

Article

DRAG: Dual-Channel Retrieval-Augmented Generation for Hybrid-Modal Document Understanding

Zhe Xin [†] , Shuyuan Xia [†] and Xin Guo ^{*} 

School of Artificial Intelligence and Data Science, Hebei University of Technology, Tianjin 300401, China; xzzhuanshu96@163.com (Z.X.); xiashuyuan@hebut.edu.cn (S.X.)

^{*} Correspondence: gxhebut@aliyun.com

[†] These authors contributed equally to this work.

Abstract

Large Language Models (LLMs) have acquired vast amounts of knowledge during pre-training. However, there are a lot of challenges when it is deployed in real-world applications, such as poor interpretability, hallucinations, and the inability to reference private data. To address these issues, Retrieval-Augmented Generation (RAG) has been proposed. Traditional RAG relying on text-based retrievers often converts documents using Optical Character Recognition (OCR) before retrieval. While testing has revealed that it tends to overlook tables and images contained within the documents. RAG, relying on vision-based retrievers, often loses information on text-dense pages. To address these limitations, we propose DRAG: Dual-channel Retrieval-Augmented Generation for Hybrid-Modal Document Understanding, a novel retrieval paradigm. The DRAG method proposed in this paper primarily comprises two core improvements: first, a parallel dual-channel processing architecture is adopted to separately extract and preserve the visual structural information and deep semantic information of documents, thereby effectively enhancing information integrity; second, a novel dynamic weighted fusion mechanism is proposed to integrate the retrieval results from both channels, enabling precise screening of the most relevant information segments. Empirical results demonstrate that our method achieves Competitive performance across multiple general benchmarks. Furthermore, performance on biomedical datasets (e.g., BioM) specifically highlights its potential in specialized, vertical domains such as elderly care and rehabilitation, where documents are characterized by dense hybrid-modal information.

Keywords: retrieval-augmented generation; large language models; multi-document retrieval; visual question answering; elderly care



Academic Editor: Arkaitz Zubiaga

Received: 19 January 2026

Revised: 12 February 2026

Accepted: 14 February 2026

Published: 16 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The global population aging process is accelerating, a trend that is altering the balance between the demographic groups requiring care and the working-age population capable of providing it. The World Health Organization (WHO) defines intrinsic capacity (IC) as the composite of all an individual's physical and mental (including psychological) capacities. In its report on aging and health, WHO emphasizes the importance of maintaining the intrinsic capacity of older adults. Approaches such as external care and exercise can promote the functioning of their organism, thereby reducing the occurrence of adverse events like falls and sudden death. Against this backdrop, providing scientifically effective

guidance for care and rehabilitation interactions for older adults and caregiver groups is both urgent and necessary.

With the advancement of large language models, they leverage their massive parameter scale and vast training datasets, and have demonstrated remarkable capabilities across numerous natural language processing (NLP) tasks [1,2] such as natural language understanding and text generation. LLMs have driven breakthroughs in intelligent information processing, with applications across domains highlighting technical and practical value. However, given that large language models are inherently parameterized, their knowledge is constrained by static training corpora. This leads to challenges in handling knowledge-intensive queries in care and rehabilitation, such as generating hallucinations [3,4], knowledge obsolescence, and a lack of verifiable sources.

While Retrieval-Augmented Generation (RAG) methods [5,6] can retrieve knowledge from external corpora, potentially reducing hallucinations and improving knowledge coverage, early RAG systems were text-centric [7–9]. They relied on traditional retrieval algorithms like BM25 or neural models such as DPR, processing only plain text data. Although effective for semantic similarity search, these systems struggled to capture relational structures. With the maturation of Optical Character Recognition (OCR), multimodal RAG enables retrieval from documents with diverse information, but the OCR process often discards crucial visual elements like layout and chart semantics [10,11]. In vertical domains represented by elderly care and rehabilitation, knowledge possesses an inherent structure—different etiologies correspond to different symptoms, while reports and documentation contain multimodal information. This is reflected in document processing: pages comprise text, graphics, charts, etc., and in content, tabular data quantifies textual descriptions, while images visually concretize written information. These diverse information modalities are interrelated and complementary, each being an indispensable component.

