Improving Industrial RAG with Cross Encoder Reranking and Image Summary Embeddings

Shervan Shahparnia
Department of Computer Engineering
San Jose State University
San Jose, CA, USA
shervan.shahparnia@sjsu.edu

Aadarsh Marahatta
Department of Computer Engineering
San Jose State University
San Jose, CA, USA
aadarsh.marahatta@sjsu.edu

Abstract— **Industrial documentation contains both dense text and highly detailed diagrams that convey essential operational procedures. Retrieval Augmented Generation (RAG) methods can help large language models answer questions grounded in such documents, but multimodal retrieval remains a difficult challenge. Prior work introduced two strategies: raw image embeddings and image summary embeddings. The original study showed that converting images into text improves downstream performance, but it also reported significant retrieval noise and poor reliability in visual relevance.**

This paper introduces an end to end multimodal RAG pipeline designed to address these issues. Our system uses MiniLM [2] for text embeddings, SigLIP2 for visual embeddings, and Gemini 2.5 Flash for both image summarization and answer generation. The primary contribution of this work is the addition of a cross encoder reranking stage using BAAI bge reranker base, which improves semantic alignment by rescoring text or summary candidates using joint query and document attention. We evaluate four configurations on a set of 35 question and answer tasks derived from six industrial manuals. Results show that summary based retrieval outperforms raw visual embedding retrieval, and that reranking improves image faithfulness, context alignment, and overall answer correctness. This work confirms earlier findings while providing a more reliable and complete multimodal RAG system suitable for technical and industrial domains.

I. INTRODUCTION

Technical manuals across industrial, mechanical, electrical, and general engineering domains are critical resources that combine textual instructions with complex visual content such as diagrams, schematics, flowcharts, and annotated illustrations. These materials are not limited to industrial settings and also include instruction manuals, maintenance guides, assembly documents, and technical handbooks used in manufacturing, construction, electronics, automotive systems, and consumer equipment. Visual elements like wiring schematics, component views, and process diagrams are essential for conveying spatial relationships, operational sequences, and system dependencies that are difficult or impossible to express through text alone. As a result, large language models often struggle when answering questions that require joint reasoning over both text and images. Common failure modes include hallucinating non existent steps, misinterpreting schematic symbols or diagram structure, overlooking visual constraints, or inferring details that are not explicitly present in the source material.

Retrieval Augmented Generation (RAG) has emerged as the state-of-the-art strategy to mitigate hallucination by providing relevant and grounded evidence to the model during the generation phase. However, traditional RAG pipelines are optimized for only text and struggle to reliably support image retrieval or multimodal indexing. Industrial diagrams present unique challenges as they are often visually cluttered, contain fine line details (e.g., flow lines, measurements), and rely on precise textual labels that standard image encoders (trained on natural images) may not capture effectively. These documents also vary widely in formatting, which further complicates the clean extraction required for vectorization.

A recent paper explored two multimodal RAG approaches. The first uses raw image embeddings and text embeddings stored in separate vector spaces. The second converted images into descriptive textual summaries to store in a unified vector storage [1]. That study discovered that image summaries had clear benefits, but also highlighted a persistent and significant retrieval problem. When images or their summaries were retrieved, the initial retrieval stage often returned context passages that were incorrect or weakly related to the query. These high recall, low precision retrieval errors then cascaded into the generation phase, resulting in poor final answers.

Our work addresses this critical failure point by building an end-to-end multimodal RAG system that incorporates a dedicated post-retrieval reranking module. Instead of using the initial similarity scores from the bi-encoder retrieval stage, we send the top k candidates to a cross-encoder reranker. The cross-encoder jointly processes the query and each candidate document, using attention mechanisms to produce a more reliable relevance score. This process refines the final context documents, which significantly enhances grounding fidelity, improves the alignment of visual evidence, and reduces hallucinations. This advancement is crucial for deploying RAG systems in technical and industrial environments where accuracy in answers is critical.

II. RELATED WORK

Multimodal retrieval has been shaped by two primary directions. The first relies on contrastive vision language models like CLIP [5] and SigLIP [3], which map images and text into a shared embedding space. These models perform well on natural images but can struggle with industrial diagrams that rely on precise labeling or geometric structure. Small changes in diagrams or text labels may be lost during contrastive training, producing weaker representations for retrieval.

