilities. We use the `sentence-transformers/all-mpnet-base-v2` implementation in our experiments.

BGE-1.5 The BGE model family is based on a BERT-like architecture and a three-stage training process, which collectively enhance its adaptability and generalization capabilities. Pre-training is performed on large-scale plain text corpora using a tailored MAE-style approach, effectively encoding polluted text and reconstructing the clean version. The model then undergoes contrastive learning with in-batch negative sampling, leveraging large batch sizes to improve embedding discriminativeness. Finally, task-specific fine-tuning is employed using labeled datasets, applying instruction-based prompts and advanced negative sampling techniques to better accommodate diverse task types. We use the `BAAI/bge-base-en-v1.5` model in our experiments, which is their large english model, version 1.5.

ColPali, ColQwen2 ColPali (Faysse et al., 2024) performs late interaction retrieval on document embeddings generated directly from document page images using Vision-Language Models (VLMs). By passing the document images through PaliGemma (Beyer et al., 2024), ColPali uses the projected token embeddings to index the document pages, eliminating the need for OCR or document parsing. The multimodal alignment learned by VLMs allows both text queries and document image embeddings to exist in a shared semantic vector space, enabling more precise and efficient retrieval. ColQwen2 is a similar model with Qwen2 (Yang et al., 2024) as the base VLM. We used the `vidore/colpali-v1.2`, `vidore/colqwen2-v0.1` implementations for our experiments.

A.1.2 LLMs

We used `Qwen/Qwen2-VL-7B-Instruct`, `chatgpt-4o-latest` and `gemini-1.5-flash` in our experiments. For ChatGPT4o and Gemini, we set the temperature as 0.5, and use the default hyperparameters. For `Qwen2-VL`, the pixel range is set to $[256 \times 28 \times 28, 640 \times 28 \times 28]$. For Long Context evaluation, we use `Qwen/Qwen2-7B-Instruct` because of the implementation availability of long context inference using YaRN (Peng et al., 2023). We report results on a single run of experiments.

A.2 Datasets

The datasets use in our benchmark are described below. Fig 6-10 represent the distribution of pages per query in all the data splits.

FetaTab FetaTab is derived from UDA (Hui et al., 2024), which sources its data from FetaQA (Nan et al., 2022). Many source datasets provide only segmented and partial content, lacking complete documents. To resolve this, UDA conducted a thorough source-document identification process, verifying and collecting the complete original document files based on metadata or content fragments. This was followed by rigorous matching and reorganization to form complete triplet data pairs consisting of document-question-answer. Additionally, UDA categorizes queries based on the source of factual evidence, filters out Q&As without available answers, converts token-based data patterns to natural language, unifies data formats and structures across datasets, and designs specific LLM prompts tailored for each dataset after experimental trials. FetaTab is licensed under the CC-BY-SA-4.0 license.

PaperTab PaperTab is also sourced from UDA (Hui et al., 2024), which obtains its data from the QASPER (Dasigi et al., 2021) dataset. Similar to the process described for FetaTab, UDA emphasizes the necessity of ensuring the integrity of

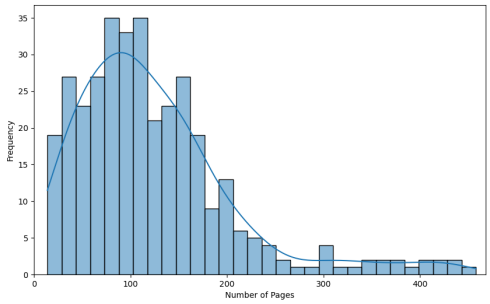


Figure 6: Distribution of pages per query for FetaTab.

original documents for effective document analysis. This involves a comprehensive process of identifying, verifying, and collecting complete original document files, followed by matching and reorganization to create document-question-answer triplets. UDA also categorizes queries, filters out unanswered Q&As, converts data patterns to natural language, unifies data formats, and designs specific LLM prompts for each dataset based on experimental evaluations. PaperTab is released under the CC-BY-SA-4.0 license.

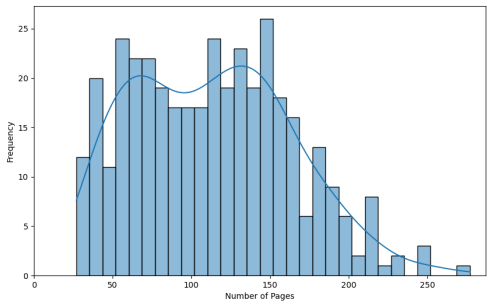


Figure 7: Distribution of pages per query for PaperTab.

SPIQA SPIQA (Pramanick et al., 2024) is a large-scale and challenging question-answering dataset that focuses on figures, tables, and text paragraphs extracted from scientific research papers across various computer science domains. The dataset encompasses a diverse array of visual elements, including plots, charts, schematic diagrams, and result visualizations. SPIQA consists of 270K questions divided between training, validation, and three different evaluation splits. To ensure the highest quality and reliability, SPIQA employs both automatic and manual curation methods. The dataset is released under the CC-BY-SA-4.0 license, allowing for broad use while ensuring proper attribution.

