Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Visual, OCR-Text, and Hybrid Evidence Retrieval for Document Visual Question Answering: A Controlled Failure Analysis

Richmond Ampah-Mensah

richmond.a.m@unimus.ac.id

Universitas Muhammadiyah Semarang

Muhammad Munsarif

Universitas Muhammadiyah Semarang

Muhammad Sam`an

Universitas Muhammadiyah Semarang

Research Article

Keywords: document visual question answering, document retrieval, visual retrieval, OCR-text retrieval, multimodal retrieval, failure attribution

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10707422/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Visual, OCR-Text, and Hybrid Evidence Retrieval for Document Visual Question Answering: A Controlled Failure Analysis

Richmond Ampah-Mensah [1,*], Muhammad Munsarif [1], Muhammad Sam`an[1]

[1]Informatics, Universitas Muhammadiyah Semarang, Semarang, Indonesia

***Corresponding Author:** Email: richmond.a.m@unimus.ac.id

ORCID Identifiers: Richmond Ampah-Mensah: 0009-0007-2124-3262; Muhammad Munsyarif: 0000-0002-3009-4234; Muhammad Sam`an: 0000-0001-5408-5562

Abstract

Purpose: Document visual question answering (DocVQA) systems can fail because relevant evidence is not retrieved or because the answer generator fails to use correctly retrieved evidence. This study investigates how lexical, dense text, visual, and hybrid page retrieval affect evidence selection and downstream answer quality, and whether correct-page access resolves remaining generation errors.

Methods: BM25, BGE-M3, MiniLM, ColQwen, and reciprocal-rank fusion were evaluated on a document-disjoint DocVQA protocol. Fusion parameters were selected using development data and fixed before final evaluation. Answer generation used Qwen2.5-VL under matched evidence conditions from BM25, ColQwen, fusion, and oracle pages. Retrieval performance was additionally evaluated on MP-DocVQA without retuning.

Results: ColQwen achieved the strongest page-retrieval performance on DocVQA, with Recall@1 of 0.559 compared with 0.335 for BM25 and 0.274 for BGE-M3. Reciprocal-rank fusion did not improve over ColQwen. Better retrieved evidence translated into higher answer quality, with ANLS increasing from 0.127 using BM25 pages to 0.213 using ColQwen pages. Oracle evidence further increased ANLS to 0.395, showing that substantial answer-generation errors remained even when the annotated page was available. The visual retrieval advantage also transferred to MP-DocVQA.

Conclusion: Visual page retrieval improved evidence selection and downstream answering in the evaluated settings, but retrieval alone did not resolve DocVQA errors. Fixed fusion provided no additional benefit, while the oracle results show that answer generation remains an important source of failure.

Keywords: document visual question answering; document retrieval; visual retrieval; OCR-text retrieval; multimodal retrieval; failure attribution;

1. Introduction

Documents communicate through more than recognized words. Typography, spatial grouping, reading order, tables, figures, and page-level context all influence how information is interpreted. Document visual question answering therefore differs from ordinary passage question answering because the system must first identify the page that contains the relevant visual evidence before it can produce a supported answer.

This distinction matters in multi-stage systems. An incorrect answer may result because the retriever selects an irrelevant page, or because the correct page is retrieved but the generator fails to locate, normalize, or reason over the relevant content. End-to-end scores merge these failure modes and can make it difficult to determine whether an improvement comes from better evidence selection or better answer generation.

Recent visual document retrievers encode page images directly and use late interaction between query and page representations [1]. Multimodal document RAG systems have shown that retaining rendered pages can help when evidence depends on layout, tables, charts, or other visual regions [2]. Recent benchmarks have also expanded evaluation to long-document, multilingual, cross-domain, and efficiency-oriented settings [3–6]. At the same time, direct comparisons report meaningful trade-offs between text and visual retrieval rather than universal visual dominance [7–10]. The relevant empirical questions are therefore when each evidence route succeeds, whether fixed fusion improves the strongest component, and how much answer error remains after the correct page is available.

We investigate how visual and OCR-derived document representations affect evidence selection and downstream answer generation under a controlled multimodal DocVQA pipeline. To isolate these effects, lexical, dense-text, and visual retrievers are evaluated on identical queries and candidate pages. Fusion parameters are selected only on development data, the final evaluation is document-disjoint, answer generation is tested on a representative paired sample with an oracle-page control, and the retrieval comparison is repeated on MP-DocVQA without retuning. Separate balanced samples are used only for conditional diagnostics and error analysis.

The study makes three contributions:

1. It provides a controlled comparison of BM25, BGE-M3, MiniLM, ColQwen, and Reciprocal Rank Fusion (RRF) on a common DocVQA corpus, with paired uncertainty estimates and efficiency measurements.
2. It evaluates answer generation on the same 1,000 randomly sampled questions under BM25, ColQwen, RRF, and oracle evidence, allowing retrieval-associated losses to be separated from failures that persist when the annotated page is supplied.
3. It combines this paired failure decomposition with diagnostic error analysis and external MP-DocVQA validation, testing whether the visual-retrieval advantage and the fusion result transfer beyond the primary split.

The resulting claim is intentionally bounded. Under the evaluated DocVQA and MP-DocVQA protocols, visual page embeddings selected evidence more reliably than the tested OCR-text baselines. Fixed RRF did not improve on the strongest visual retriever, and substantial answer error remained even with oracle evidence. The contribution is therefore diagnostic: it clarifies which failures arise during page selection and which remain after relevant evidence reaches the generator.

2. Related Work

2.1 Document visual question answering

DocVQA formalized question answering over document images and showed a pronounced gap between model and human performance, particularly for questions requiring structural document understanding [11]. Document Collection VQA expanded the task to evidence retrieval over collections [12]. MP-DocVQA then introduced multi-page documents and treated page localization as part of an explainable answer process [13]. These benchmarks motivate evaluating page selection separately from answer generation.