The second direction relies on transforming images into text through captioning or summarization. When images are expressed in natural language, they can be embedded using standard text models. This produces a single modality index, simplifies retrieval, and often improves alignment because image content is represented through descriptive language. However, the quality of summaries depends heavily on the choice of multimodal language model and the style of the prompt.

In information retrieval, reranking is a long standing technique for improving relevance. Traditional search engines often retrieve an initial set of candidates using light weight embeddings, followed by a deeper neural or rule based model that refines the ranking. Cross encoders in particular have become popular in natural language retrieval because they jointly encode the query and candidate document, providing stronger semantic matching than independent bi encoders.

Although reranking is well established in text retrieval, it has not been widely studied in multimodal RAG pipelines designed for industrial contexts. Our work fills this gap by evaluating reranking on both raw image embedding systems and image summary systems.

III. Dataset

We tested our system on six publicly available industrial manuals. These manuals include an espresso machine, a refrigerator, a water heater, a dial up telephone, and two additional household appliances. The manuals contain both text instructions and diagrams showing internal components, electrical wiring, and mechanical details.

From these manuals, we created a set of 35 question and answer tasks. The questions require grounded reasoning using text, images, or a combination of both. The variety of formats and styles provides a realistic challenge for multimodal retrieval. Questions span multiple categories. Some require understanding textual procedures and safety instructions, others demand interpretation of schematic diagrams and component layouts, and many necessitate synthesizing information from both modalities to provide accurate answers.

IV. Methodology

Our goal is to implement a complete multimodal RAG system capable of handling both text and image based evidence from industrial manuals. The methodology follows four major phases: extraction, embedding, retrieval, and post retrieval reranking, followed by grounded answer generation. This design addresses the failures noted in earlier work, which identified retrieval noise as a primary obstacle in multimodal RAG pipelines.

The central goal of this work is to implement a complete, multimodal RAG system capable of reliably processing and utilizing both text- and image-based evidence from industrial manuals and sources. The methodology is structured into four sequential and critical phases: extraction, embedding, retrieval, and post-retrieval reranking, followed by the final grounded answer generation. This systematic design explicitly addresses the observed failures in prior work, which identified initial retrieval noise as the primary impediment to multimodal RAG accuracy.

A. System Overview

Fig. 1 shows the overall workflow. The system begins by loading PDF documents and processing each file through two extraction paths. One path identifies and chunks text into segments of approximately 225 words with a 50 word overlap. A second path extracts diagrams and images for further processing.

Documents then pass through one of two configurations. Config A stores text and image embeddings in separate vector spaces. Config B converts each image into a textual summary using Gemini 2.5 Flash, allowing all content to be stored in a single index. This enables direct comparison of raw image embedding retrieval versus summary based retrieval.

After embedding, all content is stored in a ChromaDB vector database [7]. At query time, the system retrieves the top 20 candidates with high recall. Because the original research found that retrieval noise significantly harms downstream accuracy, we introduce a Cross Encoder Reranking stage to refine the candidate list. The reranker strengthens semantic alignment and improves precision. Finally, Gemini 2.5 Flash produces the grounded answer based on the selected context.

B. Config A: Multimodal Embeddings with SigLIP2

Config A uses MiniLM for text embeddings and SigLIP2 for image embeddings. Text chunks and images are stored in separate vector spaces. At retrieval time, the system performs two searches. One search finds the nearest text embeddings and the other search finds the nearest image embeddings. These results are merged and forwarded to the reranker.

This configuration preserves visual information without modifying images. However, diagrams in industrial manuals often contain dense labels and fine structure that contrastive image encoders may not capture effectively. This leads to weaker retrieval performance and supports findings from the original study regarding noisy or semantically irrelevant image matches.

C. Config B: Textual Image Summaries in a Unified Vector Space

Config B converts each image into a textual summary using Gemini 2.5 Flash. The summarization prompt requests component labels, structural descriptions, and functional relationships. The resulting summaries are embedded using MiniLM and placed into the same vector database as the text chunks.