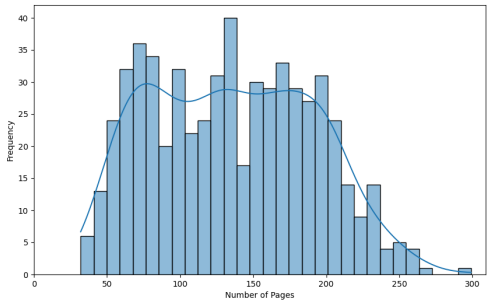


Figure 8: Distribution of pages per query for SPIQA.

SciGraphQA SciGraphQA (Li and Tajbakhsh, 2023) is a synthetic multi-turn question-answer dataset centered on academic graphs, representing a significant advancement in the field of visual question answering. At 13 times larger than the previous largest dataset, ChartVQA, it stands as the largest open-sourced chart VQA dataset with non-synthetic charts. The dataset was constructed from 290,000 Computer Science and Machine Learning papers published on ArXiv between 2010 and 2020, with the help of Palm-2 generating 295,000 samples of open-vocabulary multi-turn question-answer dialogues about the graphs. Each dialogue is contextualized with the paper title, abstract, relevant paragraphs, and rich contextual data from the graphs, achieving an average of 2.23 question-answer turns per graph. SciGraphQA is released under the MIT license.

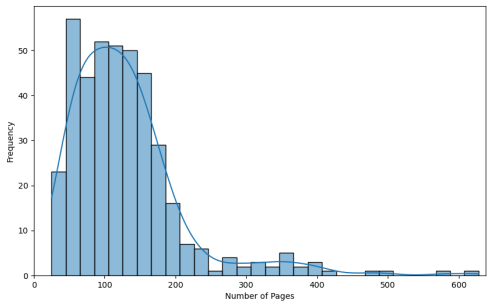


Figure 9: Distribution of pages per query for SciGraphQA.

SlideVQA SlideVQA (Tanaka et al., 2023) is a multi-image document VQA dataset that contains over 2,600 slide decks, comprising more than 52,000 slide images and 14,500 questions regarding the slide content. This dataset requires complex reasoning skills, including single-hop, multi-hop, and numerical reasoning. It also provides anno-

tated arithmetic expressions for numerical answers, enhancing numerical reasoning capabilities. More details about the dataset can be found under the license at this link.

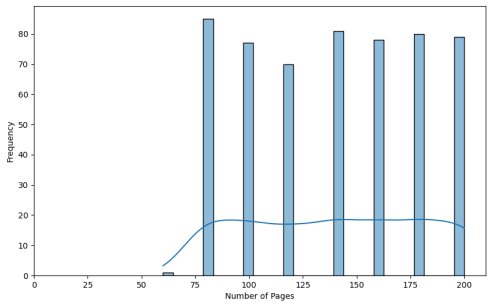


Figure 10: Distribution of pages per query for SlideVQA.

A.3 Examples

A.3.1 Query Augmentation

Tables 5 and 6 represent examples of query augmentation during dataset construction for PaperTab and SciGraphQA.

A.3.2 End-to-End QA Examples

Figures 11-16 illustrate End-to-End QA examples across the five datasets, demonstrating the performance of different LLMs.

In Figure 11, we analyze an example from the PaperTab dataset using Qwen2VL. VisualRAG fails in this instance by selecting the incorrect column for computation during reasoning. Conversely, TextualRAG identifies the correct column but overlooks samples from the test and validation sets. VisDoMRAG evaluates both outputs and produces the correct answer, demonstrating its ability to refine responses across modalities.

Figure 12 presents an example from the FetaTab dataset, where Gemini is employed as the base LLM. Here, TextualRAG successfully generates the correct answer by accurately verbalizing the OCR-processed table during evidence retrieval. Although VisualRAG underperforms in this case, VisDoMRAG integrates the evidence effectively, providing the overall correct answer.

In Figure 13, an example from SciGraphQA shows both Visual and Textual RAG producing correct responses. Consequently, VisDoMRAG corroborates the correct answers, confirming the alignment between both modalities.

Figure 14 depicts a scenario from the SPIQA dataset where VisDoMRAG fails to provide the correct answer. This error arises from its bias towards the longer response generated by VisualRAG, which itself is incorrect.

Lastly, Figure 16 showcases an example from the SlideVQA dataset. In this case, TextualRAG fails to capture the necessary evidence, whereas VisualRAG successfully employs multi-hop reasoning across two slides to derive the correct answer. VisDoMRAG recognizes the precision in VisualRAG's response, favoring its consistency with the question's context.

A.4 LLM Prompts

Fig. 17 - 19 represent prompt templates used in our experiments for query augmentation, baselines and VisDoMRAG.

A.5 Human Review Process

We addressed the challenge of trivial or underspecified queries in some datasets by augmenting the queries using ChatGPT4o and relevant context, including the title and abstract of the research paper, the relevant figure's caption, and other available metadata. We employ a human reviewer to assess the quality of the generated queries and select one of the queries or reject all queries. The reviewer is a graduate student who is paid at the hourly rate for Graduate Assistants at the university where they are a student. Fig 20 gives a brief of the instructions as well as the evaluation rubric given to the reviewer.

