Multimodal RAG and Parsing: Data Architectures for Semantic Retrieval in Complex Industrial Layouts

Renato Miyaji[1,2], Arthur Luz[1], Renato Moulin[2], Leonardo Machado[2], Pedro L. P. Corrêa[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)

[2]Grupo Visagio

{re.miyaji, pedro.correa}@usp.br

{renato.moulin,leonardo.machado}@visagio.com

Abstract. *Traditional RAG systems struggle with industrial documents by neglecting structural semantics in complex layouts. To mitigate this, we propose a formal Visual Document Understanding ingestion methodology and a Semantic Chunking framework. By orchestrating Vision-Language Models to generate high-fidelity captions for visual elements and grouping them with adjacent text, our pipeline preserves context-rich retrieval units. Experiments on an industrial dataset demonstrate 88.0% accuracy, a 0.91 Hit Rate@5, and 0.89 Context Recall, significantly outperforming OCR baselines (74.0% accuracy) and matching state-of-the-art solutions.*

1. Introduction

Large Language Models (LLMs) have achieved unprecedented success in diverse cognitive tasks. However, these models still face inherent limitations, such as a reliance on static training data and hallucinations. Retrieval-Augmented Generation (RAG) has emerged as a robust solution to mitigate these issues [Gao et al. 2023]. While RAG is well-consolidated for text-based information retrieval, its application has been largely restricted to plain-text knowledge bases, often neglecting the critical information available in other modalities. In industrial applications, base documents are inherently multimodal and visually rich structures, such as technical reports, complex schematics, and scientific papers [An et al. 2024]. Standard text-centric RAG pipelines frequently filter out visual elements, leading to a significant loss of structural semantics and cross-modal cues [Gao et al. 2025]. Consequently, building an effective retrieval system for these domains requires a paradigm shift toward Visual Document Understanding (VDU) [Kim et al. 2022].

A bottleneck for efficient multimodal document retrieval is the data ingestion and parsing pipeline [Gao et al. 2025]. Traditional Optical Character Recognition (OCR)-based methods fail to extract data from complex elements and are unable to fully capture the qualitative information embedded in charts and figures [Gu et al. 2021]. While recent research has moved toward OCR-free architectures to enable faster indexing, most RAG systems still struggle to integrate textual content with visual metadata [Gao et al. 2025]. This gap creates a ”visual information bottleneck”where the relationship between text and layout is severed during the segmentation process.

In this paper, we address these limitations by proposing a modular architecture for Multimodal RAG and Parsing specifically designed for technical documents with complex layouts. Unlike traditional pipelines that prioritize raw extraction, our objective is to formalize a methodology for Semantic Chunking that preserves the relation between content and layout. By integrating Vision-Language Models (VLMs) as orchestration agents, visual elements are indexed and recovered as context-aware retrieval units without information loss.

The main scientific and technical contributions of this work are summarized as follows: (1) a modular VDU-based ingestion methodology that formalizes the transition from unstructured multimodal layouts to structured semantic representations; (2) a Semantic Chunking algorithm for complex documents that maintains the text-image relationship through hierarchical decomposition and detailed metadata enrichment; (3) a hybrid retrieval pipeline that demonstrates how the orchestration of open-source and proprietary VLMs can outperform or match integrated enterprise solutions by mitigating conversion errors; and (4) an experimental validation in a Decision Intelligence case study.

2. Related Works

Traditional RAG grounds LLM responses in factual data but often fails to leverage the context within visually rich documents. Multimodal RAG (MRAG) addresses this by integrating diverse data types [Mei et al. 2025]. Because standard OCR pipelines face high computational costs and error propagation [Gao et al. 2025], the field has shifted toward OCR-free architectures. Models like LayoutLM combine spatial and textual data [Xu et al. 2020], while others map images to semantics or vision-space embeddings. Recent research divides these multimodal systems into closed-domain and open-domain architectures [Gao et al. 2025]. Closed-domain approaches utilize teacher-student LLM setups, closed-loop extraction, or multi-stage pipelines. In contrast, open-domain methodologies tackle the complex generation and storage of knowledge across diverse collections. Solutions include embeddings for unified processing, alongside multi-agent frameworks that employ iterative reasoning, dynamic thresholds, and consistency-based voting.