Recent years have witnessed the rapid rise of Vision-Language Models, which have brought new technological breakthroughs for Multimodal Retrieval-Augmented Generation [12–14]. These models can directly process document images and capture multi-dimensional information—including text, layout, and charts—through visual embeddings. However, their comprehension of textual semantics remains relatively weak, a limitation that proves particularly constraining in the context of care and rehabilitation, where rigorous information is paramount. An alternative emerging approach employs large vision-language models to extract document information, leveraging their detailed perceptual capabilities to generate textual summaries of non-textual elements like charts while also recognizing and extracting document text. Nevertheless, the inherent tendency of these models to hallucinate can introduce significant noise into the retrieval results [15–17]. Consequently, relying solely on either a visual or a textual channel is insufficient for achieving precise retrieval in scenarios dealing with complex, hybrid-modal documents. While our research is motivated by the urgent needs of the elderly care and rehabilitation domain—where accurate interpretation of illustrated manuals is critical—the proposed solution addresses a fundamental limitation in current RAG systems applicable to any text-graphic dense corpora.

To address this challenge, this paper proposes DRAG: Dual-channel Retrieval-Augmented Generation for Hybrid-Modal Document Understanding that integrates both visual and textual channels. Channel 1 is built on a visual encoder, which directly extracts features from the original images to avoid information loss. It excels particularly in understanding structured visual content such as tables and charts. By introducing indexing and feature fusion mechanisms during the retrieval process, this channel significantly enhances retrieval speed and accuracy, ultimately returning the image segments with the highest visual similarity. Channel 2, on the other hand, is based on a vision-language model. It

first converts the entire document into a textual representation and then employs semantic retrieval to match and return the text paragraphs most relevant to the query. Finally, a novel dynamically weighted fusion mechanism is proposed to balance the retrieval results from the dual channels. The primary contributions of this study are as follows: (1) A dual-channel RAG architecture based on visual semantic retrieval and text vector retrieval is integrated into a coherent pipeline. (2) The adoption of a dynamically weighted hybrid evidence fusion mechanism to balance results from textual and visual retrieval. Empirical evaluation demonstrates significant improvements in accuracy, semantic consistency, and latency.

2. Materials and Methods

Text retrieval-augmented techniques utilize retrieval strategies to extract external knowledge [5], dynamically supplementing the high-quality and precise information necessary during the generation of model responses, thus mitigating model hallucinations [18]. Early Retrieval-Augmented Generation (RAG) systems predominantly relied on sparse retrieval methods [19], such as BM25 [20], which identified relevant information through keyword matching and statistical metrics [21–23]. Although computationally efficient, these approaches operated solely at the lexical level and were unable to capture the deeper semantic relationships in text. Recent RAG research has concentrated on the following aspects: (1) Dense retrieval [24,25], which encodes queries and documents into dense vectors through deep neural networks [26], followed by high-dimensional semantic matching using similarity algorithms; (2) Generative retrieval [27–29], in which sequence-to-sequence models directly generate document identifiers relevant to queries [30,31], and contextual augmentation strategies [32–35]. However, current research on text retrieval augmentation remains narrowly focused on performance improvements for specific tasks within fixed datasets. Real-world documents predominantly consist of multimodal content interwoven with text-centric tasks, necessitating the development of novel and effective multimodal RAG pipelines.

The fundamental breakthrough of Vision-Language Models lies in establishing semantic alignment between visual and textual modalities, enabling unified comprehension of multimodal information. Since CLIP pioneered cross-modal mapping between text and images through contrastive learning, VLMs have rapidly evolved from basic alignment to complex task adaptation. Subsequent models like BLIP and FLAVA enhanced generalization capabilities for tasks such as image captioning and visual question answering via multi-task pretraining. Models such as LLaVA and Qwen2-VL further integrated visual encoders with large language models, empowering them to directly process visual inputs and generate natural language responses. This integration allows capturing semantic information, including textual content, layout structures, and graphical elements from high-resolution images. The VISPLAY model introduces a self-evolution reinforcement learning framework, enabling VLMs to autonomously enhance reasoning capabilities by leveraging large-scale unlabeled image data [36]. The advancement of vision models serves as the fundamental cornerstone for multimodal technologies.