A.6 Computational Resources

Table 7 describes the Computational Resources used for running this paper's experiments.

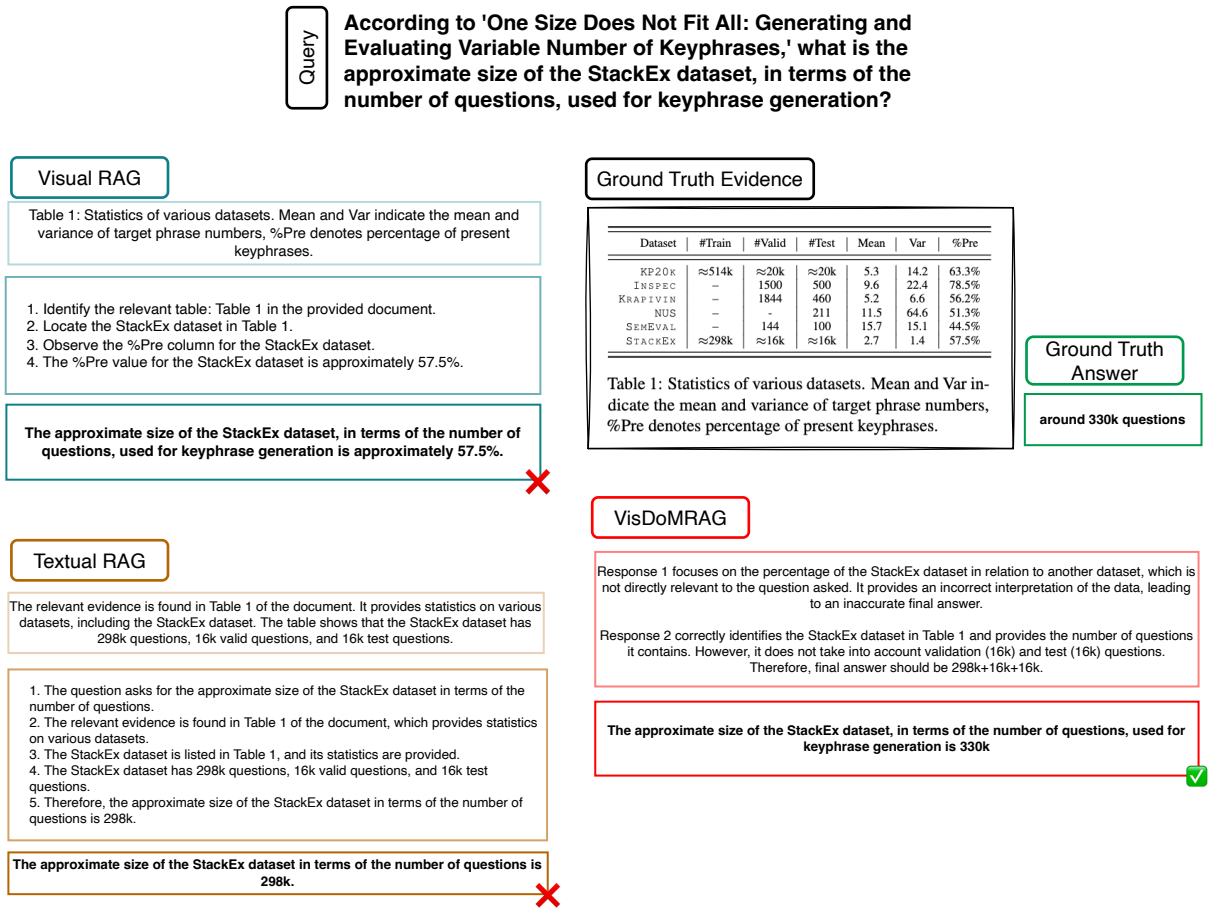


Figure 11: Qualitative example from the PaperTab dataset, comparing VisDoMRAG with unimodal RAG strategies, with Qwen2VL as the base LLM.

Original Query	**Augmented Query**
What baselines did they consider?	What baseline approaches using state-of-the-art PDTB taggers were employed for the evaluation of causality prediction in the automatic causal explanation analysis pipeline?
What is the average length of the claims?	What is the average token count of claims as reported in Table 2 of the PERSPECTRUM dataset?
What was the performance on the self-collected corpus?	What F1 scores did the CM-Net achieve for intent detection and slot filling on the CAIS dataset as detailed in Table 6?
Do they test their framework performance on commonly used language pairs, such as English-to-German?	Does the paper report results for English-to-German translation in a simulated under-resourced scenario using their proposed multilingual NMT framework?

Table 5: Example of query augmentation from PaperTab dataset.

Original Query	**Augmented Query**
What is the main difference between the two scheduling algorithms compared in this graph?	In this paper, what scheduling algorithms are analyzed in Figure 8 for WCRT comparison?
What does the phase diagram indicate about the stability of the different phases?	What does Figure 4.18 reveal about the phase boundaries for different choices of Jt and k?
What does the graph show about the impact of the load-changing attack on the frequency of the system?	What does the figure show about frequency limits during the 2019 and 2020 load-changing attacks?
What are some of the implications of the graph for the design of fuzzing tools?	What relationship does Fig. 3 suggest between performance and resources in fuzzing tools?