3. Methodology

3.1. Proposed Method

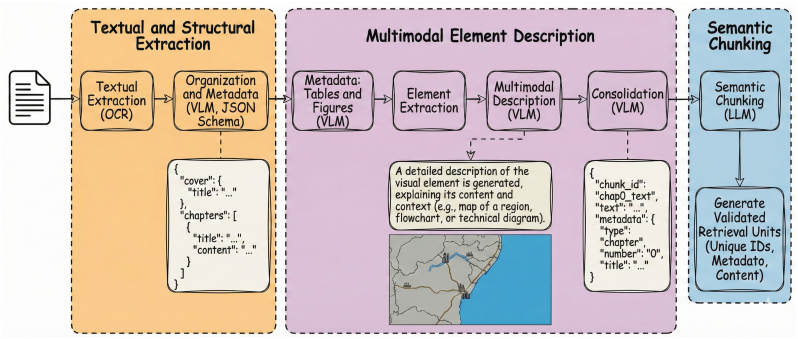

Figura 1. Architecture of the multimodal VDU pipeline for industrial document processing.

The proposed architecture implements a modular, model-agnostic pipeline for VDU, specifically engineered to overcome the primary performance bottlenecks in data

ingestion for industrial documents. Unlike traditional RAG systems that rely on monolithic proprietary solutions, our framework democratizes access to multimodal retrieval by orchestrating interchangeable VLMs via LangGraph. This ensures that the structural semantics of complex layouts are preserved without locking the pipeline into a closed-source ecosystem. As Figure 1 presents, the workflow is divided into three main stages: textual and structural extraction, multimodal element description, and semantic chunking.

The first stage begins with Textual Extraction utilizing an OCR engine (Fitz [McKie 2024]), which retrieves the integral textual content while preserving the original spatial sequence of information. This raw content is subsequently processed by a VLM, specifically Gemini 3 Flash [DeepMind 2025a], in the Organization and Metadata module. Our engineering contribution at this stage lies in the prompt design, which utilizes zero-shot chain-of-thought heuristics to instruct the VLM to transform unstructured text into a deterministic, strictly validated JSON schema. This structuring captures the logical hierarchy (cover, titles, and chapters) and spatial anchors, which is vital because the relative position of document elements significantly contributes to the semantic representation.

Concurrent with textual organization, the pipeline identifies and extracts visually rich elements through the Multimodal Metadata module. The VLM is employed to detect bounding regions and classify document segments, ensuring that tables and technical diagrams are not fragmented. This stage mitigates the typical information loss seen in traditional OCR-based parsing. A core methodological innovation of this framework is the Multimodal Description stage, powered by a state-of-the-art VLM, Gemini 3 Pro [DeepMind 2025b]. For every extracted visual element, the VLM generates a high-fidelity textual summary guided by instruction prompts specifically engineered for technical layouts (e.g., explicitly demanding the extraction of tabular relationships and flowchart flows). This captioning strategy acts as a bridge between the visual and textual modalities.

In the Consolidation module, the integral textual data and the multimodal descriptions are unified based on a strict spatial-textual alignment rule. By cross-referencing the bounding boxes of visual elements with the positional markers of the OCR text, the architecture maps figures and tables to their adjacent narrative context, ensuring that distributed correlation signals across a page are preserved.

The final stage formalizes the Semantic Chunking process. Rather than relying on naive text segmentation, this module produces contextually autonomous retrieval units. Formally, the algorithm transforms the document into a set of retrieval units U, where each chunk is defined as $u_i = \{T_i, V_i, M_i\}$. Here, T_i represents the integral adjacent text, V_i represents the high-fidelity multimodal description generated by the VLM, and M_i encapsulates the hierarchical metadata (e.g., specific document type, chapter, figure number, and title). This formalization ensures high-precision retrieval during the online querying phase, as each unit u_i contains sufficient context and structural tags to ground the generative model in the most relevant documentary evidence without semantic leakage.

3.2. Experiments

The primary objective of this experiment was to quantitatively evaluate the efficacy of the proposed VDU-based ingestion and Semantic Chunking framework in mitigating the vi-

sual information bottleneck inherent in complex industrial layouts. The methodology was validated using a high-stakes dataset comprising 80 visually rich technical reports from a consulting firm operating across diverse sectors (e.g., banking, mining, and oil and gas). These reports convey critical information not only through text but also through complex charts, technical tables, and schematics. To ensure reproducibility while complying with double-blind review guidelines, a representative, anonymized subset of the proprietary industrial dataset and the orchestration code have been prepared. These artifacts will be made publicly available in a repository upon the paper's acceptance.

To ensure a rigorous assessment, an Expert-in-the-Loop evaluation framework was established. Domain experts, specialized in the themes covered by the technical reports, were tasked with proposing a diverse set of queries and establishing the corresponding ground truth. This approach ensured that the benchmark accurately reflected realistic information-seeking scenarios. The initial validation of the generated responses was conducted by human evaluators using a binary scoring system focused strictly on factual correctness. To complement this manual validation, we implemented an evaluation pipeline. The retrieval stage was assessed using Hit Rate and Mean Reciprocal Rank (MRR). For the generation stage, we adopted the RAGAS framework to measure Faithfulness, Answer Relevance, and Context Recall with GPT-5.2.

