Multimodal Retrieval-Augmented Generation: Unified Information Processing Across Text, Image, Table, and Video Modalities

Nazarii Drushchak[1,2], Nataliya Polyakovska[1], Maryna Bautina[1], Taras Semenchenko[1,3], Jakub Koscielecki[1], Wojciech Sykala[1], Michal Wegrzynowski[1]

[1]SoftServe Inc., [2]Ukrainian Catholic University, [3]Taras Shevchenko National University of Kyiv

Correspondence: ndrus@softserveinc.com

Abstract

Retrieval-augmented generation (RAG) is a powerful paradigm for leveraging external data to enhance the capabilities of large language models (LLMs). However, most existing RAG solutions are tailored for single-modality or limited multimodal scenarios, restricting their applicability in real-world contexts where diverse data sources—including text, tables, images, and videos—must be integrated seamlessly. In this work, we propose a unified *Multimodal Retrieval-augmented generation (mRAG)* system designed to unify information processing across all four modalities. Our pipeline ingests and indexes data from PDFs and videos using tools like Amazon Textract, Transcribe, Langfuse, and multimodal LLMs (e.g., Claude 3.5 Sonnet) for structured extraction and semantic enrichment. The dataset includes text queries, table lookups, image-based questions, and videos. Evaluation with the Deepeval framework shows improved retrieval accuracy and response quality, especially for structured text and tables. While performance on image and video queries is lower, the multimodal integration framework remains robust, underscoring the value of unified pipelines for diverse data.

1 Introduction

The exponentially growing volume of digital content in various forms, including text, tables, images, and videos, has created new challenges. Traditional information retrieval systems typically focus on a single modality, such as text or images, limiting their ability to process complex queries that require insight from multi-modal data sources. However, real-world applications, such as enterprise data analytics, troubleshooting equipment through video manuals, or processing product specifications, need a framework to manage various data types.

Retrieval-augmented generation (RAG) systems have emerged as a powerful paradigm combining retrieval mechanisms with generative models to enhance information access and synthesis. However, conventional RAG frameworks were not designed initially to handle multimodal data, restricting their utility in environments where diverse data forms must be unified and processed seamlessly. This limitation underscores the need for an evolved approach that extends the capabilities of RAG systems to accommodate and integrate multiple modalities effectively.

This paper presents an mRAG system that unifies information across text, tables, images, and videos. Using tools like AWS, LangChain, and multimodal LLMs, it provides a robust pipeline for data ingestion, retrieval, and response generation.

2 Background and Related Work

The landscape of information retrieval has evolved significantly with the advent of large-scale digital data across diverse modalities. Traditional information retrieval systems focus mainly on single modalities, such as text-based search engines (Amati and Van Rijsbergen, 2002; Karpukhin et al., 2020; Khattab and Zaharia, 2020) or image retrieval systems (Lin et al., 2015; Chen et al., 2023), each optimized for their specific data type.

Multimodal information retrieval (MMIR) aims to bridge the gap between different data types, facilitating comprehensive searches that span text, images, videos, and other formats (Baltrusaitis et al., 2019). Researchers have successfully applied deep learning techniques for multimodal information retrieval (Hu et al., 2019).

RAG systems represent a paradigm shift in combining retrieval mechanisms with generative models. Introduced by Lewis et al. (2020), RAG leverages LLMs to generate contextually relevant responses by retrieving pertinent information from extensive external knowledge bases. RAG research has rapidly expanded, tackling efficiency bottlenecks (Borgeaud et al., 2021), memory constraints

Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025), pages 59–64

(Qian et al., 2024), and self-reflection strategies (Asai et al., 2023).

Recent advances in RAG have begun integrating multiple modalities to enhance retrieval and generation, as seen in MuRAG (Chen et al., 2022). However, most work remains limited to small, domain-specific datasets (e.g., healthcare) and only two modalities (Xia et al., 2024).