Table 6: Example of query augmentation from SciGraphQA dataset.

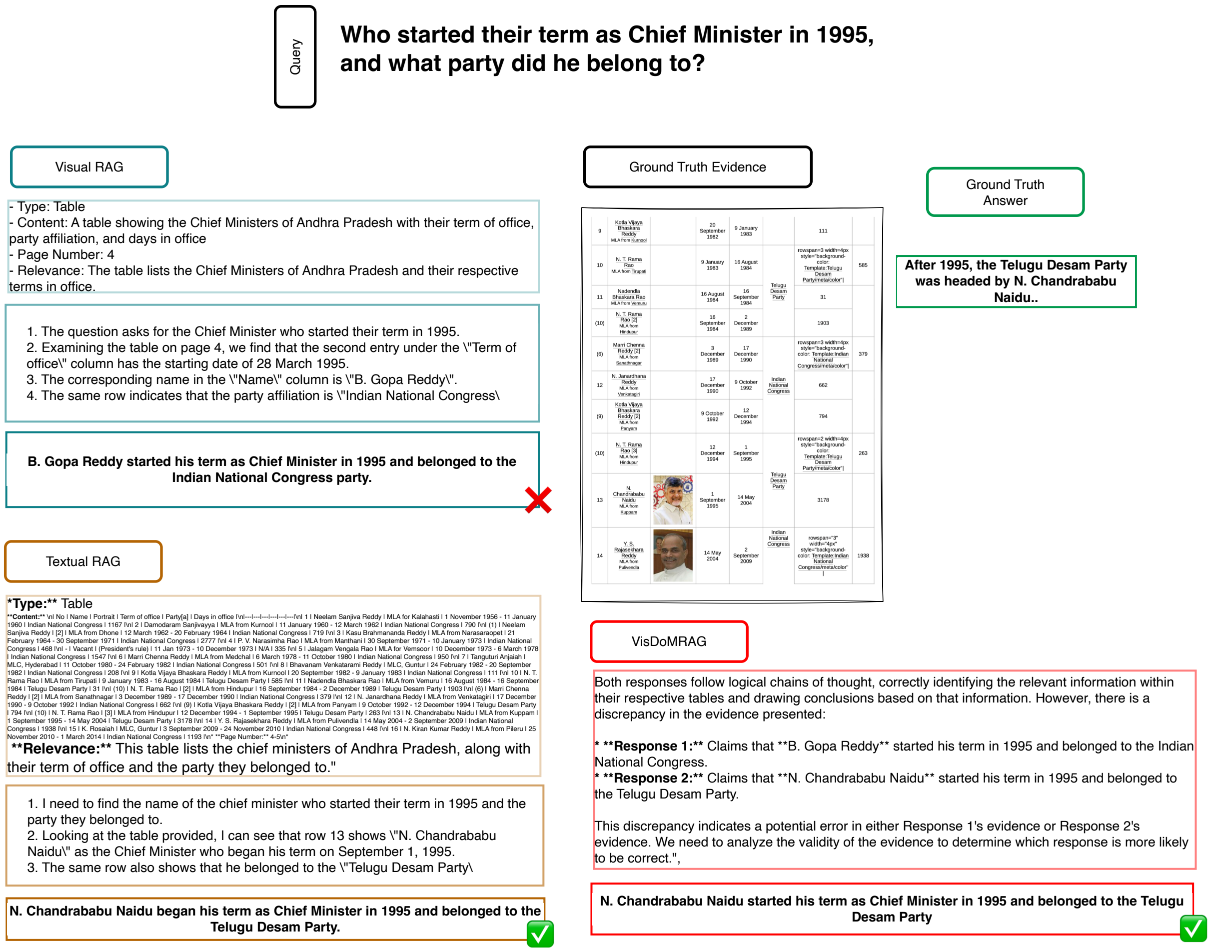


Figure 12: Qualitative example from the FetaTab dataset, comparing VisDoMRAG with unimodal RAG strategies, with Gemini as the base LLM.

Query

What observation about aggregate gap in relation to DBS number is presented in the paper's Fig. 4?

Visual RAG

Fig. 4
The figure shows the aggregate gap achieved through the proposed scheme versus the number of DBSs.

1. **Observation**: The figure illustrates that as the number of DBSs increases, the aggregate gap decreases.
2. **Explanation**: This indicates that the proposed scheme becomes more efficient in terms of aggregate gap as more DBSs are utilized.
3. **Implication**: The performance of the proposed scheme improves with an increase in the number of DBSs.

The paper's Fig. 4 presents the observation that the aggregate gap decreases as the number of DBSs increases. This suggests that the proposed scheme becomes more efficient in terms of aggregate gap with an increase in the number of DBSs.

Ground Truth Evidence

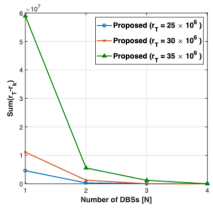


Fig. 4. Plot of the aggregate gap achieved through the proposed scheme vs. the number of DBSs.

Ground Truth Answer

The graph shows that the aggregate gap decreases as the number of DBSs increases. This is because as more DBSs are used, the communication rates of the terminals are closer to the target rate. This is because the DBSs can provide more resources to the terminals, which allows them to communicate at higher rates.