OCR-first document models combine recognized tokens with layout or image features. LayoutLM and its successors demonstrated that two-dimensional position and visual context improve document understanding beyond plain text [14–16]. OCR-free models such as Donut and Pix2Struct instead learn directly from rendered

documents or screenshots [17, 18]. These lines of work establish that visual structure is informative, but they do not by themselves identify which stage failed in a retrieval-augmented QA pipeline.

2.2 Text and visual document retrieval

Text retrieval provides mature lexical and dense baselines. BM25 remains effective when exact names, identifiers, dates, and rare terms are decisive [19]. Sentence-BERT and MiniLM provide compact dense representations [20, 21], while BGE-M3 supports longer inputs and multiple retrieval functions [22]. We therefore use BM25 and BGE-M3 as the principal OCR-text baselines and retain MiniLM as a compact reference point.

ColBERT introduced token-level late interaction for text retrieval [23]. ColPali extended this principle to page images, using a vision-language encoder to represent visually rich documents without an OCR indexing stage [1]. The ColQwen checkpoint evaluated here belongs to that model family. M3DocRAG combined multimodal retrieval with a vision-language generator for multi-page and multi-document QA and showed the value of preserving visual evidence [2]. MMDocRAG broadened evaluation to multi-page, cross-modal evidence chains and highlighted persistent retrieval, selection, and integration failures [24].

Recent comparisons show that modality effects depend on the corpus and task. López et al. reported different gains for text and visual RAG variants across multi-page DocVQA benchmarks [7]. Most et al. identified efficiency and generalization trade-offs between OCR-dependent and vision-based retrieval under changing document quality [8]. IRPAPERS found complementary text and image retrieval errors and a positive hybrid gain on scientific papers, while its text-based QA pipeline outperformed its image-based counterpart [10]. Related OCR robustness work further shows that low character or word error does not necessarily preserve downstream retrieval quality when structural information is lost [9]. Together, these findings motivate dataset-specific comparison rather than a universal claim that one modality is always superior.

2.3 Fusion and answer generation

Combining rankings is attractive when modalities fail on different pages, but text and visual retrieval scores are not naturally comparable. RRF avoids direct score addition by combining the source ranks of each candidate [25]. Whether fusion improves retrieval remains an empirical question because a weaker or correlated ranking can also displace strong candidates. We therefore select RRF parameters on development data and keep them fixed for all subsequent evaluation.

Qwen2.5-VL provides document, table, chart, and spatial understanding capabilities [26] and is used here as the answer generator. Correct page retrieval does not, however, guarantee a correct answer. Benchmarks such as ChartQA and TAT-DQA illustrate persistent challenges in visual, tabular, and numerical reasoning [27, 28]. The oracle-page condition in this study measures the residual error that remains when the annotated evidence page is supplied, while the retrieved-page conditions measure the downstream effect of evidence selection.

2.4 Relation to recent benchmarks and adaptive systems

Recent benchmarks broaden visual document retrieval beyond the settings considered here. MMDocIR evaluates page- and layout-level retrieval across visual and OCR-derived representations [3], while VisR-Bench extends evaluation to long multilingual documents and multiple evidence types [6]. UniDoc-Bench compares text-only, image-only, fusion, and joint multimodal retrieval over cross-domain PDFs and

reports consistent gains from text-image fusion in its evaluated settings [4]. These resources establish broader benchmark coverage than the present study.

Method-oriented work has also explored more adaptive evidence selection. AVIR combines page relevance, score-distribution clustering, and adaptive filtering before invoking a vision-language model [5]. In contrast, our experiments use off-the-shelf text and visual rankings and a fixed RRF rule so that the effect of evidence modality can be isolated without introducing a new retriever or adaptive selector.

The comparison with these studies is therefore one of experimental focus rather than competing scope. Broad benchmark and adaptive-method contributions ask how to improve multimodal retrieval across diverse settings; the present study asks where a controlled document-QA pipeline fails when retrieval, fusion, and generation are evaluated separately.

2.5 Position of this study

Visual document retrieval is already an established research direction, and text and page images are known to provide complementary signals. The contribution of this study lies in controlled failure attribution: stronger text baselines are compared on an identical corpus, fusion is selected on a separate development split, answer generation is evaluated under paired retrieved and oracle evidence, and the retrieval comparison is transferred to an external dataset without retuning. This design tests whether the retrieval advantage survives stronger baselines, whether fixed fusion adds value, and how much end-to-end error remains after the relevant page reaches the generator.

3. Study Design

3.1 Research questions

The experiments address four questions:

- **RQ1:** How do lexical OCR-text, dense OCR-text, and visual page retrieval compare on the same questions and candidate corpus?
- **RQ2:** Does a development-selected text–visual fusion rule improve over its strongest component on held-out and external data?
- **RQ3:** How much does evidence-page choice change answer quality, and how much failure remains when the gold page is supplied?
- **RQ4:** What kinds of errors occur among low-scoring answers, and which reported distributions support population inference versus conditional diagnosis only?

Figure 1 summarizes the evaluation framework. Development data are used only for fusion selection, while the document-disjoint final test and external evaluation measure the fixed retrieval systems. Answer-generation conditions are paired on the same randomly sampled questions, and balanced audit samples are reserved for diagnostic analysis.