Key challenges remain in the development of multimodal RAG systems. Most existing approaches lack unified frameworks capable of reasoning across more than two modalities, such as text, tables, images, and videos. Scalability is also limited, as adding new modalities often requires separate training pipelines (Chen et al., 2022). Furthermore, current evaluation benchmarks primarily focus on single- or dual-modality tasks, making it difficult to assess systems designed for more complex, fully multimodal scenarios (Chen et al., 2024; Es et al., 2024; Krishna et al., 2024).

This work addresses these gaps by proposing a unified framework for building and evaluating multimodal RAG systems.

3 Methodology

3.1 Dataset Description

We test system capabilities by using 36 publicly available Dell server documents, including specifications, service manuals, and installation guides. These documents cover a range of modalities, including plain text, complex tables, and images, ensuring diverse data for testing.

Additionally, the dataset contains 82 video manuals of the servers, including one more modality. The dataset was selected to provide all the required modalities of varying complexities, reflecting real-world challenges in the technical documentation[1].

3.2 System Architecture

Information retrieval is structured into three primary layers: Data Processing, Embedding and Indexing, and Retrieval Engine. All operate within a cloud environment provided by Amazon Web Services (AWS) [2]. The generative module is built on the information retrieval component to support multimodal RAG scenarios.

The architecture of the main AWS components is represented in Figure 1.

[1]PDFs and videos can be shared upon request.

[2]https://aws.amazon.com/

The following sections of this research describe the detailed architecture of the retrieval and generation engines and the guardrails.

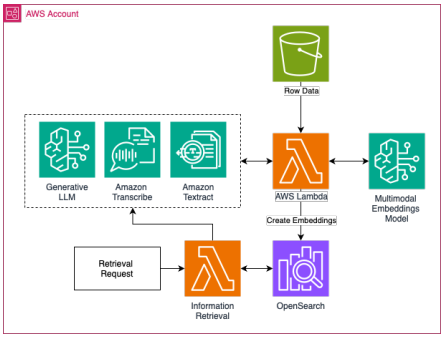

Figure 1: Main Services for Information Retrieval

3.3 Information Retriever

In this part, we explain how we create information retrievers. These pipelines are designed to prepare data for retrieval from various sources such as PDFs and videos.

3.3.1 PDF-based retriever

The PDF-based retriever processes PDFs to extract and index textual, tabular, and image data for efficient search. It is built on the **AWS stack** for scalability and performance, as illustrated in Figure 2.

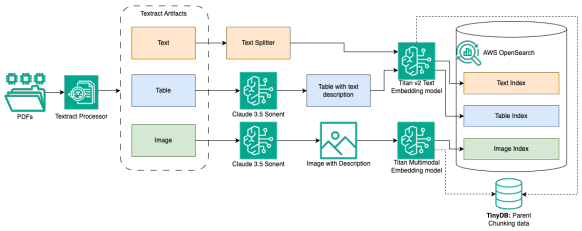

Figure 2: Pipeline for PDF-based Information Retrieval

Pipeline Overview:

1. **PDF Processing: Amazon Textract**[3] extracts text, tables, and images from PDFs.

2. **Text Splitting: LangChain**[4] split the text into contextually relevant chunks.

3. **Table Processing: Claude 3.5 Sonnet**(Team, 2024a,b) LLM generates semantic summaries for table data.

4. **Image Processing: Claude 3.5 Sonnet** LLM creates descriptive image metadata.

[3]https://docs.aws.amazon.com/textract/latest/dg

[4]https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/

5. **Embedding and Indexing:** Text and images are embedded using **Amazon Titan Text Embeddings V2**[5] and **Amazon Titan Multimodal Embeddings G1 models**[6] and indexed in **Amazon OpenSearch**[7].

6. **Metadata Tracking: TinyDB**[8] stores parent-child relationships between data chunks.

3.3.2 Video-based retriever

The video-based retriever extracts and indexes keyframe and textual data from videos using the **AWS stack**. The pipeline process is illustrated in Figure 3.

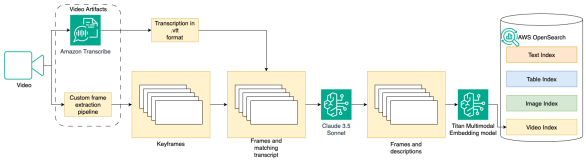

Figure 3: Pipeline for Video-based Information Retrieval

Pipeline Overview:

1. **Transcription:** Uses **Amazon Transcribe** to transcribe videos.