Textual RAG

The observation about the aggregate gap in relation to DBS number is presented in the paper's Fig. 4. The figure illustrates the aggregate gap between the communication rates and the target rate of terminals with respect to the change in the number of utilized DBSs.

1. The question asks for an observation about the aggregate gap in relation to DBS number.
2. The relevant figure is Fig. 4, which shows the aggregate gap between the communication rates and the target rate of terminals.
3. The figure demonstrates the benefits of utilizing DBSs and the benefits of optimizing the 3D positions of the DBSs.
4. The figure also shows that as the number of DBSs increases, the performance difference between the proposed scheme and the compared schemes also increases.
5. This implies that the proposed scheme could be utilized more efficiently in the cases where most terminals in the coverage of the MBS require wireless services with high data rates.

The observation about the aggregate gap in relation to DBS number presented in the paper's Fig. 4 is that as the number of DBSs increases, the performance difference between the proposed scheme and the compared schemes also increases. This implies that the proposed scheme could be utilized more efficiently in the cases where most terminals in the coverage of the MBS require wireless services with high data rates.

VisDoMRAG

Both responses provide logical chains of thought and final answers based on the evidence presented in the paper's Fig. 4. Response 1 focuses on the visual representation of the figure, while Response 2 provides a more detailed textual explanation of the figure's content. Both responses are consistent with the question and the evidence provided in the figure.

The paper's Fig. 4 presents the observation that the aggregate gap decreases as the number of DBSs increases. This suggests that the proposed scheme becomes more efficient in terms of aggregate gap with an increase in the number of DBSs.

Figure 13: Qualitative example from the ScigraphQA dataset, comparing VisDoMRAG with unimodal RAG strategies, with Qwen2VL as the base LLM.

Query

How does the shape of the IRLS weight function change as the shape parameter α increases?

Visual RAG

Element 1:
- **Type:** Paragraph
- **Content:**
 - "Curiously, these IRLS weights resemble a non-normalized form of Student's t-distribution."
 - "A visualization of our loss's Ψ-functions and weight functions for different values of α can be seen in Figure 7."
- **Page Number:** 2
- **Explanation:** This paragraph indicates that the IRLS weight function changes as the shape parameter α increases and directs us to Figure 7 for visualization.

Element 2:
- **Type:** Figure
- **Content:** Figure 7, which shows "Our general loss's IRLS weight function (left) and Ψ-function (right) for different values of the shape parameter α ."
- **Page Number:** 2
- **Explanation:** The figure visually represents how the IRLS weight function changes with different values of α , which is crucial to answering how the shape of the weight function changes as α increases.

Element 3:
- **Type:** Paragraph
- **Content:**
 - "Ψ(z, α) = (1 – z) for α = 2"
 - "Ψ(z, α) is not defined when α ≥ 2 because for those values the loss is no longer robust, and so it is not well described as a process that rejects outliers."
- **Page Number:** 2
- **Explanation:** This section explains that for values of α equal to or greater than 2, the loss becomes less effective at rejecting outliers, implying changes in the weight function behavior.

Element 4:
- **Type:** Paragraph
- **Content:**
 - "The shape of the derivative gives some intuition as to how α affects behavior when our loss is being minimized by gradient descent or some related method."
 - "For all values of α the derivative is approximately linear when |x| << c, so the effect of a small residual is always linearly proportional to that residual's magnitude."
 - "As α becomes more negative, and as α approaches –∞ an outlier whose residual magnitude is larger than 3c is almost completely ignored."
- **Page Number:** 5
- **Explanation:** This paragraph explains how the shape parameter α affects the behavior of the weight function, particularly for outliers.

1. **Understanding the Role of α in the Weight Function:** From the paragraph on page 2, we learn that the IRLS weight function is dependent on α , and its behavior can be visually analyzed using Figure 7. Moreover, as α increases, the function becomes less robust at rejecting outliers.

2. **Analyzing Figure 7:** The left side of Figure 7 shows the IRLS weight function for different values of α . As α increases (from negative to positive values), the weight function becomes flatter around the origin, indicating that higher α values reduce the penalization of residuals, especially for larger residuals.

3. **Impact on Outliers:** From the paragraph on page 5, we know that as α becomes more negative, the weight function heavily penalizes large residuals (outliers), but as α increases, the "influence" of outliers diminishes, making the weight function flatter and less responsive to large residuals.

4. **α = 2 and Beyond:** For α values greater than or equal to 2, the weight function is no longer robust, meaning it does not effectively reject outliers, as described on page 2.

As the shape parameter α increases, the IRLS weight function becomes flatter, indicating that the penalization of large residuals (outliers) decreases, making the function less robust to outliers. ✗

Ground Truth Evidence

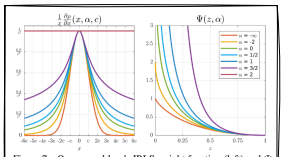


Figure 7. Our general loss's IRLS weight function (left) and Ψ-function (right) for different values of the shape parameter α.

Ground Truth Answer

The IRLS weight function becomes more peaked and concentrated around zero as the shape parameter α increases.