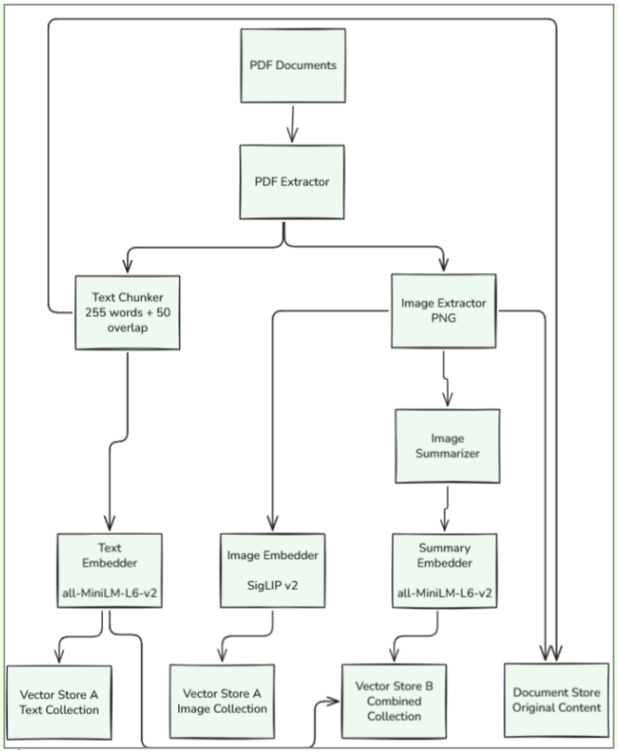

PDF Documents
PDF Extractor
Text Chunker 255 words + 50 overlap
Image Extractor PNG
Image Summarizer
Text Embedder all-MiniLM-L6-v2
Image Embedder SigLIP v2
Summary Embedder all-MiniLM-L6-v2
Vector Store A Text Collection
Vector Store A Image Collection
Vector Store B Combined Collection
Document Store Original Content

Fig. 1. Overview of the multimodal RAG pipeline including extraction, embedding, retrieval, reranking, and generation.

This removes the need to align two embedding spaces and creates a uniform retrieval structure. Prior work suggested that summaries provide more robust semantic representations than raw image embeddings. Our findings confirm that summary based retrieval generally yields stronger relevance scores and more accurate final answers.

D. Reranking Pipeline

Fig. 4 shows the reranking process. After initial retrieval, the top 20 candidates are paired with the user query and evaluated using the BAAI bge reranker base model. This model is a Cross Encoder that processes the query and candidate together in a single forward pass. This allows deeper semantic matching compared to independent bi encoder embeddings.

The reranker outputs a relevance score for each candidate. The system selects the top 4 candidates as the final context. The addition of a reranker directly addresses the retrieval noise problem reported in the original study and substantially improves grounding quality.

E. Generation

Gemini 2.5 Flash receives the selected context and produces the final answer. The generation prompt instructs the model to avoid speculation and rely only on retrieved content. This helps ensure that incorrect answers usually result from retrieval failures rather than hallucination. All prompts are included in the appendix.

V. Experiments

A. Evaluation Metrics

We evaluate the four configurations using six metrics that reflect both retrieval quality and answer quality. These metrics are scored using an LLM judge for consistency and replicability.

- Answer Correctness: Measures whether the generated answer matches the ground truth.
- Answer Relevancy: Evaluates whether the answer directly addresses the question.
- Text Faithfulness: Checks whether the answer is supported by the retrieved text.
- Image Faithfulness: Measures whether the answer relies correctly on the visual content or its summary.
- Text Context Relevancy: Evaluates whether retrieved text chunks are relevant to the query.
- Helpfulness: Assesses practical usefulness and clarity of the answer.

These metrics match the evaluation protocol from the original paper, enabling direct comparison [1].

B. Dataset and Experimental Setup

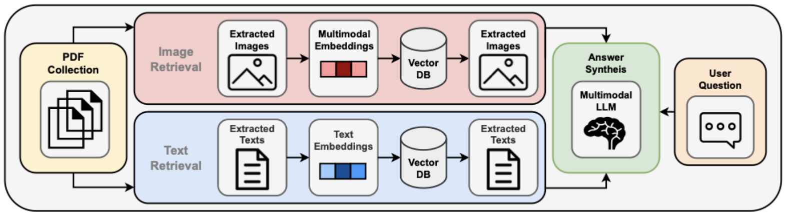
We use six industrial manuals that include mechanical devices, electrical appliances, and household equipment. These documents contain dense procedural text and complex diagrams. From these manuals, we created 35 question and answer tasks that represent realistic industrial queries requiring text based reasoning, image based reasoning, or both.