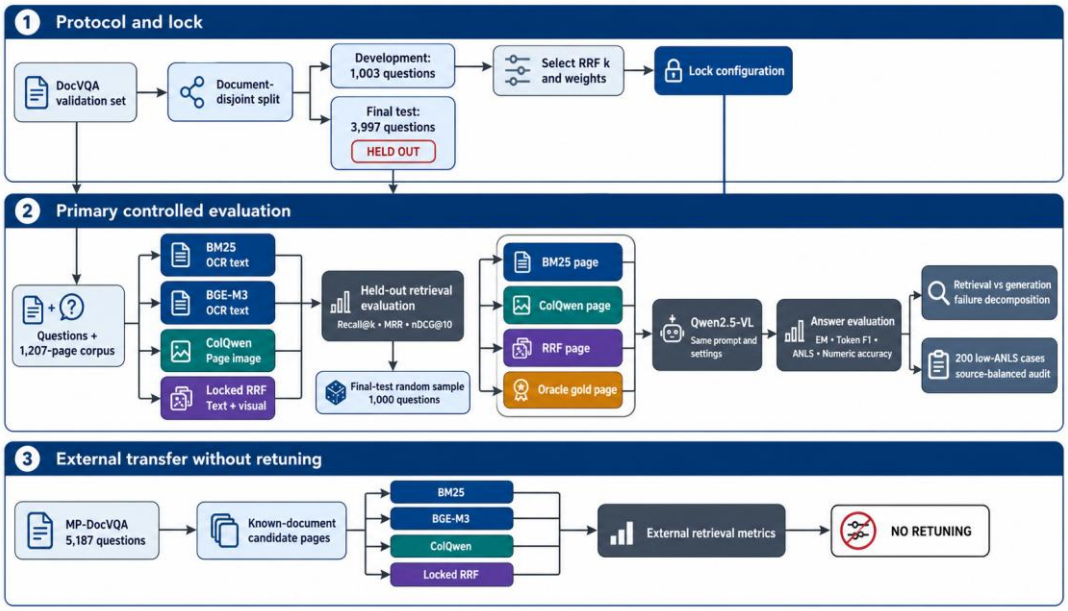


Figure 1. Controlled evaluation framework. Development questions are used only to select the RRF rule. The document-disjoint DocVQA final test and MP-DocVQA external evaluation use the same fixed retrieval configuration. On the random final-test sample, BM25, ColQwen, RRF, and oracle pages are passed to the same Qwen2.5-VL generator to separate evidence-selection effects from errors that remain with annotated evidence. Balanced audits are used only for conditional diagnosis.

3.2 DocVQA corpus and leakage-safe split

The source set comprised 5,000 questions from the DocVQA validation data [11]. Duplicate page records were removed before indexing, producing a canonical corpus of 1,207 pages. Questions were split by document_id rather than independently by question so that questions from the same document could not appear in both parameter selection and final testing. The resulting split contained 1,003 development questions from 241 document groups and 3,997 final-test questions from 966 groups, with no group overlap. The final set contained 1,939 numeric, 1,525 short-text, 509 long-text, and 24 yes/no answers.

All retrieval systems used the same canonical page identifiers, queries, candidate corpus, and gold page labels. Development data were used for fusion selection only; final-test queries were not used for parameter selection. The retrieval outputs used in the answer-generation experiments were fixed before generation was run.

3.3 Retrieval systems

BM25. The lexical OCR-text baseline indexed page text using Okapi BM25 with $(k_1 = 1.2)$, $(b = 0.75)$, and the tokenizer pattern \w+(?:[./:-]\w+)*. This tokenizer retains common identifier, date, path, and punctuation-linked forms. The system returned the top 100 pages.

BGE-M3. The principal dense OCR-text baseline used BAAI/bge-m3, normalized embeddings, maximum sequence length 2,048, and GPU inference. Inner-product search over normalized vectors returned the top 100 pages.

MiniLM. The compact dense baseline used sentence-transformers/all-MiniLM-L6-v2 with normalized embeddings, maximum sequence length 256, CPU inference, and top-100 retrieval. MiniLM is included as an additional reference

baseline; the text-versus-visual conclusions rely primarily on BM25 and BGE-M3.

ColQwen. Visual retrieval used vidore/colqwen2-v1.0-hf. Each rendered page was encoded as a multi-vector visual representation and matched to the question through late interaction. The same 1,207-page corpus was used, and the system returned the top 100 candidates per question.

3.4 Hybrid retrieval

Raw BM25 and ColQwen scores were not combined directly because they are produced on different scales. The primary hybrid therefore used weighted reciprocal-rank fusion. For candidate page d,

$$s_{RRF(d)} = \frac{w_t}{k + r_{t(d)}} + \frac{w_v}{k + r_{v(d)}}$$

where $r_{t(d)}$ and $r_{v(d)}$ are the one-based BM25 and ColQwen ranks. A missing source rank contributes zero. Candidates were drawn from the top 100 results of each source and ordered by decreasing fused score, with source rank and page identifier used for deterministic tie-breaking.

On the development set, k was evaluated at 10, 30, 60, and 100; the text weight at 0.25, 0.50, 0.75, and 1.00; and the visual weight was fixed at 1.00. Recall@5 was the selection metric and MRR the tie-break metric. The selected configuration used k=10 with text and visual weights of 0.25 and 1.00, respectively. Its development Recall@1, Recall@5, and MRR were 0.5015, 0.7248, and 0.6028. This configuration was then kept fixed for the final and external evaluations.

A secondary development ablation tested robust score normalization followed by weighted score fusion. The best normalized configuration reached Recall@5 of 0.6381, below the selected RRF result of 0.7248, and was not evaluated on the final test.

3.5 Retrieval metrics and uncertainty

Retrieval was correct at rank k when the gold page appeared among the top-k candidates. We report Recall@1, Recall@5, Recall@10, Recall@20, MRR, and nDCG@10. Paired differences were estimated at the query level with 10,000 bootstrap resamples. Reported latency is descriptive wall-clock search time from the recorded hardware and is not intended as a hardware-independent benchmark.

3.6 Answer-generation protocol

Answer generation used Qwen/Qwen2.5-VL-3B-Instruct [26]. Full-precision inference exceeded the available 8 GB GPU memory, so every evidence condition used the same 4-bit bitsandbytes NF4 quantization with bfloat16 computation and double quantization. The model, prompt, decoding procedure, query samples, and retrieval predictions were otherwise unchanged across evidence conditions.