2. **Keyframe Extraction:** A custom pipeline based on **OpenCV**[9] extracts keyframes by detecting scene changes and analyzing content using entropy (≥ 4.5), edge ratio (≥ 0.02), contrast variation (≥ 600), and pixel changes ($\geq 5\%$). **Perceptual hashing** prevents redundancy, ensuring keyframes differ with a similarity threshold of 0.95.

3. **Context:** Matches keyframes with transcripts (±10-30s window).

4. **Description: Claude 3.5 Sonnet** generates enriched descriptions of keyframes, incorporating the visual details and corresponding transcript context.

5. **Indexing:** Embeds content via **Amazon Titan Multimodal Embeddings G1 model** and stores in **Amazon Open-Search**.

[5] https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html

[6] https://docs.aws.amazon.com/bedrock/latest/userguide/titan-multiemb-models.html

[7] https://aws.amazon.com/opensearch-service/

[8] https://tinydb.readthedocs.io/en/latest/

[9] https://opencv.org/

3.4 Multimodal Retrieval Augmented Generation

This section outlines the mRAG system's core components for answering user queries using file data.

1. **User Input Processing:** Queries are analyzed by **Claude 3.5 Sonnet** by checking the conversation history and the new user query. It then has two options: rephrase the query based on the context for continuity, or return the original message if the new query is unrelated to previous discussions.

2. **Independent Retrieval:** Relevant text, tables, and images are retrieved from **AWS OpenSearch** using a unified parent-child chunking strategy: smaller embedding-based chunks for search, with associated larger parent chunks provided to the model. Video modalities use only embedding retrieval. The top 10 textual results and the top 5 for other modalities are selected.

3. **Answer Generation:** Retrieved data and the user query are structured for **Claude 3.5 Sonnet** to generate responses.

4. **Citation and Traceability:** To ensure transparency, sources are cited with links to document pages or video timestamps.

3.5 Monitoring, Guardrail, and Feedback Loop

The system integrates monitoring, guardrails, and feedback to ensure ethical compliance. User interactions are tracked using **LangFuse**[10], with personally identifiable information (PII) anonymized by **Amazon Comprehend**[11]. **Amazon Bedrock Guardrails**[12] enforce safeguards to prevent harmful content and ensure AI safety (Chua et al., 2024). User feedback is analyzed based on the provided category, such as good, inconsistent, irrelevant, incomplete, confusing, or other. This feedback is processed with **Claude 3.5 Sonnet** to identify potential issues, and bugs are logged for resolution, enabling continuous system improvement.

[10] https://langfuse.com/

[11] https://aws.amazon.com/comprehend/

[12] https://aws.amazon.com/bedrock/guardrails/

Modality	Method	Correct Sources	Contextual Precision	Contextual Recall	Contextual Relevancy	Must Mention	LLM as Evaluator	Answer Relevancy	Faithfulness	Hallucination
All	Base	**0.652**	**0.349**	0.653	0.655	0.283	0.708	0.946	**0.677**	0.356
All	Opt	0.644	0.336	**0.690**	**0.702**	**0.290**	**0.717**	**0.951**	0.668	**0.314**
Text	Base	0.828	**0.493**	0.846	**0.846**	**0.068**	**0.812**	0.964	**0.672**	0.233
Text	Opt	**0.830**	0.491	**0.860**	0.846	0.058	0.809	**0.968**	0.636	**0.202**
Table	Base	**0.970**	**0.292**	**0.849**	0.818	0.702	0.752	**1.000**	**0.617**	**0.273**
Table	Opt	0.939	0.195	**0.849**	**0.879**	**0.742**	**0.782**	0.995	0.591	0.364
Image	Base	**0.694**	**0.332**	0.537	0.536	**N/A**	0.593	**1.000**	**0.718**	0.630
Image	Opt	0.685	0.313	**0.573**	**0.628**	**N/A**	**0.650**	0.994	0.662	**0.537**
Video	Base	**0.293**	0.190	0.399	0.417	**N/A**	**0.619**	0.876	0.682	0.394
Video	Opt	0.281	**0.193**	**0.481**	**0.496**	**N/A**	0.613	**0.891**	**0.737**	**0.323**

Table 1: Experimental results across different modalities comparing Base and Optimized (Opt) Q&A prompts. Bold values indicate the best performance for each metric within each modality.

