Technical Report: Multi-Modal Document Intelligence using RAG

1. Architecture Overview

The system is designed as a modular **Multi-Modal Retrieval-Augmented Generation (RAG) pipeline** to answer questions over complex financial and policy documents. The architecture consists of five stages:

1. **Multi-Modal Ingestion**
2. **Chunking and Normalization**
3. **Unified Embedding and Vector Indexing**
4. **Cross-Modal Retrieval**
5. **Answer Generation with Citations**

Each stage operates independently and communicates through structured JSON artifacts, enabling scalability and ease of debugging.

---

2. Design Choices

2.1 Multi-Modal Ingestion

Separate ingestion pipelines are implemented for text, tables, and images. Text is extracted page-wise from PDFs, tables are parsed using layout-aware extraction, and images are processed using OCR to capture information from charts and scanned pages. All extracted elements are normalized into a unified schema containing content, page number, modality, and source document.

This separation ensures modality-specific processing while maintaining a consistent downstream representation.

2.2 Chunking Strategy

A modality-aware chunking strategy is adopted to balance retrieval accuracy and context efficiency:

- Text is split into paragraph-level chunks of approximately 400–600 tokens.
- Tables are retained as single chunks to preserve structural semantics.
- OCR-derived image text is stored as compact standalone chunks.

Page-level metadata is preserved in all chunks to enable precise citation.

2.3 Unified Embedding Space

All chunks are embedded into a single semantic vector space using a Sentence Transformer model. This design enables **cross-modal retrieval**, allowing text queries to retrieve relevant tables or OCR-derived evidence without maintaining separate indices.

A FAISS vector index is used to support fast similarity search with low latency.

2.4 Retrieval and Answer Generation

User queries are embedded and matched against the vector index to retrieve the top-k most relevant chunks across modalities. Retrieved context is passed to a Large Language Model via **Groq inference**.

Prompting strictly enforces:

- context-only answering
- prohibition of external knowledge
- explicit citation of retrieved sources

This constraint-based generation significantly reduces hallucinations.

2.5 User Interface and Memory

A Streamlit-based interface provides interactive question answering, expandable citations, and session-level conversational memory. Memory is maintained only at the UI layer to avoid contaminating retrieval context, preserving factual grounding.

---

3. Benchmarking and Evaluation

The system is evaluated using a small benchmark set of queries designed to test:

- **Multi-modal coverage** (text, tables, OCR)
- **Citation presence**
- **End-to-end latency**

For each query, retrieved modalities and response time are recorded. The evaluation demonstrates:

- correct modality-appropriate retrieval
- consistent citation inclusion

- sub-second response latency suitable for interactive use

The evaluation focuses on system-level behavior rather than language metrics, aligning with real-world document intelligence requirements.

---

4. Key Observations

- OCR-derived image content significantly improves answer completeness for charts and scanned sections.
- Modality-aware chunking yields higher retrieval precision compared to naive text splitting.
- A unified embedding space simplifies cross-modal retrieval without increasing system complexity.
- Groq inference provides low-latency generation suitable for real-time applications.

---

5. Limitations

- Table extraction accuracy depends on the structural quality of source PDFs.
- Visual rendering of images is not enabled by default.
- Learning-based reranking and retrieval fine-tuning are not implemented and are left as future enhancements.

---

6. Conclusion

The implemented system demonstrates a practical and scalable **multi-modal document intelligence pipeline** capable of answering complex questions over real-world documents. By integrating structured ingestion, unified embeddings, cross-modal retrieval, and citation-grounded generation, the system effectively addresses key challenges in enterprise document analysis.