For each run, the model received one page image and a fixed prompt instructing it to answer only from visible information and to return 'not enough evidence' when the page did not support an answer.

Decoding was deterministic (do_sample=false) with at most 48 new tokens. The main evaluation used a simple random sample without replacement of 1,000 questions from the 3,997-question final test. Each question was evaluated under four evidence conditions: the BM25 top-1 page, ColQwen top-1 page, RRF top-1 page, and the gold/oracle page. This produced 4,000 paired generation runs.

A separate conditional diagnostic sample contained 40 questions from each of four top-1 retrieval outcomes: text-only hit, visual-only hit, both hit, and neither hit. The same four evidence conditions were evaluated for each question. Because these retrieval outcomes were deliberately

balanced, the resulting analysis describes conditional behavior rather than population frequencies.

3.7 Answer metrics and paired comparisons

Answers were normalized consistently and scored against all accepted DocVQA answers, retaining the best score. We report exact match, token F1, average normalized Levenshtein similarity (ANLS), and numeric relaxed accuracy. Numeric relaxed accuracy was evaluated on the 476 numeric questions in the main sample. Paired system differences used 10,000 query-level bootstrap resamples.

For the failure decomposition, ANLS ≥ 0.5 was treated as a useful-answer threshold. Each retrieved-page outcome was assigned to one of five mutually exclusive categories:

1. evidence and generation success: the retrieved page was gold and the answer met the threshold;
2. generation failure given gold page: the retrieved page was gold but the answer failed;
3. answer success despite a non-gold page;
4. retrieval-associated loss supported by oracle: the retrieved-page answer failed but the oracle-page answer succeeded; or
5. unresolved wrong-page and oracle failure: the retrieved page was non-gold and both answers failed.

The retrieval-associated category indicates that the oracle answer succeeded when the retrieved-page answer did not; it does not establish retrieval as the sole cause. Cases in which both the retrieved-page and oracle answers failed remain unresolved because evidence selection and answer generation cannot be separated uniquely.

3.8 Manual diagnostic audit

Two hundred low-ANLS cases were sampled equally from the BM25, ColQwen, RRF, and oracle evidence conditions. Each case received one final label: wrong evidence page; answer-formatting or normalization mismatch; correct-page numeric or unit reasoning failure; missed text span; chart or visual interpretation failure; or table/reading-order failure. Because the sources were balanced by design, this audit identifies error mechanisms but does not estimate their prevalence in the full evaluation population.

All 200 cases received a computer-assisted primary review. Forty cases were reviewed a second time, and ten disagreements were adjudicated after the source key was opened. Before adjudication, label agreement was 77.5% (Cohen's kappa = 0.686) and visibility agreement was 77.5% (kappa = 0.582). Because AI-generated suggestions were visible during review and adjudication was computer-assisted, the audit was not fully blind or independent.

3.9 Exploratory OCR diagnostic

An exploratory audit examined 300 DocVQA pages, including table/form and prose/other pages and a subset of visual-only retrieval cases. The supplied OCR text was compared with computer-assisted verified transcriptions using character error rate (CER), word error rate (WER), token coverage, and numeric-token accuracy.

The reference transcriptions were initialized from machine-generated drafts, so this audit cannot provide an independent comparison of OCR engines. It is used only to characterize mismatch between the supplied OCR representation and visible page content. Apparent accuracy estimates for the machine system that contributed to the reference drafts are therefore not treated as valid benchmark results.

3.10 External validation

External retrieval validation used 5,187 MP-DocVQA validation questions over 6,510 pages [13]. Candidate pages were restricted to pages belonging to each question's document, corresponding to a page-localization setting rather

than open-corpus retrieval. BM25, BGE-M3, ColQwen, and the RRF configuration selected on DocVQA development data were evaluated without external retuning. Paired intervals used 10,000 query-level bootstrap resamples.

3.11 Implementation and reproducibility controls

Experiments were run in a WSL Python environment. The recorded dense and visual retrieval runs used an NVIDIA GeForce RTX 4060 Laptop GPU, and Qwen generation used the uniform 4-bit configuration described above. The retrieval environment used PyTorch 2.12.0 with CUDA 13.0. Reproducibility materials include the data-split manifests, resolved configurations, model versions, prompts, retrieval predictions, generation outputs, bootstrap tables, and annotation records. Dataset images remain subject to their original licenses and redistribution terms.

4. Results

4.1 Dataset and evaluation scope

Table 1 summarizes the role of each dataset and sample. The 1,000-question simple random sample supports population-oriented answer estimates for the sampled final-test population, while the balanced samples are used only for conditional and diagnostic analyses.

Table 1. Evaluation datasets and sampling roles.

Analysis	Dataset	Pages	Questions	Role
Fusion development	DocVQA	1,207 shared corpus	1,003	Parameter selection only
Primary retrieval	DocVQA	1,207 shared corpus	3,997	Held-out final test
Main answer generation	DocVQA	top-1 or oracle page	1,000 x 4 sources	Simple random paired sample
Conditional answer diagnostic	DocVQA	top-1 or oracle page	160 x 4 sources	40 per retrieval bucket; not population weighted
Manual low-ANLS audit	DocVQA	case-specific page	200	50 per source; not a population distribution
OCR audit	DocVQA	300	n/a	Stratified assisted transcription audit
External retrieval	MP-DocVQA	6,510	5,187	External validation; document-restricted candidates

4.2 Primary DocVQA retrieval

Table 2 reports retrieval on the held-out DocVQA test set. ColQwen was the strongest system across the listed effectiveness metrics, reaching Recall@1 of 0.559 compared with 0.335 for BM25 and 0.274 for BGE-M3. This gain came with a substantial efficiency cost: ColQwen required 349.5 ms per query and a 444 MB page-embedding cache, whereas BM25 required 1.60 ms per query and a 0.73 MB index.