All variables outside the retrieval and reranking configurations were held constant. The same extraction rules, embedding models, prompts, and generation settings were used across all experiments. The system was tested on a single workstation using ChromaDB as the vector store and Gemini 2.5 Flash for both summarization and answer generation.

C. Relation to Observations From Prior Work

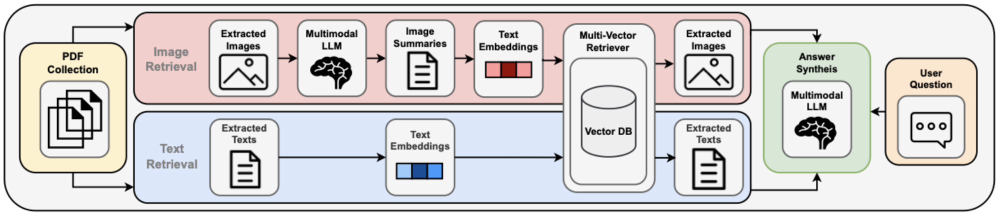
Our experiments reproduce several key findings emphasized in the slides and in the original research.

- Summary based retrieval is more stable and semantically precise than raw image embedding retrieval. This matches the results shown in the presentation where Config B consistently performed better.
- Retrieval noise is a major source of error. Both our work and the original study show that visual embeddings often return semantically irrelevant diagrams. Reranking substantially reduces this issue.
- Prompt quality significantly affects performance. Summaries depend on the clarity of the summarization prompt. Poorly written prompts lead to incomplete summaries and weakened retrieval.



(a) Multimodal RAG with Multimodal Embeddings and Separate Vector Stores.

Fig. 2. Config A uses MiniLM for text embeddings and SigLIP2 for image embeddings with separate vector stores.



(b) Multimodal RAG with Image Summaries and Combined Vector Store.

Fig. 3. Config B converts images into text summaries that are embedded with MiniLM in a unified index.

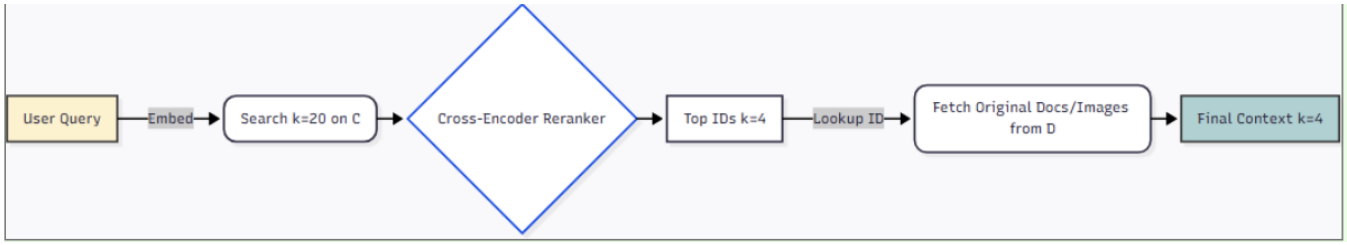


Fig. 4. Reranking pipeline from initial retrieval through Cross Encoder scoring to final context selection.

D. *Qualitative Observations*

The weakest results occur in Config A where SigLIP2 sometimes retrieves diagrams based on superficial visual similarity rather than functional relevance. For example, general wiring diagrams were retrieved for questions about specific heating components. Reranking mitigates this by prioritizing semantically aligned passages over visually similar ones, as reflected in Fig. 6.

Config B occasionally fails when Gemini summaries omit fine details such as the number of valves or the orientation of switches. In such cases, the missing details lead to inaccurate answers even though the retrieval mechanism behaved correctly, as illustrated in Fig. 5.

VI. Results

Table I shows the quantitative averages.