Multimodal retrieval systems are designed to process diverse data modalities—including text, images, and videos—within a unified framework. Representative works such as UniIR, MuRAG, and M3DocRAG achieve multimodal retrieval through shared embedding spaces or cross-modal attention mechanisms [37,38]. In document retrieval, approaches like DSE and ColPali propose directly processing document images as input and leveraging VLMs to generate embeddings, significantly streamlining retrieval pipelines while improving visual information retention. Additionally, SERVAL introduces a zero-shot visual document retrieval method that employs VLMs to generate document descriptions followed

by text-encoder-based retrieval, achieving state-of-the-art performance on multiple benchmarks. Despite these notable achievements in multimodal document retrieval, several limitations persist: while ColPali excels in visual retrieval, its multi-vector embedding mechanism incurs substantial storage overhead and underutilizes textual information; though supporting multi-page and multi-document retrieval, M3DocRAG primarily relies on visual embeddings during retrieval, lacking deep integration of textual semantics; SERVAL achieves zero-shot retrieval through generated descriptions, yet its performance depends heavily on the quality of VLM-generated captions and exhibits instability in semantically complex tasks. Consequently, there remains a clear absence of multimodal RAG systems that effectively balance dual visual-textual information integration while maintaining equilibrium between storage efficiency and semantic understanding. Recent studies also highlight the importance of disciplined architectural choices in RAG systems to ensure reliability [39].

3. Methods

We propose a Dual-channel Retrieval-Augmented Generation for a Hybrid-Modal Document Understanding framework that integrates visual and textual channels. This architecture can flexibly process diverse information within documents—such as text and charts—through its two parallel channels and adjusts the ranking of retrieval results based on the proportional weight of each information modality. Figure 1 illustrates the workflow of our method, which consists of two main phases: (1) offline dual-channel document embedding (Section 3.2); (2) LckBorda merging and generation of two-channel retrieval (Section 3.3).

3.1. Task Definition

A query is defined as Q , and we consider only textual question inputs analogous to traditional search settings. The corpus is defined as $C = \{D_1, D_2, \dots, D_m\}$, where m denotes the total number of documents. Each document D_i consists of a set of pages P_j . Each P_j represents an image rendered from a page screenshot, thus the entire corpus can be represented as $C = \{P_1, P_2, \dots, P_j, \dots, P_n\}$, where n is the total number of page images across all documents in the corpus. The document retrieval task aims to identify the top- K most relevant page images for query Q , denoted as P_K^Q . This result set P_K^Q comprises two components: the output from the visual retrieval channel, P_K^{QV} , and the output from the textual retrieval channel, P_K^{QT} . To enable efficient search, all units in corpus C are pre-encoded into embeddings offline. Finally, the combined page information from retrieval, along with the original query Q , is fed into a question-answering model to generate the final answer.

3.2. Dual-Channel Document Embedding

Within the vision-centric embedding retrieval channel, we employed an improved version of the ColPali model in our experiments. This enhanced model retains the core capability of understanding document visual layouts and demonstrates strong performance in processing tables and charts. However, in Figure 2, we also observed that for images containing a high density of text—specifically, in text-intensive documents—it still struggles to capture fine-grained textual information. Furthermore, in purely text-dense retrieval tasks, its depth of semantic understanding does not match that of dedicated traditional text retrieval models.