Table 2. Retrieval on the 3,997-question DocVQA final test.

System	Evidence representation	R@1	R@5	R@10	nDCG@10	MRR	Search ms/query	Index/cache MB

BM25	OCR text, lexical	0.3350	0.4691	0.5144	0.4218	0.3983	1.60	0.73
BGE-M3	OCR text, dense	0.2737	0.4071	0.4656	0.3629	0.3390	6.27	4.71
MiniLM	OCR text, dense legacy	0.1749	0.3005	0.3628	0.2605	0.2366	2.57	1.77
ColQwen	Page image, late interaction	**0.5592**	**0.7378**	**0.7866**	**0.6716**	**0.6398**	349.47	444.03
RRF	BM25 plus ColQwen ranks	0.5347	0.7346	0.7863	0.6610	0.6256	n/a	component indexes

The paired ColQwen-minus-BM25 gain was 0.224 at Recall@1 (95% CI 0.210 to 0.239), 0.269 at Recall@5 (0.254 to 0.283), and 0.250 at nDCG@10 (0.238 to 0.262). Against BGE-M3, the Recall@1 gain was 0.286 (0.271 to 0.301). The primary paired intervals exclude zero.

The advantage persisted across heuristic question types. For example, ColQwen versus BM25 Recall@1 was 0.350 versus 0.214 for chart/figure questions (117 cases), 0.635 versus 0.396 for number/date questions (1,521 cases), and 0.508 versus 0.283 for text-lookups (1,991 cases). These subgroup labels are rule-based and should be interpreted descriptively.

4.3 Fusion outcome and modality overlap

RRF did not improve on ColQwen in the final test. Recall@1 was lower by 0.025 (95% CI 0.017 to 0.032 in absolute magnitude), while the Recall@5 difference was small and its interval included zero. The same pattern was observed for nDCG@10 and MRR, where ColQwen remained higher.

At rank 1, BM25 and ColQwen were both correct for 1,255 queries, BM25 alone for 84, ColQwen alone for 980, and neither for 1,678. RRF preserved all 1,255 shared successes but recovered only 26 of the 84 BM25-only cases while retaining 820 of the 980 ColQwen-only cases. Thus, the text ranking sometimes displaced pages that the visual retriever had ranked correctly, explaining why component complementarity did not translate into a net fusion gain.

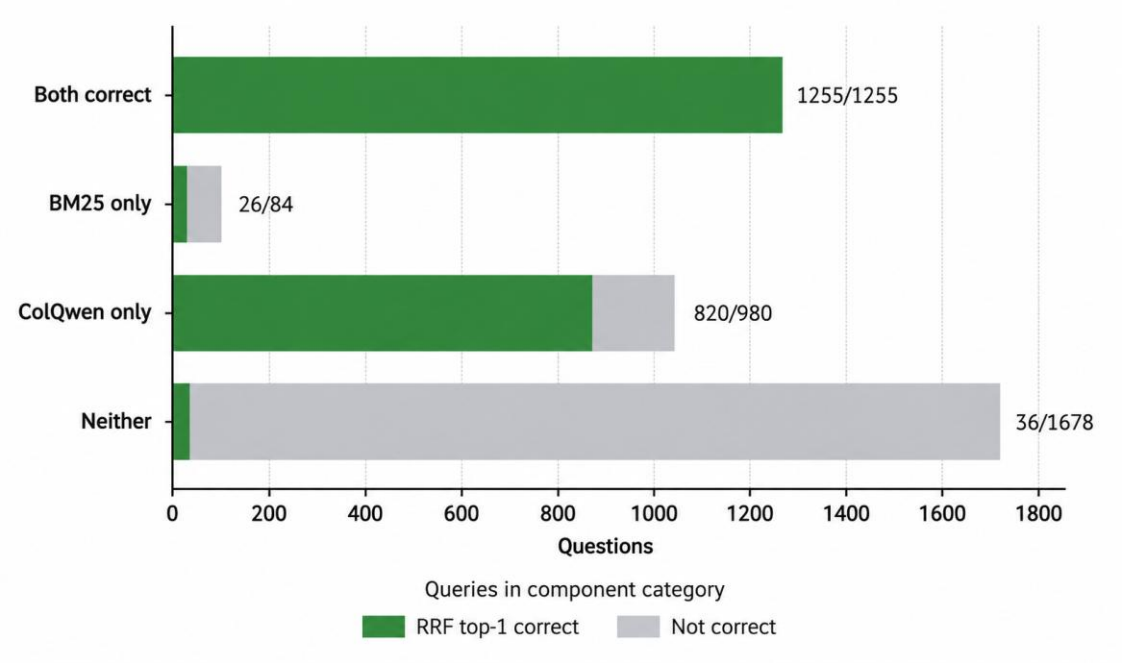


Figure 2. Rank-1 overlap between BM25 and ColQwen and the corresponding RRF successes. Gray bars show all queries in each component category; green bars show cases in which RRF ranked the gold page first.

4.4 Representative answer generation

Table 3 reports answer quality on the 1,000-question random sample. ColQwen evidence improved all reported answer metrics over BM25 evidence. ANLS increased by 0.087 (95% CI 0.068 to 0.107), exact match by 0.070 (0.053 to 0.088), and numeric relaxed accuracy by 0.219 (0.179 to 0.261). RRF and ColQwen produced similar downstream answer quality; their paired ANLS and exact-match differences both had intervals that included zero.

Table 3. Qwen2.5-VL results on the paired 1,000-question random sample.