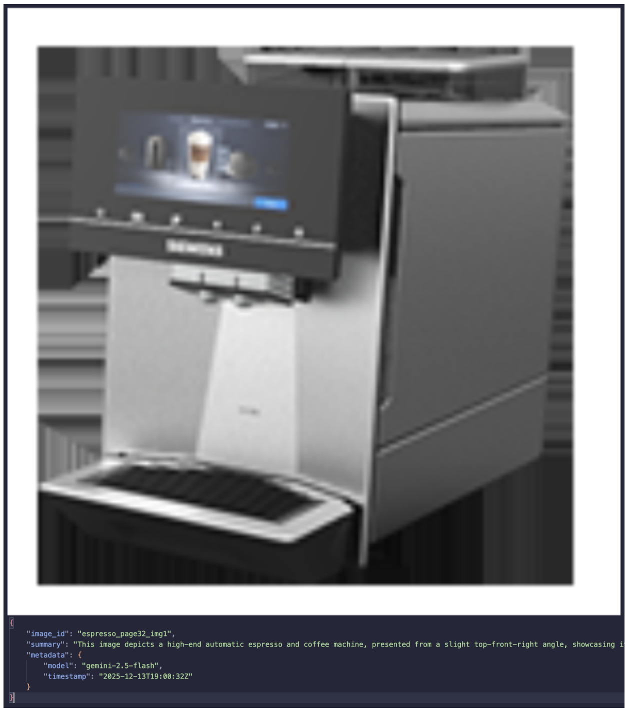
The summary based configurations show stronger performance than raw image embedding configurations. Reranking produces noticeable improvements in image faithfulness and text context relevancy, consistent with the aggregate trends in Fig. 7.

TABLE I
Evaluation Metrics by Configuration

Config	Corr	Rel	TxtF	ImgF	CtxR
A	0.556	1.000	0.889	0.278	0.722
A + Rerank	0.500	1.000	1.000	0.278	0.722
B	0.611	0.944	1.000	0.333	0.722
B + Rerank	0.556	1.000	0.944	0.444	0.778

VII. Discussion

The experimental findings validate that optimization of the retrieval step significantly enhances the reliability of MM-RAG in technical domains. Through the testing, we found that Configuration B (Unified Vector Store with Textual Summaries) was the superior choice when compared to Configuration A (Separate Multimodal Embeddings). This current finding confirms the challenge of combining separate embedding spaces, specifically one for text and one for raw images (SigLIP2). Dealing with the complex language of industrial diagrams exposed this fundamental flaw of trying



```
{
    "image_id": "espresso_page32_img1",
    "summary": "This image depicts a high-end automatic espresso and coffee machine, presented from a slight top-front-right angle, showcasing i
    "metadata": {
        "model": "gemini-2.5-flash",
        "timestamp": "2025-12-13T19:00:32Z"
    }
}
```

Fig. 5. Example industrial diagram (top) and the corresponding Gemini generated textual summary (bottom).

to force these separate modalities to retrieve a combined, aligned document.

The key breakthrough and confirmation was the text summarization configuration. By utilizing Gemini 2.5 Flash to transform visual diagrams into textual summaries, we converted the visuals into a format that allowed for a single text-based embedding space to be utilized. This summarization step avoided the issue of needing the separate embedding spaces to be aligned, creating a more stable knowledge base for the entire pipeline. The summarization step further enabled for precise customization of the retrieval query. During the summarization process, we are able to engineer the query to capture specific visual features, descriptions, or part names. This granular control is lost when utilizing raw image embeddings, where retrieval is simply based on visual similarity.

The integration of the BAA/bge-reranker-base [6] cross encoder served as the necessary cleanup to address the retrieval noise that the initial configuration couldn't eliminate. The initial bi-encoder retrieval (Chroma DB similarity search) was designed to get all documents, essentially casting a wide net to ensure that the correct context was in the top documents retrieved. The following reranking process uses a joint attention mechanism. Unlike the simple distance-based score of the bi-encoder, the cross-encoder evaluates the query and the retrieval documents together, which leads to a much more reliable assessment of actual relevance. This filtering step reduces the initial k=20 documents down to the final k=4 documents, increasing the precision of retrieval. The impact can be clearly seen with the improvements in the Image Faithfulness and Context Relevancy for Config B + Rerank setting highlight the reranker's success in capturing the most relevant passage. This provides the Gemini 2.5 Flash model with documents that are highly relevant and aligned with the user's query.

Qualitative analysis further highlighted these strengths and demonstrated distinct vulnerabilities with each configuration. Config A often failed because SigLIP2, the image encoder, retrieved diagrams based on low level visual similarity like retrieving a generic schematic when a specific pressure gauge diagram with measurements was needed. This was the visual noise that the reranking helped in eliminating. Config B had a slightly different issue as its retrieval noise came form the summarization. The Gemini 2.5 Flash model, when describing the image, sometimes omitted fine details like a specific label, and structural relationships, this leads to incomplete contexts. This was a key issue, as while a combined embedding space simplified the search process, the system's performance bottleneck becomes dependent on the details provided by the initial LLM used for summarisation.