To isolate the specific impact of the multimodal parsing and semantic chunking modules, a series of comparative experiments were conducted against two distinct baselines. The first baseline relied exclusively on standard textual extraction via the Fitz engine [McKie 2024], representing a conventional text-only OCR pipeline that neglects layout semantics and visual cues, leading to anticipated information loss. The second baseline evaluated an open-source multimodal approach by substituting the Gemini models with Nanonets-OCR-s [Mandal et al. 2025] in the parsing and description stages. Powered by Qwen2.5-VL-3B [Qwen 2025], this model specializes in image-to-markdown conversion and intelligent image description, allowing us to test the framework's modularity and evaluate the effectiveness of different visual backbones in establishing fine-grained grounding.

Across all experimental configurations, the RAG system was constructed utilizing Chroma as the vector database in conjunction with the LangGraph framework for modular pipeline orchestration. GPT-5.1 served as the base LLM for the generation stage, while a retrieval parameter of $K = 5$ was consistently applied to ensure that the model received context-rich retrieval units for grounded synthesis.

4. Results

Tabela 1. Comparative evaluation results including human annotation accuracy, Information Retrieval metrics, and automated RAGAS metrics.

Configuration	Acc (%)	Hit Rate@5	MRR@5	Faithfulness	Relevance	Recall
OCR only	0.74	0.68	0.55	0.87	0.69	0.65
Nanonets-OCR-s	0.82	0.79	0.68	0.85	0.81	0.78
Gemini 3	0.88	0.91	0.82	0.90	0.88	0.89

The experiments revealed significant performance variances across the three tested configurations. The baseline system, which relied exclusively on conventional OCR-

based extraction via the Fitz engine [McKie 2024], achieved an accuracy rate of 74.0%. As detailed in Table 1, this configuration struggled particularly in retrieval effectiveness, yielding a Hit Rate@5 of 0.68 and an MRR@5 of 0.55. Furthermore, automated RAGAS metrics indicated low Context Recall (0.65) and Answer Relevance (0.69). This lower performance underscores the critical limitations of text-centric RAG pipelines.

Performance improved notably when the pipeline was modified to incorporate the Nanonets-OCR-s [Mandal et al. 2025] configuration, which yielded an accuracy of 82.0%. By substituting the primary VLM backbones with this open-source model designed for image-to-markdown conversion, the system demonstrated greater modularity and an enhanced ability to structure elements that traditional OCR typically fragments or ignores. The improvement is clearly reflected in its retrieval metrics, achieving a Hit Rate@5 of 0.79 and an MRR@5 of 0.68, alongside significant increases in Context Recall (0.78) and Answer Relevance (0.81).

The proposed modular architecture utilizing the Gemini 3 models [DeepMind 2025b] achieved the highest performance, reaching a peak accuracy of 88.0%. This superior result is strongly supported by the experimental metrics, peaking at a Hit Rate@5 of 0.91 and an MRR@5 of 0.82, as well as RAGAS scores including a Faithfulness of 0.90 and a Context Recall of 0.89. This performance is attributed to the integration of multimodal captioning and context-aware semantic chunking, which preserved the logical hierarchy of the technical reports more effectively than the alternative models. By maintaining the intricate relationship between narrative text and visual evidence, the Gemini-based pipeline provided the most reliable grounding for decision-making.

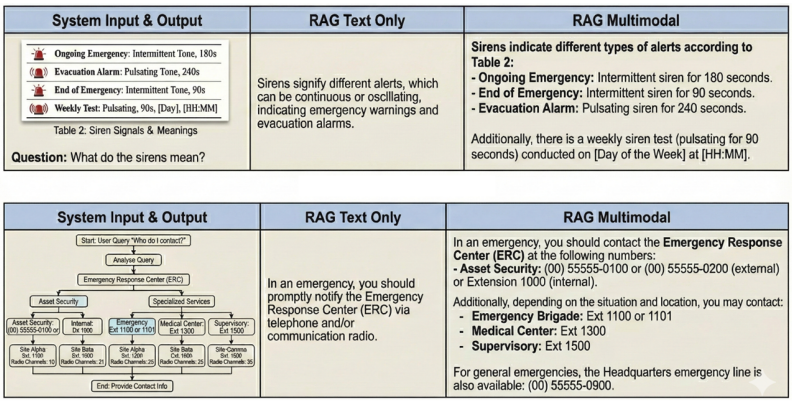

Figura 2. Case study validation: Unimodal vs. Multimodal RAG responses.