Textual RAG

1. **Figure and Text Description (Page 26)**:
 - "A visualization of our loss's IRLS weight functions for different values of α can be seen in Figure 7. [...] This curve is roughly piecewise-linear in [0,4] with a slope of ~1 at α = 0 and α = 4, but with a slope of ~10 at α = 2. When α > 4, the curve becomes logarithmic."
 - **Relevance**: This passage directly describes the shape of the IRLS weight function as the shape parameter α increases, providing key insights into how the function changes.

1. The passage describes the weight function for different values of α , starting with a piecewise-linear shape in the range [0, 4].
2. At α = 0 and α = 4, the slope is ~1, indicating a relatively flat response.
3. At α = 2, the slope is much steeper (~10), indicating a more pronounced response to changes in the input.
4. For α > 4, the function transitions to a logarithmic shape, indicating that the growth of the weight function becomes slower and sub-linear.

As the shape parameter α increases, the IRLS weight function transitions from a roughly piecewise-linear shape to a logarithmic shape, with a steeper slope at intermediate values like α = 2. ✓

VisDoMRAG

Both responses provide consistent reasoning on how the shape of the IRLS weight function changes as the shape parameter **α ** increases, but they approach the question from slightly different angles.

1. **Response 1:**
 - The evidence is drawn from a visual assessment (Figure 7) and various descriptive paragraphs about the behavior of the weight function as α changes.
 - The explanation clearly states that as the shape parameter α increases, the weight function becomes flatter, reducing the penalization of large residuals (outliers), hence making the function less robust to outliers.
 - The explanation considers specific behavior for very negative α values, and even points out that the weight function is no longer robust for α greater than or equal to 2.

2. **Response 2:**
 - Builds mostly on textual evidence and describes in more numerical detail the shape changes of the IRLS weight function (e.g., piecewise-linear with specific slopes at α = 0, 2, and 4, becoming logarithmic when α > 4).
 - This response emphasizes how the weight function becomes logarithmic as the shape parameter α exceeds 4.

There is no fundamental disagreement between the two responses. However, **Response 1** emphasizes the relationship to robustness to outliers and the visual aspects of the behavior change, while **Response 2** provides more numerical information but does not cover robustness or specific outlier effects as well as Response 1.

- **Content Completeness**: Response 1 provides a broader explanation, including different α ranges (negative α , small α , and large α) and discussions about outlier behavior.
- **Specificity of Behavior Across α Values**: Response 2 offers precise descriptions (slope changes and logarithmic behavior), but lacks the nuance about outliers and robustness.

As the shape parameter α increases, the IRLS weight function becomes flatter, indicating decreased penalization of large residuals (outliers), making the function less robust to outliers. ✗

Figure 14: Qualitative example from the SPIQA dataset, comparing VisDoMRAG with unimodal RAG strategies, with ChatGPT4o as the base LLM.