VIII. Limitations

A. Image Faithfulness

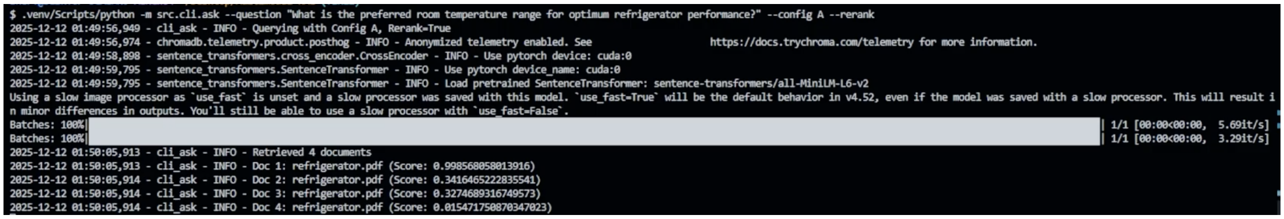
Even with our most successful configuration (Config B + Rerank), the Image faithfulness metric had the lowest performance metrics, lagging behind all others. This points to a fundamental limitation, that the Gemini 2.5 Flash generation layer, even when given correct context, failed to accurately generate details directly from the retrieved visual context or the textual summary. While we were focused on the retrieval problem, the model's ability to ensure that the generated answers are aligned with the image context remains a significant challenge.

B. Scope of the Data

Evaluation was conducted with a highly specific, limited collection of just 35 question-answer tasks generated from six consumer appliance manuals (e.g., espresso machine, water heaters). While these are aligned with our experiments, they limit our ability to make specific claims about this multimodal pipeline's performance in industrial environments. Real-world specialized engineering documents are often dense, cluttered with complex notations, and include specialized structures that were not fully tested here. We have only proven the effectiveness of this pipeline with a relatively small subset of technical documents.

C. Fixed Parameters

To ensure the controlled comparison across all configurations, the initial retrieval (k=20) and the final reranked (k=4) parameters remained constant. This was a necessary constraint, but it could have resulted in suboptimal performance. For example, in Configuration B's combined index, a query might retrieve 4 highly relevant text summaries but exclude a document that holds a single critical procedure step. The fixed size of k=4 might not be the optimal context length for every query, and it highlights a tradeoff between experimentation and real-world use cases.



```
$ .venv/Scripts/python -m src.cli.ask --question "What is the preferred room temperature range for optimum refrigerator performance?" --config A --rerank
2025-12-12 01:49:56,949 - cli_ask - INFO - Querying with Config A, Rerank=True
2025-12-12 01:49:56,974 - chromadb.telemetry.product.posthog - INFO - Anonymized telemetry enabled. See                    https://docs.trychroma.com/telemetry for more information.
2025-12-12 01:49:58,898 - sentence_transformers.cross_encoder.CrossEncoder - INFO - Use pytorch device: cuda:0
2025-12-12 01:49:59,795 - sentence_transformers.SentenceTransformer - INFO - Use pytorch device_name: cuda:0
2025-12-12 01:49:59,795 - sentence_transformers.SentenceTransformer - INFO - Load pretrained SentenceTransformer: sentence-transformers/all-MiniLM-L6-v2
Using a slow image processor as `use_fast` is unset and a slow processor was saved with this model. `use_fast=True` will be the default behavior in v4.52, even if the model was saved with a slow processor. This will result i
n minor differences in outputs. You'll still be able to use a slow processor with `use_fast=False`.
Batches: 100%|██████████████████████████████████████████████████| 1/1 [00:00<00:00,  5.69it/s]
Batches: 100%|██████████████████████████████████████████████████| 1/1 [00:00<00:00,  3.29it/s]
2025-12-12 01:50:05,913 - cli_ask - INFO - Retrieved 4 documents
2025-12-12 01:50:05,913 - cli_ask - INFO - Doc 1: refrigerator.pdf (Score: 0.998568058013916)
2025-12-12 01:50:05,914 - cli_ask - INFO - Doc 2: refrigerator.pdf (Score: 0.3416465222835541)
2025-12-12 01:50:05,914 - cli_ask - INFO - Doc 3: refrigerator.pdf (Score: 0.3274689316749573)
2025-12-12 01:50:05,914 - cli_ask - INFO - Doc 4: refrigerator.pdf (Score: 0.015471750870347023)
```

Fig. 6. CLI output showing the top k retrieved documents with reranking enabled.