The experimental results highlight a significant performance disparity between traditional text-only RAG baselines and the proposed multimodal approach when processing visually rich industrial documents, as Figure 2 presents. Because critical safety and operational data are often embedded in complex tables and flowcharts, text-centric systems suffer from a visual information bottleneck, frequently yielding generic or superficial answers—such as failing to retrieve specific siren durations or navigate hierarchical emergency contact structures. The proposed architecture overcomes this limitation by employing VLMs to generate detailed, qualitative textual descriptions of these visual

elements. By translating complex layouts and spatial semantics into context-rich metadata, the system successfully bridges the semantic gap, enabling models to extract granular data accurately and preventing the severe consequences of incomplete information in high-stakes environments.

These findings confirm that the initial data ingestion pipeline is the primary bottleneck in technical document retrieval, and that Semantic Chunking is essential for modern Decision Intelligence applications. Unlike standard OCR methods that fragment keywords and lose structural detail, the proposed VDU framework maintains the logical flow between narrative text and supporting visual data by creating image-aware retrieval units. This hybrid strategy represents multimodal data within the textual modality, allowing for high-precision semantic search without the heavy computational overhead of processing raw pixels during the inference stage. Ultimately, by moving beyond a purely text-centric paradigm to automate the extraction of structured data, organizations can transform dense industrial reports into highly reliable, searchable knowledge bases that significantly enhance corporate decision-support systems.

5. Conclusions

This paper introduced a formal VDU-based ingestion methodology and a Semantic Chunking framework specifically designed to mitigate the visual information bottleneck in complex industrial documents. Our approach demonstrated how the model-agnostic orchestration of VLMs can successfully map distributed visual and textual semantics into context-aware retrieval units. Experimental validation confirmed the scientific and practical efficacy of this methodology. Achieving a peak accuracy of 88.0% , alongside robust retrieval metrics (Hit Rate@5 of 0.91 and MRR@5 of 0.82) and superior generation quality scores (Faithfulness of 0.90 and Context Recall of 0.89), the proposed framework not only significantly outperformed traditional OCR baselines but also established a robust alternative capable of matching or exceeding the performance of state-of-the-art VDU solutions.

By transforming complex visual elements into context-rich metadata, the system enables high-precision retrieval without the need of processing raw pixels during inference. This allows LLMs to synthesize grounded answers from previously inaccessible data. Furthermore, to democratize access to multimodal retrieval and promote reproducibility beyond proprietary industrial silos, the orchestration artifacts and a representative dataset subset have been open-sourced. Future research should aim to scale this architecture to larger datasets and further reduce information loss during retrieval. Additionally, subsequent studies should explore the trade-offs between hybrid metadata-driven strategies and native multimodal models, alongside developing evaluation methods for VLM-generated descriptions.

Acknowledgments

This study was supported by the Amigos da Poli – project 2025_021. It was also financed in part by CAPES – Finance Code 001.

Referências

An, Z., Ding, X., Fu, Y.-C., Chu, C.-C., Li, Y., and Du, W. (2024). Golden-retriever: High-fidelity agentic retrieval augmented generation for industrial knowledge base. *arXiv preprint arXiv:2408.00798*.

DeepMind, G. (2025a). Gemini 3 flash: Frontier intelligence built for speed. `https://blog.google/products/gemini/gemini-3-flash`.

DeepMind, G. (2025b). Gemini 3 pro: State-of-the-art multimodal ai model. `https://blog.google/products/gemini/gemini-3`.

Gao, S., Zhao, S., Jiang, X., Duan, L., Chng, Y. X., Chen, Q.-G., Luo, W., Zhang, K., Bian, J.-W., and Gong, M. (2025). Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding. *arXiv preprint arXiv:2510.15253*.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gu, J., Kuen, J., Morariu, V. I., Zhao, H., Barmpalios, N., Jain, R., Nenkova, A., and Sun, T. (2021). Unified pretraining framework for document understanding. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. (2022). Ocr-free document understanding transformer. In *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)*, pages 498–517, Berlin, Heidelberg. Springer-Verlag.

Mandal, S., Talewar, A., Ahuja, P., and Juvatkar, P. (2025). Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging. `https://huggingface.co/nanonets/Nanonets-OCR-s`.

McKie, J. X. (2024). *PyMuPDF: Python bindings for the MuPDF library*. Version 1.24.x.

Mei, L., Mo, S., Yang, Z., and Chen, C. (2025). A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*.

Qwen (2025). Qwen2.5-vl-3b: Instruction-tuned vision-language model. `https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct`.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pages 1192–1200, New York, NY, USA. Association for Computing Machinery.