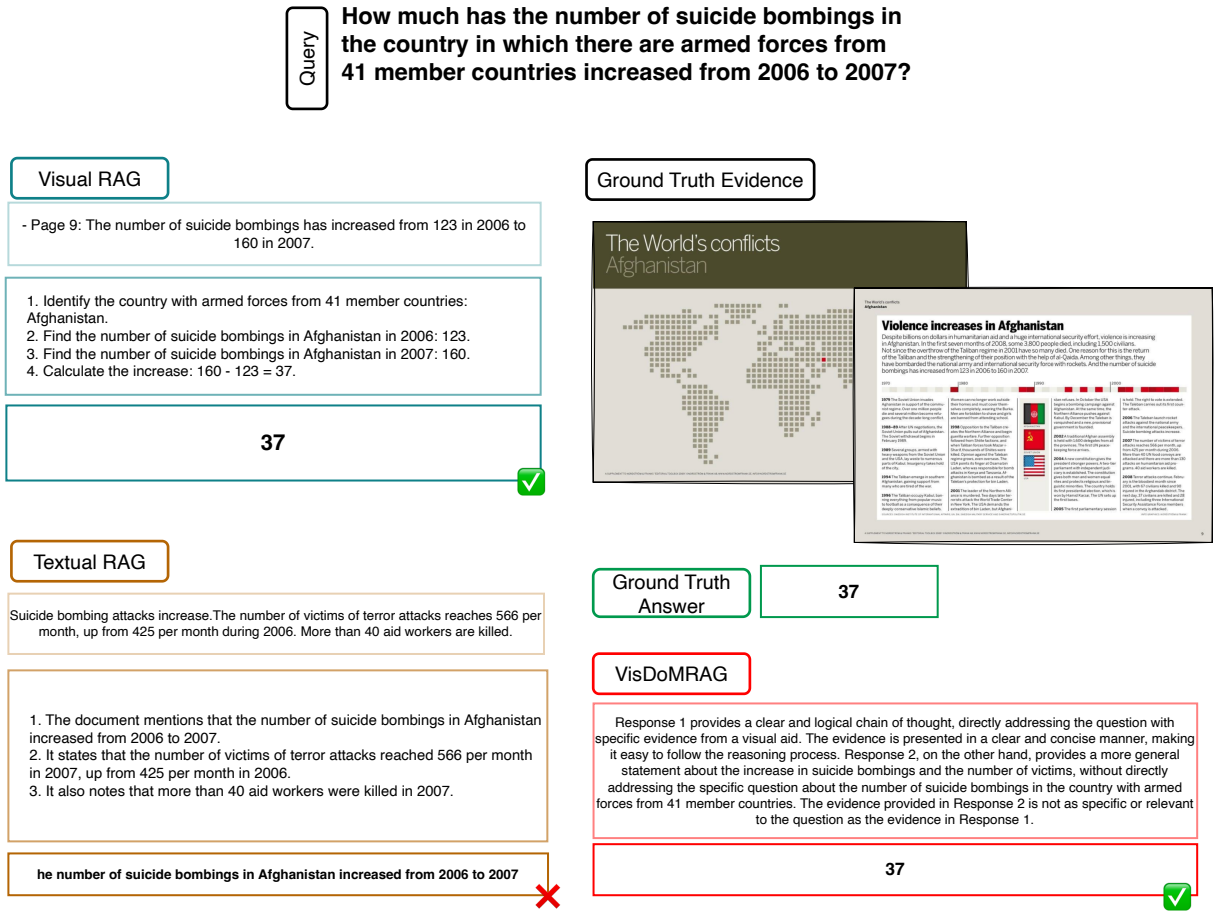


Figure 15: Qualitative example from the SlideVQA dataset, comparing VisDoMRAG with unimodal RAG strategies, with ChatGPT4o as the base LLM.

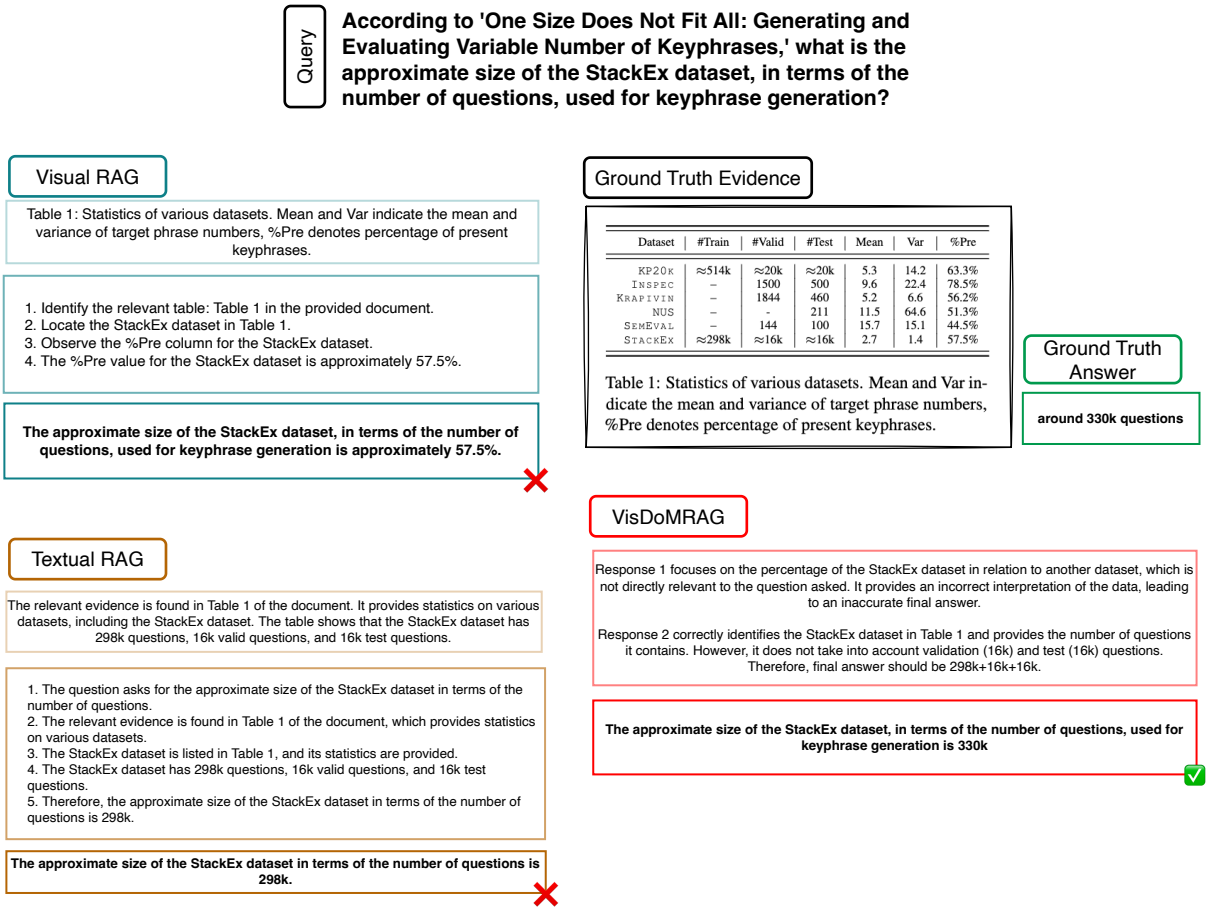


Figure 16: Qualitative example from the PaperTab dataset, comparing VisDoMRAG with unimodal RAG strategies, with Qwen2VL as the base LLM.