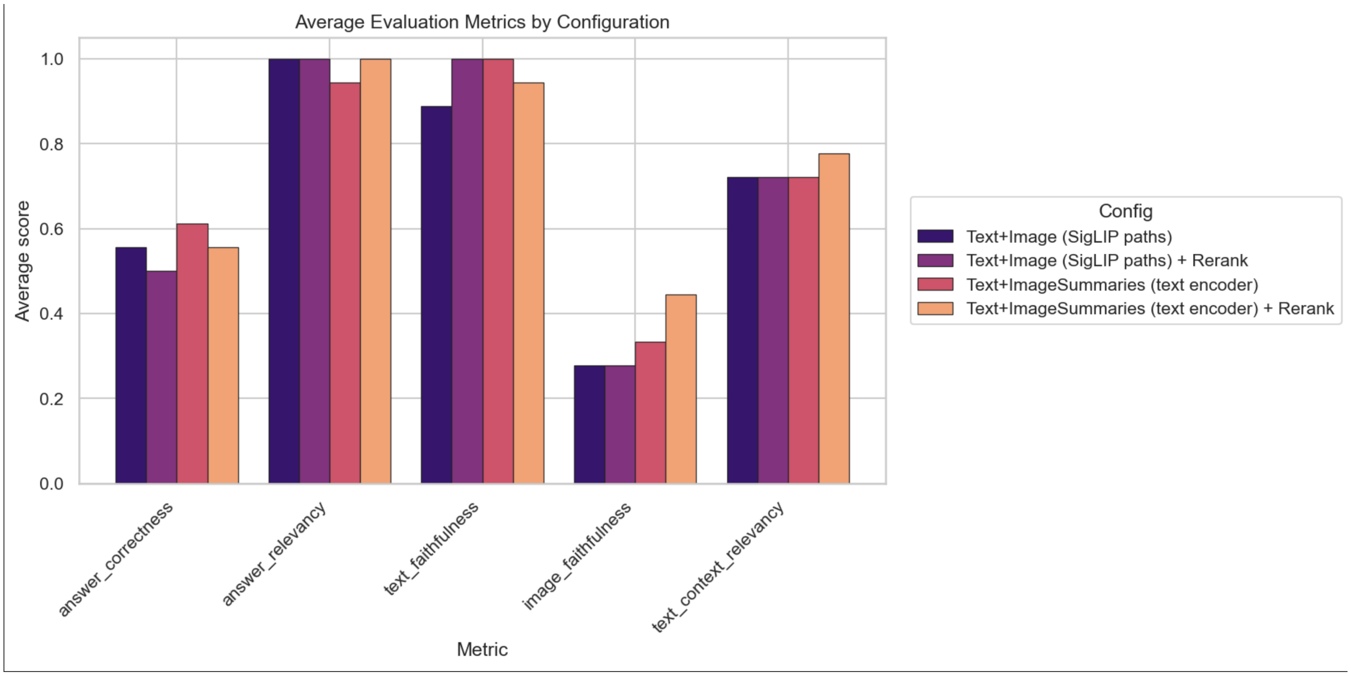


Fig. 7. Average scores across six metrics for all four configurations.

D. Model Dependency

Finally the success of Configuration B creates a vulnerability in latency and dependency. The pipeline's accuracy is dependent upon the Gemini 2.5 Flash summaries providing highly detailed descriptions. Generating these summaries is a prerequisite and it creates a const in time before actual indexing can begin. Another issue is that being reliant on a proprietary MLLM for this function introduces a dependency that we cannot control or modify, unlike if we used a custom, in house image encoder.

IX. Future Work Roadmap

With the limitations identified, such as the issues regarding Image Faithfulness, they provide a clear roadmap for the next phase of research. Future work must focus on delving beyond the limitations of text-only process in the reranking step.

A. Multimodal Rerankers

The current architecture is reliant on the BAAI/bge-reranker-base cross-encoder which is specialized in reranking textual content and this is a point of information loss. The next critical step is to investigate and integrate true multimodal rerankers, models that are capable of processing the original image simultaneously with the query and the textual content. Using direct visual processing at the reranking step could help avoid the information bottleneck introduced during the MLLM summarization step, potentially leading to a boost in the Image Faithfulness metric.