4 Experiments

4.1 Experiments Setup

We evaluated our system using a dataset of 36 PDF documents and 82 videos, based on **Dell server specifications and service manuals**. Four participants were involved in the question creation process, with each person generating queries across all modalities: text, table, image, and video.

The benchmarking set includes 116 questions[13]: 43 for text, 22 for tables, 18 for images, and 33 for videos. We executed the system three times for each question and averaged the scores to obtain stable results.

An example question format is:

```
{
 "query":"How to set up T150 system?",
 "answer":"Perform the following steps to set up the system:
    1. Unpack the system.
    2. Connect the peripherals.
    3. Power on the system.",
 "sources":["Dell EMC PowerEdge T150 Installation
    and Service Manual.pdf"],
 "type":"text"
}
```

Langfuse[14] was used to track experiments, and **Deepeval**[15] as core evaluation framework.

4.2 Evaluation Metrics

The evaluation used two sets of metrics: retrieval and response. Retrieval metrics included the percentage of correct sources retrieved, contextual precision and recall, and the relevancy of retrieved contexts. Response metrics assessed keyword inclusion ("must mention"), LLM as evaluator score (rated by **GPT-4o** (OpenAI and et al., 2024)), answer relevancy, faithfulness to sources, and the presence of hallucinations.

4.3 Experimental Results

We evaluated the system's performance using two experimental setups: a baseline prompt (Base) and a manually optimized prompt based on providing additional limitations (Opt). Table 1 summarizes the results, with the best metric for each category/modality highlighted in bold.

Overall, the optimized prompt slightly outperformed the baseline in most metrics, particularly in contextual recall, relevancy, and hallucination reduction. However, performance varied by modality. Text and table modalities demonstrated the highest accuracy and stability, benefiting from the structured nature of their data. Image and video modalities showed lower performance, reflecting the challenges of interpreting and retrieving unstructured visual content.

Notably, video retrieval had the lowest scores in correct sources and contextual metrics, indicating room for improvement in handling video data. Despite this, optimized prompts improved performance metrics for both image and video modalities.

5 Conclusion and Future Work

This work presents a methodology for building an mRAG system, focusing on pipelines for extracting and indexing text, tables, images, and videos. Experimental results show improved contextual relevancy, LLM evaluation scores, and reduced hallucinations, while performance variations highlight challenges with unstructured data.

Future work will focus on enhancing mRAG with improved LLM capabilities, fine-tuning embeddings for better domain understanding, incorporating user feedback, and adding visual modalities for input.

6 Ethical Consideration

This study builds an mRAG system processing text, images, tables, and videos, ensuring data privacy and security. It uses only open-source PDFs, anonymizes all requests and feedback, and uses feedback solely to improve system performance.

[13] Evaluation dataset and script can be shared upon request.

[14] https://langfuse.com/

[15] https://docs.confident-ai.com/

We used ChatGPT[16] and Grammarly[17] to help refine the writing of this work, ensuring the language is straightforward.

References

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *ArXiv*, abs/2310.11511.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, T. W. Hennigan, Saffron Huang, Lorenzo Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and L. Sifre. 2021. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S. Lew. 2023. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7270–7292.

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

[16] https://chat.openai.com/

[17] https://www.grammarly.com/

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *Preprint*, arXiv:2407.18369.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 635–644, New York, NY, USA. Association for Computing Machinery.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *Preprint*, arXiv:2409.12941.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. 2015. Deep learning of binary hash codes for fast image retrieval. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 27–35.

OpenAI and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *Preprint*, arXiv:2409.05591.

Anthropic Team. 2024a. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Anthropic Team. 2024b. Claude 3.5 sonnet model card addendum. Technical report, Anthropic.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.