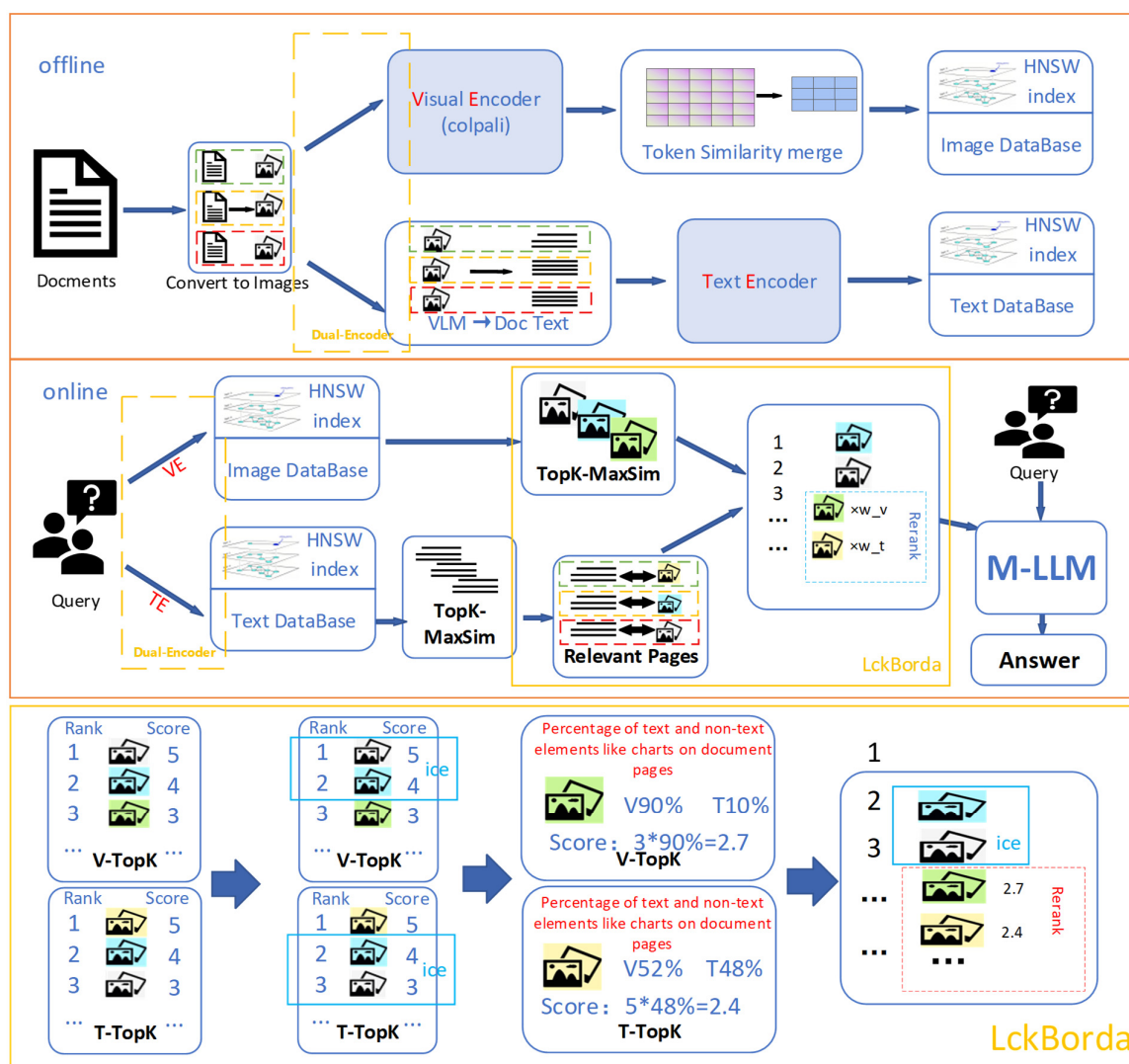


Figure 1. DRAG (Dual-Encoder Multimodal Document Retrieval). The dual encoding channel (yellow dashed box) operates during the offline processing phase, where page transitions are encoded and embedded through both text and visual channels. In the online phase, queries are retrieved from two separate databases. The LckBorda merged retrieval (yellow solid box) combines the weight distribution of different modalities in the page with Borda scores to reorder the retrieval results.

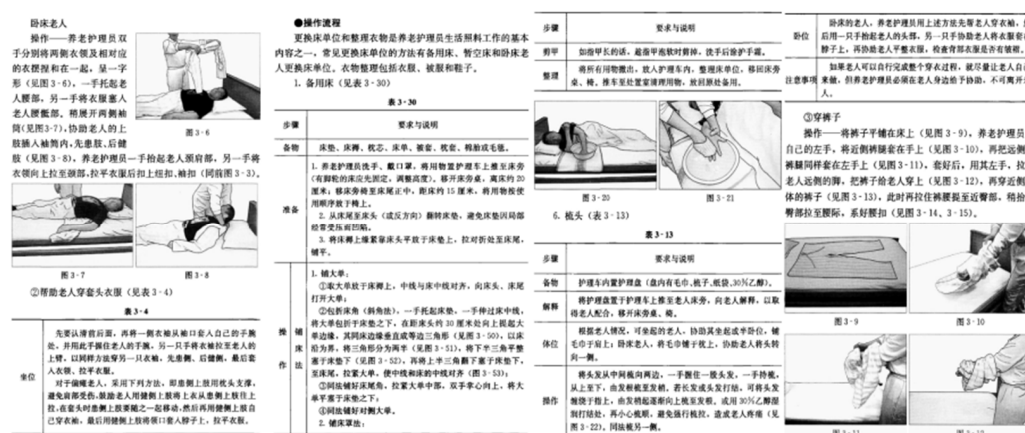


Figure 2. Document example in care and rehabilitation. The Chinese text in the figure explains the steps and key knowledge points related to elderly care, while the small illustration beside it demonstrates specific nursing actions.