B. Specialized Encoding and Indexing

Future work should move beyond general models like SigLIP2, and more specialized tools should be built. This means fine-tuning image encoders on large, diverse datasets of industrial diagrams and schematics. The training process should involve the model recognising technical features specific to the industrial domain like flow lines, circuit symbols, and precise measurements. To further manage the metadata associated with large documents, exploration of structured retrieval indexes that includes hierarchical document structures (e.g., chapter/section titles) and page alignment should be done to optimize retrieval beyond simple similarity.

C. Performance Benchmark

Finally, future work must move beyond the current limited document size. The system needs to be tested rigorously on larger, more complex, and diverse groups of industrial documents to definitively confirm the results. Experiments on how well the system scales when handling thousands of manuals, checking both performance and how quickly it can answer queries to ensure it meets real-world expectations. The goal is to move this research from a proof of concept to a deployable, enterprise grade solution.

Conclusion

In this work, we present a multimodal retrieval and generation system designed to support question answering over technical manuals that combine textual instructions with complex visual content such as diagrams and schematics. By explicitly separating and comparing different indexing strategies for text and images, and by evaluating the impact of reranking, we provide a systematic analysis of how design choices affect grounded reasoning in technical domains.

Our experiments on publicly available manuals demonstrate that multimodal retrieval is necessary for accurately answering questions that rely on both procedural text and visual structure. The results show that reranking consistently improves retrieval quality across configurations, leading to more relevant context and more faithful answers. In particular, configurations that better align retrieved content with the query reduce hallucination and improve both text and image faithfulness.

This study highlights the importance of careful prompt design, evidence based generation, and retrieval strategies when applying large language models to technical documentation. While our system uses simple and constrained prompts, the evaluation shows that reliable performance can be achieved without aggressive prompt engineering. Future work may explore larger datasets, domain specific adaptations, and alternative summarization or reranking approaches to further improve robustness and scalability.

APPENDIX: PROMPT ENGINEERING

D. Image Summary Prompt

Describe the major components, labels, and functional purpose of the following diagram. Focus strictly on what is visually present, including identifiable parts, connections, and annotations. Be concise but complete. Avoid speculation, inferred behavior, or assumptions beyond the visible content.

E. Generation Prompt

Use only the retrieved textual and visual context to answer the question. Cite specific pieces of evidence from the provided sources to support your response. If the retrieved context does not contain sufficient information to answer the question with confidence, explicitly state that the answer is unknown rather than inferring or hallucinating details.

F. Prompt Design and Modifiability

Prompt wording directly affects how the model summarizes diagrams and generates answers. In technical manuals, loose or overly detailed prompts can lead to the introduction of components, steps, or behaviors that are not present in the source material. To reduce this risk, we use concise and clearly constrained prompts that emphasize observable content, explicit evidence, and acknowledgment of missing information.

The image summary prompt restricts output to visible elements and labeled structures, helping prevent imagined parts or functions. The generation prompt enforces reliance on retrieved context and discourages unsupported inference by requiring explicit citation and allowing unknown answers when evidence is insufficient. This design prioritizes accuracy and faithfulness over verbosity.

All prompts in the system are modular and can be adjusted without changing the underlying retrieval or generation pipeline. This allows future work to explore alternative prompt formulations, such as adding domain specific language or requesting more detailed descriptions, while using the current prompt set as a conservative baseline that favors grounded and reliable outputs.

References

[1] M. Riedler, S. Ma, S. Jiang, Y. Dong, et al., "Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications," arXiv preprint arXiv:2410.21943, 2024.

[2] H. Wang, Y. Lan, and W. Zhu, "MiniLM: Deep Self-Attention Distillation for Transformer Compression," in *NeurIPS*, 2020.

[3] A. Alayrac et al., "Scaling Vision-Language Models with Momentum Distillation," arXiv preprint arXiv:2306.07915, 2023.

[4] M. Tschannen et al., "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding and Localization," arXiv preprint arXiv:2502.14786, 2025.

[5] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. ICML*, 2021.

[6] BAAI Team, "BGE Rerankers: Efficient Cross Encoders for Scoring Document-Query Pairs," Model Documentation, 2024.

[7] Chroma, "Chroma: The Open-Source Embedding Database," Software Documentation, 2025.

[8] Gemini Team, "Gemini 2.5: Advanced Reasoning, Multimodality, Long Context, and Agentic Capabilities," arXiv preprint arXiv:2507.06261, 2025.