You are given a question about a figure in a research paper. Make that question more specific to the research paper, such that it can only be answered by that paper.

Question:

Metadata:
Paper Title:
Paper Abstract:
Figure Caption:
Other context (if any):

- Generate 5 questions that have the same answer as below and ask the question that the original question asks.
- Make sure to include some hints in the question like the paper, or what experiment/dataset is being asked about here (but no hint that significantly overlaps with the answer)
- Max 15-20 words long.
- Make the question uniquely specific to this paper and figure.

The answer to your question should be:

Format your response as a Python list.

Figure 17: Prompt Template used for Query Augmentation.

You are tasked with answering a question based on the relevant pages of a PDF document. Provide your response in the following format:

Evidence:

Chain of Thought:

Answer:

Instructions:

1. Evidence Curation:
- Extract relevant elements (such as paragraphs, tables, figures, charts) from the provided pages and populate them in the "evidence_curation" section.
- For each element, include the type, content, page number, and a brief explanation of its relevance.

2. Chain of Thought:
- In this section, list out each logical step you take to derive the answer, referencing the evidence where applicable.
- You should perform computations if you need to to get to the answer.

3. Answer:
- The answer should be a [short natural sentence/ just the answer without explanation/ could be sourced from the context].

Context:

Question:

Figure 18: Prompt Template used for Unimodal RAG and Long Context experiments.

Analyze the following two responses to the question:

Response 1:
Evidence:
Chain of Thought:
Final Answer:

Response 2:
Evidence:
Chain of Thought:
Final Answer:

- Response 1 is based on a visual q/a pipeline, and Response 2 is based on a textual q/a pipeline.

- Evaluate the answers from the two responses based on their chain of thought. You must try to check if both the chains of thoughts are consistent with respect to each other, the evidence provided and the final answer.

- If one of the responses has declined giving a clear answer, please weigh the other answer more unless there is reasonable thought to not answer, and both thoughts are inconsistent.

- Language of the answer should be short and direct, usually answerable in a single sentence, or phrase, similar to the language in the responses. You should give direct responses without explanation in the final answer.

Consider both chains of thought and final answers. Provide your analysis in the following format:

Analysis:
[Your detailed analysis here, evaluating the consistency of both the chains of thoughts, with respect to each other, the question and their respective answers, as well as validity of the evidence.]

Conclusion:
[Your conclusion on which answer is more likely to be correct, or if a synthesis of both is needed]

Final Answer:

Figure 19: Prompt Template used for VisDoMRAG.

Metric	Details
GPU Hours	100
GPU Specification	RTXA600
Number of GPU(S)	1
Max Model Parameters	7B

Table 7: Computationa Resources for VisDoM RAG experiments.

Instructions for Human Evaluator: Selecting the Optimal Query

When evaluating a list of AI-generated queries, it's essential to consider not only how well each query fits the context but also how natural it sounds and whether it maintains a balance between clarity and specificity. As you review the queries, your goal is to select the one that best matches the context, feels human-like, and doesn't over-include information from the answer.

In some cases, none of the queries may be optimal. Therefore, if you find that all the AI-generated queries are inadequate, you also have the option to **reject all of them** and instead return the **original query**. This ensures that you always have a fallback and aren't restricted to using the AI-generated versions.

To guide your decision-making, the table below outlines specific criteria and a rubric that will help you evaluate each query on its quality. You'll weigh the naturalness of the query, its relevance to the context, whether it's overly detailed, and its adaptability for reuse across different contexts.

Evaluation Rubric

Criteria	Score: 5 (Excellent)	Score: 4 (Good)	Score: 3 (Moderate)	Score: 2 (Weak)	Score: 1 (Poor)
Natural Sounding	Extremely natural and conversational	Generally natural with minor awkwardness	Somewhat natural but noticeably rigid	Clearly generated or awkward phrasing	Highly unnatural, difficult to interpret
Context Specificity and 1-1 Mapping	Perfectly matches context, specific and clear	Highly relevant, slight flexibility in meaning	Somewhat relevant, could apply to other contexts	Vague or general, could apply to many contexts	Completely irrelevant or too broad
Not Over-Augmented	Balanced, requires document to answer	Slightly detailed, but still needs document	Some unnecessary hints, but answerable	Too much detail, answer can be inferred	Over-detailed, answer obvious without a document

Step-by-Step Evaluation Process:

1. **Review the Context**: Begin by understanding the context or document that the queries reference, as this will be key to judging relevance and specificity.
2. **Score Each Query**: Evaluate each of the five AI-generated queries based on the rubric above. Assign a score from 1 to 5 for each criterion.
3. **Sum the Scores**: Add up the scores for each query across all four criteria.
4. **Select or Reject**: Choose the highest-scoring query, but if none of the queries perform satisfactorily (e.g., if all are under a threshold like 12/20), you have the option to **reject all queries** and refer back to the original.
5. **Optional Feedback**: If necessary, provide a brief rationale for your decision, especially if you rejected all queries or if two or more had similar scores.

Figure 20: Brief of Reviewer Instructions, including the Evaluation Rubric.