Visual model embedding: First, page images P are encoded into a unified multimodal embedding space using ColPali. While ColPali draws on the training methodology of the ColBERT model and is capable of encoding both textual and visual inputs, it generates patch-level embeddings for each page, resulting in excessive memory usage. The research by Ma et al. [40], reduces embeddings per page with minimal performance degradation, identifying semantic clustering as the most effective method.

Within the document setup, each page P is formatted as an image with fixed pixel dimensions (both width and height). For a given page image, ColPali extracts a dense visual embedding $E_V^P \in \mathbb{R}^{n_P \times d}$, where n_P denotes the fixed number of tokens per page and d represents the embedding dimension (reduced to 128, consistent with ColPali's configuration). Similarly, the query Q is embedded as $E^Q \in \mathbb{R}^{n_Q \times d}$, where n_Q is the number of text tokens.

Merging token: To reduce the memory footprint of multi-vector representations, we employ hierarchical clustering to merge visually and semantically redundant patches. Following the empirical findings of 'Towards Storage-Efficient Visual Document Retrieval', this process focuses on maximizing semantic information density. Although explicit grid coordinates are relaxed, the method retains the essential feature activations required for high-recall retrieval, efficiently filtering out background noise. For a given page E_V^P , the initial embedding set $S^{(0)} = \{\{e_1\}, \{e_2\}, \dots, \{e_{n_P}\}\}$ is iteratively merged to produce the final compressed semantic clusters $S = \{S_1, S_2, \dots, S_k\}$, where k is the target number (set to $n_P/25$ to balance compression ratio and performance preservation). The iterative procedure is as follows: Two clusters $C_a = \{e_{a_1}, e_{a_2}, \dots, e_{a_m}\}$ (containing m number of embeddings) and $C_b = \{e_{b_1}, e_{b_2}, \dots, e_{b_n}\}$ (containing n number of embeddings) are defined. Their inter-cluster similarity is calculated as: $\text{sim}(C_a, C_b) = \frac{1}{m \cdot n} \sum_{i=1}^m \sum_{j=1}^n \text{sim}(e_{a_i}, e_{b_j})$. The pair of clusters with the highest similarity, $C_{\max-a}$ and $C_{\max-b}$, are merged into a new cluster $C_{\text{new}} = C_{\max-a} \cup C_{\max-b}$. The compressed visual embeddings are denoted as $E_{VK}^P \in \mathbb{R}^{k \times d}$. After processing each page in the corpus as described above, the final high-dimensional embedding set for the corpus is obtained: $D_V = \{E_{VK1}^P, E_{VK2}^P, \dots, E_{VKn}^P\}$.

In constructing the text-dominant retrieval channel, we aim to complement the shortcomings of the visual channel. While traditional text retrieval models offer significant compensatory benefits, our analysis of the care and rehabilitation domain revealed a prevalence of documents where textual and graphical information are interwoven—such as tables containing extensive symptom descriptions. Consequently, we opted to implement a text-prompted description generation method based on a large vision-language model. This approach avoids the direct discarding of structural and non-textual information inherent in conventional text-only retrieval pipelines.

Text model embedding: Unlike traditional OCR methods that often result in fragmented text without structural context, we leverage Qwen2.5-VL to generate holistic semantic descriptions of page content. This approach aligns with the SERVALL framework, where explicit predictive text generation serves to bridge the semantic gap that purely visual embeddings (like ColPali) may overlook. In the textual retrieval channel, for the corpus $C = \{P_1, P_2, \dots, P_n\}$, we employ a vision-language model to generate detailed descriptions of each page image. As illustrated in Figure 3, the model is instructed to perform comprehensive textual extraction of the image content, including descriptions of figures, charts, and tables, followed by summarization and categorization. For graphical elements such as charts and tables, descriptions are generated based on their semantic characteristics and distinctive features. We acknowledge that VLM-generated descriptions may contain hallucinations. However, in our dual-channel design, this risk is mitigated by the parallel visual channel. The visual embeddings operate on raw image patches, preserving exact structural fidelity. During the fusion phase (Section 3.3), a page is priori-

tized only if it receives high relevance scores from both the semantic understanding (text channel) and visual similarity (visual channel), effectively filtering out noise introduced by potential generative hallucinations. Subsequently, both the query and the generated document descriptions are encoded using the same encoder—any off-the-shelf text encoder can be selected for this step. This shared encoder projects queries and documents into a unified semantic space, ensuring closely aligned representations of relevant information and eliminating semantic deviation between queries and documents.