Evidence source	Gold-page rate	Exact match	Token F1	ANLS	Numeric relaxed accuracy (n=476)
BM25 top-1	0.309	0.104	0.175	0.127	0.305
ColQwen top-1	0.541	**0.174**	**0.282**	**0.213**	**0.523**
RRF top-1	0.515	0.172	0.277	0.211	0.506
Gold/oracle page	1.000	0.336	0.494	0.395	0.824

Oracle evidence substantially improved answer quality but did not eliminate error. Relative to ColQwen evidence, oracle ANLS increased by 0.182 (95% CI 0.158 to 0.206) and exact match by 0.162 (0.139 to 0.186). Even so, oracle exact match was 0.336, and only 424 of 1,000 oracle answers reached ANLS $\geq$ 0.5. Substantial residual error therefore remained under the evaluated model, prompt, image-processing pipeline, quantization, and answer metric even when the annotated page was supplied.

4.5 Conditional diagnostic behavior

The balanced diagnostic sample showed the expected dependence on evidence choice. In text-only retrieval successes, BM25-page ANLS was 0.377 compared with 0.147 for ColQwen pages; in visual-only successes, ColQwen-page ANLS was 0.432 compared with 0.025 for BM25 pages. For neither-hit cases, retrieved-page ANLS remained below 0.100 while oracle ANLS rose to 0.450. These controlled contrasts demonstrate that evidence-page choice materially affects answer quality, but the balanced bucket sizes should not be interpreted as final-test frequencies.

4.6 Retrieval and generation error distribution

Table 4 separates the 1,000 paired outcomes using the ANLS ≥ 0.5 threshold. ColQwen reduced retrieval-associated losses relative to BM25 (195 versus 293 cases) and increased successes obtained from the gold page (211 versus 115). Because ColQwen retrieved the gold page more often, more cases also exposed answer failure despite correct page retrieval (330 versus 194), making the residual answer-generation problem easier to observe.

Table 4. Error decomposition on 1,000 paired questions using ANLS at least 0.5. Counts are also percentages because n=1,000 per source.

Mutually exclusive outcome	BM25	ColQwen	RRF
Gold page and answer success	11.5	21.1	20.7
Generation failure given gold page	19.4	33.0	30.8
Answer success despite non-gold page	2.5	2.4	2.6
Retrieval-associated loss supported by oracle	29.3	19.5	19.9
Non-gold page and oracle failure; unresolved	37.3	24.0	26.0

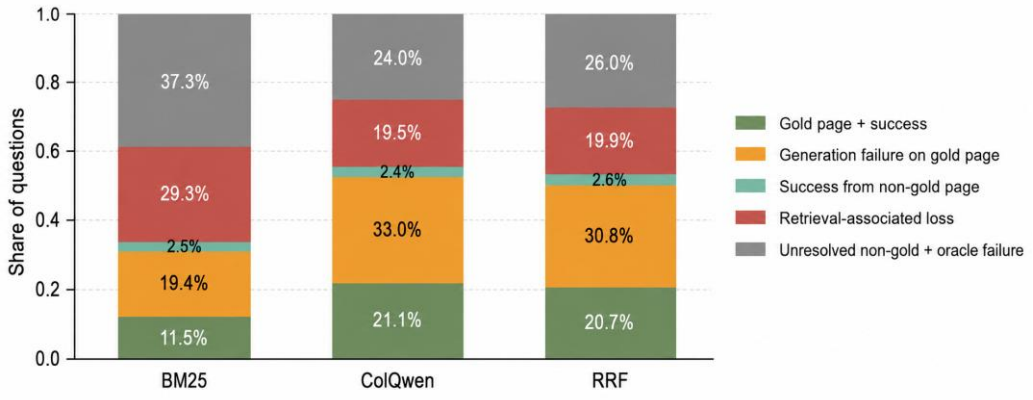


Figure 3. Mutually exclusive retrieval and generation outcomes using ANLS at least 0.5. Each bar contains 1,000 paired queries.

The oracle condition succeeded in 424 cases and failed to reach the threshold in 576. The retrieved-page outcomes therefore cannot be interpreted simply as a count of wrong-page errors: some non-gold pages still supported successful answers, while cases that also failed under oracle evidence cannot be attributed uniquely to retrieval.

4.7 Manual low-ANLS taxonomy

Among the 200 source-balanced low-ANLS cases, 82 were labelled as answer-formatting or normalization mismatches, 72 as wrong-evidence

pages, 26 as correct-page numeric or unit failures, 13 as missed text spans, five as chart or visual interpretation failures, and two as table or reading-order failures. Answer-bearing information was judged visible in 110 cases, absent in 72, and uncertain in 18.

Because each evidence source contributed the same number of audited cases, these percentages describe the diagnostic sample rather than the population of DocVQA failures. The audit nevertheless shows that low automatic scores can arise from several distinct mechanisms, including semantically equivalent answer forms and correct answers accompanied by additional wording.

4.8 Exploratory OCR diagnostic

In the assisted 300-page audit, the supplied OCR had macro CER of 0.550, WER of 0.722, token coverage of 0.735, and numeric-token accuracy of 0.599. The sample deliberately included difficult pages, including visual-only retrieval cases, so these values characterize the audit sample rather than an unbiased corpus-wide OCR benchmark. Because the reference transcriptions were machine-assisted, the analysis supports only the narrower conclusion that the supplied OCR representation was incomplete on many of the sampled pages.

4.9 External MP-DocVQA validation

Table 5 reports external MP-DocVQA retrieval with document-restricted candidates and no retuning. ColQwen again achieved the highest Recall@1, MRR, and nDCG@10. Its Recall@1 advantage was 0.132 over BM25 (95% CI 0.120 to 0.143) and 0.144 over BGE-M3 (0.132 to 0.157). RRF remained slightly below ColQwen at Recall@1.

Table 5. MP-DocVQA validation retrieval with document-restricted candidates and no external retuning.