Process the document within the image by performing summarization and description. Firstly, classify the document based on the relevance of its included charts and textual content, identifying key features and providing a summary. Subsequently, offer a detailed explanation for each chart, analyzing its underlying significance. Finally, extract all textual and numerical information from the document, ensuring that data from charts is clearly distinguished from the textual paragraphs.

Figure 3. Prompt for text description of VLM.

Unlike conventional approaches, the textual pipeline still treats each page as a fundamental unit. All descriptions for a given page image are encoded into a common vector space, yielding a textual embedding $E_T^P \in \mathbb{R}^{n_T \times d}$, where n_T denotes the fixed number of tokens per page after encoding, and d represents the embedding dimension (set to 128 in our experiments, using ColPali as the embedding model). The textual corpus embedding set is thus obtained: $D_T = \{E_{T1}^P, E_{T2}^P, \dots, E_{Tn}^P\}$.

3.3. Retrieval, LckBorda, and Generation

HNSW-based Index Retrieval: During the inference phase, to ensure low-latency retrieval across the large-scale document corpus, we employ the Hierarchical Navigable Small World (HNSW) algorithm for Approximate Nearest Neighbor (ANN) search, defined as $\mathcal{R}: (Q, D_V) \rightarrow C_V$. Consistent with the architecture shown in Figure 1, HNSW constructs a multi-layered graph structure. The lower layers capture fine-grained connections between data points (compressed page embeddings), while the upper layers serve as coarse navigation bridges, allowing for logarithmic time complexity during the search process.

Specifically, we build independent HNSW indices for both channels. For the visual channel, the compressed patch embeddings (derived in Section 3.2) are indexed to facilitate rapid similarity matching despite the high dimensionality. During retrieval, the query embedding E_Q enters the graph from the top layer and greedily traverses to the neighbor closest to the query, moving downwards until converging at a local minimum in the bottom layer. This mechanism allows us to efficiently identify the optimal Top-K candidates C_V with high recall, significantly outperforming linear scan methods in terms of query latency. Both inter-page similarity and retrieval similarity are computed using the same maximum similarity summation method. Taking the calculation between two pages P_1 and P_2 as an example:

$$\text{sim}(P_1, P_2) = \sum_{i=1}^{n_p} \max_{j \in n_p} \langle E_{VK1}^P(i) * E_{VK2}^P(j) \rangle$$

where $*$ denotes the dot product, and $E_{VK1}^P(i)$ represents the i -th row vector of the compressed embedding matrix.

Text channel retrieval is performed based on the query embedding encoded by the same model, defined as $\mathcal{R} : (Q, D_T) \rightarrow C_T$, where C_T denotes the set of candidate embeddings retrieved via the textual channel, keep only the TopK results. The similarity computation during retrieval follows the same method as used in the visual pipeline.

LckBorda: After obtaining retrieval results from both channels, we design an enhanced cross-modal fusion module named **LckBorda** (Layout-aware, content-weighted Borda count) to enable synergistic reasoning over visual and textual evidence by merging the two candidate embedding sets. For the textual channel results, a pre-processing step is performed to trace the candidate set C_T back to their original page images using page number labels, yielding C_{TP} . Unlike the standard Borda count, our method first computes the intersection between the two result sets, C_V and C_{TP} . Pages present in both sets are directly added to the final fused set C_Q .

For the remaining non-overlapping results, ranks are assigned descending scores (e.g., 10 points for rank 1, 9 for rank 2, etc.). A key innovation is the introduction of content-aware weighting coefficients: for visual channel results, the weight of a page (w_{vis}) is quantitatively defined using a layout analysis module. specifically, $w_{vis} = \frac{\sum Area_{charts} + \sum Area_{images}}{Area_{total_page}}$ and the textual weight is $w_{text} = 1 - w_{vis}$. This heuristic ensures that visual retrieval scores contribute more significantly when the page is dominated by charts or figures, while textual scores dominate for text-heavy pages. For the textual channel, the weight is the proportion of textual content. Each non-overlapping page's score is then multiplied by its corresponding weight. These weighted scores are pooled and sorted across both channels. Finally, the $Top(K - a)$ items from this sorted list are appended to C_Q , where a is the number of overlapping pages already in C_Q . The text channel page in C_Q retains the corresponding page description content.