System	R@1	R@5	R@10	MRR	nDCG@10
BM25	0.7179	0.9265	0.9719	0.8097	0.8479
BGE-M3	0.7054	0.9381	0.9751	0.8043	0.8449
ColQwen	**0.8496**	**0.9740**	0.9902	**0.9054**	**0.9260**
RRF	0.8442	0.9728	**0.9907**	0.9024	0.9239

The MP-DocVQA result supports transfer of the visual-over-text retrieval ordering under a within-document page-localization protocol. It does not establish open-corpus retrieval performance. RRF also remained below ColQwen, showing that the fixed fusion configuration did not improve on the visual retriever in either evaluated dataset.

5. Discussion

5.1 Visual evidence retrieval survived stronger text baselines

ColQwen outperformed both lexical BM25 and dense BGE-M3 on identical held-out DocVQA queries and candidate pages, and the same ordering transferred to MP-DocVQA. This strengthens the interpretation that the visual retriever captured useful page-level evidence that was not consistently preserved in the tested OCR-text representations. The finding is specific to the evaluated protocols rather than a general ranking of retrieval modalities.

The effectiveness gain also involved a clear deployment trade-off. BM25 was substantially cheaper in search time and index size, and it retrieved 84 final-test pages at rank 1 that ColQwen missed. Other corpora can favor text retrieval or produce positive hybrid gains [7, 8, 10]. In practice, the results support visual retrieval when

evidence coverage is the priority, while highlighting the value of cheaper text retrieval for exact lexical matches and resource-constrained settings.

5.2 Complementarity does not guarantee a fusion gain

The overlap analysis confirms that BM25 and ColQwen make complementary errors, but complementarity alone was not enough for RRF to improve the strongest component. On the DocVQA final test, the fused ranking recovered fewer than one third of BM25-only rank-1 successes and displaced 160 ColQwen-only successes. The same ordering reappeared in the external evaluation.

This result contrasts with the positive hybrid gain reported by IRPAPERS [10]. One important difference is the balance between the component retrievers. In IRPAPERS, the reported text and image systems had similar Recall@1 values, whereas ColQwen was substantially stronger than BM25 and BGE-M3 in our DocVQA evaluation. The corpora, question types, candidate settings, and fusion implementations also differ. These results suggest that fusion is most useful when component rankings contribute sufficiently strong and complementary evidence; the presence of exclusive hits alone does not guarantee a net gain.

More adaptive fusion could condition weights on query type, OCR coverage, or retrieval confidence. Such methods may be promising, but they would require a new development and evaluation protocol rather than additional tuning on the current final-test results.

5.3 Retrieval improvements propagate, but generation remains limiting

The paired generation experiment shows that retrieval quality propagates downstream. ColQwen pages improved ANLS by 0.087 over BM25 pages and produced a large gain in numeric relaxed accuracy. The conditional sample reinforces this pattern: when only one modality retrieved the gold page, answers generated from that modality's page were substantially better.

Retrieval, however, is only part of the remaining error. Oracle ANLS reached 0.395 and oracle exact match 0.336, with fewer than half of the oracle answers crossing the ANLS threshold. The manual audit indicates that these residual errors include normalization mismatches, numeric and unit handling, missed spans, and visual interpretation failures. Some low scores therefore reflect evaluation or answer-form issues rather than reasoning failure alone, but the oracle gap still shows that correct page retrieval is not sufficient for reliable answering under the evaluated setup.

5.4 What the manual audit adds

Automatic answer metrics can make formatting mismatches and substantive failures look similar. The audit suggests several distinct improvement targets: normalization of equivalent answer forms, better abstention when evidence is absent, stronger numeric and unit handling, and more reliable localization in tables and charts. Because the audit is source-balanced and computer-assisted, it is used to identify mechanisms rather than estimate their population prevalence.

5.5 Implications for evaluation

The results support three evaluation practices. First, evidence-page retrieval should be reported separately from final answer quality. Second, an oracle-page condition is valuable for estimating how much error remains after the annotated evidence is available. Third, balanced diagnostic samples should be clearly distinguished from random or population-weighted samples. These controls reduce the risk of attributing gains to the wrong pipeline stage or interpreting diagnostic proportions as population frequencies.

6. Limitations

The study has several limitations. The primary evaluation uses a 5,000-question subset of the DocVQA validation set, while the MP-DocVQA evaluation is restricted to pages within each question's document and therefore reflects page localization rather than unrestricted retrieval. The diagnostic audits were computer-assisted and were intended to identify failure mechanisms rather than estimate population-level error prevalence.

The experiments evaluate one visual retriever family and one answer-generation model. Qwen2.5-VL was used with 4-bit NF4 quantization because full-precision inference exceeded the available 8 GB GPU memory, and the effect of quantization relative to full precision was not measured. Answer generation was also evaluated on a random sample of 1,000 of the 3,997 final-test questions rather than the complete set.

Finally, gold-page identity is an imperfect proxy for sufficient evidence because some non-gold pages may contain enough information to answer a question, while some gold pages may require additional context. The findings should therefore be interpreted within the evaluated datasets, models, and experimental conditions.

7. Conclusion

This study treats document visual question answering as two connected but separable problems: selecting the page that contains relevant evidence and generating a supported answer from that page. On held-out DocVQA and document-restricted MP-DocVQA, ColQwen retrieved the gold page more reliably than BM25 and BGE-M3, and the stronger retrieval translated into better downstream answers. Fixed RRF did not improve on ColQwen, despite complementary text and visual errors. Oracle evidence further showed that substantial answer error remained even when the annotated page was available. Reliable document QA therefore depends on both evidence selection and answer generation, and multimodal fusion should be evaluated rather than assumed to improve the strongest component.

Statements and Declarations

Funding: The authors declare that this research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests: The authors declare that they have no competing financial or non-financial interests that could have influenced the work reported in this paper.

Data and code availability

No new dataset was created for this study. The publicly available DocVQA and MP-DocVQA benchmarks were used under their respective terms. The code, configurations, aggregate results, and reproducibility records are available at https://github.com/Richmiz/docvqa-evidence-retrieval.

Author Contributions

R.A-M. conceived the study, developed the methodology, implemented the experimental framework, conducted the experiments and data analysis, interpreted the findings, and prepared the original manuscript. M.M provided supervision and research guidance, contributed to methodological refinement and validation, and reviewed and edited the manuscript. M.S. contributed to supervision, interpretation and validation of the results, project administration, and manuscript review and editing. All authors read and approved the final manuscript.

References

[1] M. Faysse et al., ColPali: Efficient Document Retrieval with Vision Language Models, arXiv:2407.01449, 2024. https://doi.org/10.48550/arXiv.2407.01449

[2] J. Cho, D. Mahata, O. Irsoy, Y. He, and M. Bansal, M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding, arXiv:2411.04952, 2024. https://doi.org/10.48550/arXiv.2411.04952

[3] K. Dong, Y. Chang, D. Goh Xin Deik, D. Li, R. Tang, and Y. Liu, MMDocIR: Benchmarking Multimodal Retrieval for Long Documents, Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 30971–31005, 2025. https://doi.org/10.18653/v1/2025.emnlp-main.1576

[4] X. Peng, C. Qin, Z. Chen, R. Xu, C. Xiong, and C.-S. Wu, UniDoc-Bench: A Unified Benchmark for Document-Centric Multimodal RAG, arXiv:2510.03663, 2025. https://doi.org/10.48550/arXiv.2510.03663

[5] Z. Li, Y. Li, L. Kang, D. Karatzas, and W. Ma, AVIR: Adaptive Visual In-Document Retrieval for Efficient Multi-Page Document Question Answering, Proceedings of the 7th ACM International Conference on Multimedia in Asia, 2025. https://doi.org/10.1145/3743093.3771020

[6] J. Chen, M. Li, J. Kil, C. Wang, T. Yu, R. A. Rossi, T. Zhou, C. Chen, and R. Zhang, VisR-Bench: An Empirical Study on Visual Retrieval-Augmented Generation for Multilingual Long Document Understanding, arXiv:2508.07493, 2025. https://doi.org/10.48550/arXiv.2508.07493

[7] E. Lopez, A. Llabres, and E. Valveny, Enhancing Document VQA Models via Retrieval-Augmented Generation, ICDAR Workshop on Machine Learning in Document Analysis and Recognition, 2025. https://doi.org/10.48550/arXiv.2508.18984

[8] A. Most et al., Lost in OCR Translation? Vision-Based Approaches to Robust Document Retrieval, arXiv:2505.05666, 2025. https://doi.org/10.48550/arXiv.2505.05666

[9] L. Sun et al., When Good OCR Is Not Enough: Benchmarking OCR Robustness for Retrieval-Augmented Generation, arXiv:2605.00911, 2026. https://doi.org/10.48550/arXiv.2605.00911

[10] C. Shorten et al., IRPAPERS: A Visual Document Benchmark for Scientific Retrieval and Question Answering, arXiv:2602.17687, 2026. https://doi.org/10.48550/arXiv.2602.17687

[11] M. Mathew, D. Karatzas, and C. V. Jawahar, DocVQA: A Dataset for VQA on Document Images, Proceedings of WACV, 2021. https://doi.org/10.1109/WACV48630.2021.00225

[12] R. Tito, D. Karatzas, and E. Valveny, Document Collection Visual Question Answering, arXiv:2104.14336, 2021. https://doi.org/10.48550/arXiv.2104.14336

[13] R. Tito, D. Karatzas, and E. Valveny, Hierarchical Multimodal Transformers for Multi-Page DocVQA, arXiv:2212.05935, 2022. https://doi.org/10.48550/arXiv.2212.05935

[14] Y. Xu et al., LayoutLM: Pre-training of Text and Layout for Document Image Understanding, Proceedings of KDD, 2020. https://doi.org/10.1145/3394486.3403172

[15] Y. Xu et al., LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding, Proceedings of ACL-IJCNLP, 2021. https://doi.org/10.18653/v1/2021.acl-long.201

[16] Y. Huang et al., LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking, Proceedings of ACM Multimedia, 2022. https://doi.org/10.1145/3503161.3548112

[17] G. Kim et al., OCR-free Document Understanding Transformer, Proceedings of ECCV, 2022. https://doi.org/10.1007/978-3-031-19815-1_30

[18] K. Lee et al., Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding, Proceedings of ICML, 2023. https://proceedings.mlr.press/v202/lee23g.html

[19] S. Robertson and H. Zaragoza, The Probabilistic Relevance Framework: BM25 and Beyond, Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333-389, 2009. https://doi.org/10.1561/1500000019

[20] N. Reimers and I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Proceedings of EMNLP-IJCNLP, 2019. https://doi.org/10.18653/v1/D19-1410

[21] W. Wang et al., MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers, Advances in Neural Information Processing Systems, vol. 33, 2020. https://arxiv.org/abs/2002.10957

[22] J. Chen et al., BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, arXiv:2402.03216, 2024. https://doi.org/10.48550/arXiv.2402.03216

[23] O. Khattab and M. Zaharia, ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, Proceedings of SIGIR, 2020. https://doi.org/10.1145/3397271.3401075

[24] K. Dong et al., Benchmarking Retrieval-Augmented Multimodal Generation for Document Question Answering, arXiv:2505.16470, 2025. https://doi.org/10.48550/arXiv.2505.16470

[25] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods, Proceedings of SIGIR, pp. 758-759, 2009. https://doi.org/10.1145/1571941.1572114

[26] S. Bai et al., Qwen2.5-VL Technical Report, arXiv:2502.13923, 2025. https://doi.org/10.48550/arXiv.2502.13923

[27] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, Findings of ACL, 2022. https://doi.org/10.18653/v1/2022.findings-acl.177

[28] F. Zhu et al., Towards Complex Document Understanding by Discrete Reasoning, arXiv:2207.11871, 2022. https://doi.org/10.48550/arXiv.2207.11871