The fused result C_Q , along with the original query Q , is then input into a multi-modal language model capable of processing multiple images to generate the final answer autoregressively:

$$a \leftarrow VLM - Multi(Q, C_Q)$$

4. Experiment Setup

To evaluate our framework, we employ four distinct benchmark methodologies. First, we conduct experimental assessments on two multi-page document visual question answering (DVQA) datasets: MPDocVQA and DUDE [41,42]. Both datasets constrain answers to specific page ranges, making it critical to precisely identify the page(s) containing the answer before attempting generation. In the MPDocVQA dataset, all answers are extractive, while the DUDE dataset includes four answer types: extractive, abstractive, list-based, and unanswerable. Furthermore, MPDocVQA provides both the answer and its corresponding document page, whereas DUDE only supplies the answer, which significantly increases the difficulty of locating answers within large-scale document image collections.

Given that the two evaluation methods above focus on fixed-range visual question answering, we supplement with the following experiment. This study adopts the Visual Document Retrieval benchmark ViDoRe, which consists of diverse page-level retrieval tasks spanning multiple domains, languages, and practical scenarios. Compared to ViDoRe-v1, the ViDoRe-v2 benchmark presents a more challenging test set [43]. ViDoRe-v2 employs a hybrid synthetic query generation process involving context-free blind queries, long-text cross-document queries, and human intervention to construct retrieval tasks that better simulate real-world, high-difficulty scenarios. It comprises four distinct multilingual datasets and provides clear evaluation guidelines. Furthermore, as shown in Figure 4, the selection of the ViDoRe-v2 benchmark is based on practical considerations relevant to our application, the MIT Biomedical Multilingual (BioM) dataset, selected as a proxy for the

elderly care domain. While distinct in specific content, BioM shares critical typological challenges with elderly care rehabilitation manuals, specifically: (1) the dense integration of professional medical terminology, and (2) the heavy reliance on illustrative diagrams (e.g., molecular structures or anatomical parts) alongside explanatory text. This makes it an ideal public benchmark to test the model's capability in handling the 'text-graphic' dependencies typical of care documentation.

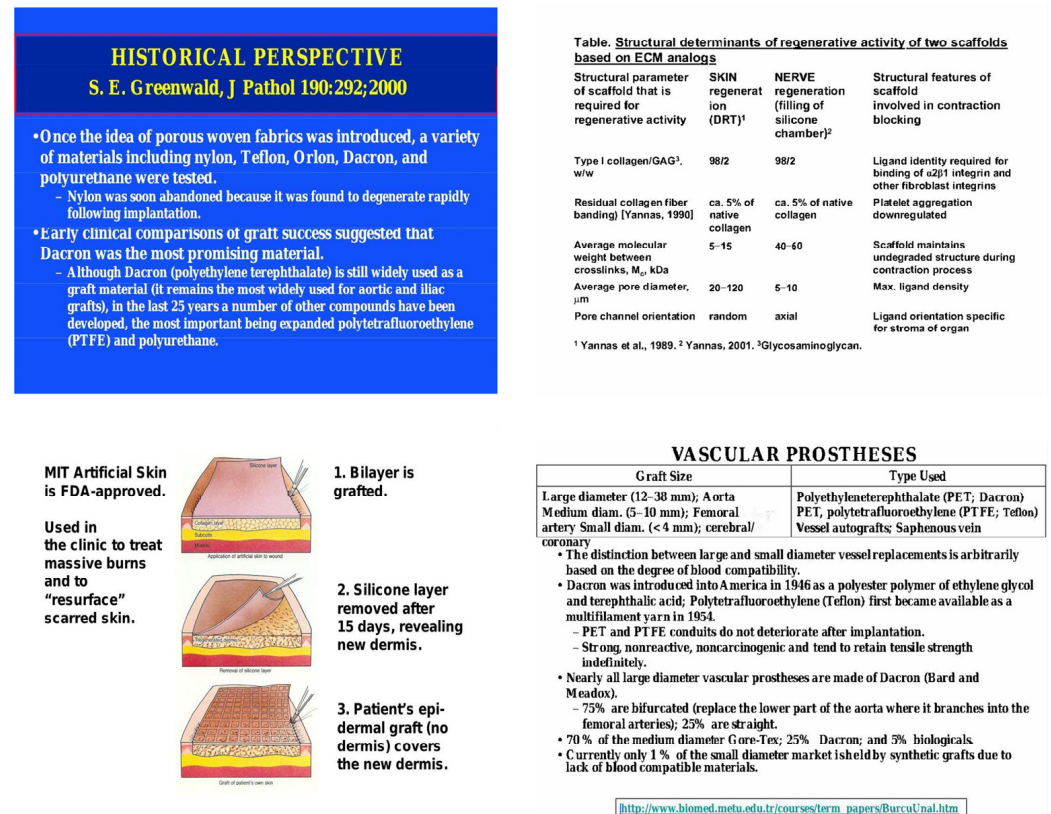


Figure 4. BioM Dataset Samples. The dataset is derived from Massachusetts Institute of Technology anatomy course materials, primarily consisting of biomedical content with relevance to elderly care and nursing. Additionally, its structure predominantly features a tightly interwoven format of integrated text and graphics.

Furthermore, this study extends performance evaluation to the MMLongBench-DOC benchmark [44]. This dataset is designed to assess the long-context comprehension capabilities of large vision-language models (LVLMs). Unlike prior benchmarks that predominantly focused on single-page or short-text scenarios with minimal cross-page questions (e.g., DUDE's 2.1% rate) and insufficient information density, results obtained on MMLongBench-DOC more effectively demonstrate the superiority of the tree-indexed retrieval mechanism in our proposed framework. The dataset is characterized by the following features: (1) **Prominent Long-context Attributes:** The 135 documents average 47.5 pages and 21,214.1 tokens per document, substantially exceeding the scale of existing datasets. (2) **Multimodal Coverage:** Answering questions requires extracting evidence from five distinct sources: text, layout, charts, tables, and images. (3) **Comprehensive Question Typology:** It includes 33.7% cross-page questions requiring multi-page reasoning and 20.6% unanswerable questions for hallucination detection, with high annotation quality ensured through expert labeling and a three-round review process.

To investigate the fusion of multimodal results from the dual channels, we conduct an ablation study on the visual document retrieval benchmark ViDoRe, comparing different fusion strategies. In the experiment, the visual channel employs ColPali as the retrieval

model, while the textual channel utilizes the Qwen model. Four distinct fusion strategies are evaluated: **Simple Merge:** Results from both channels are merged and deduplicated. If the merged set exceeds the preset K value, pages with the lowest ranks from each channel are iteratively removed. **Averaged Weighting:** Based on the Borda count method, each channel assigns a score of $K+1-\text{rank}$ to its results. Each score is then multiplied by a weight of 0.5. If a page appears in both channels (i.e., in the intersection), its final score is the weighted sum from both channels. All pages are then sorted by this composite score, and the top K results are selected. **Vision-Dominant:** Similar to Method 2, but assigns a weight coefficient of 1.5 to the visual channel and 0.5 to the textual channel. **Text-Dominant:** Similar to Method 2, but assigns a weight coefficient of 1.5 to the textual channel and 0.5 to the visual channel.

Experiments were conducted on a cloud server equipped with an Intel Xeon(R) Platinum 8458P CPU, dual NVIDIA H800 80GB GPUs, and 100 GB of system RAM, running Ubuntu 20.04 and Python 3.10. The implementation leveraged PyTorch 2.2.2, the Transformers library, and FlashAttention-2 for model execution. PDF pages were converted to RGB images using the pdf2image library at a resolution of 144 DPI. To ensure consistency and fairness across experiments, page sizes from the ViDoRe-v2, MPDocVQA, DUDE, and MMLongBench-DOC datasets were uniformly resized to 1700 (width) $\times$ 2200 (height) pixels, despite slight variations in their original dimensions. In the experiment, the textual retrieval channel employs Qwen2.5VL-7B as the vision-language model, specifically utilizing a quantized version with Activation-aware Weight Quantization. The visual channel employs the ColPali v3 retrieval model. In the comparative analysis of experimental results across different datasets, variations exist in the baseline models selected for each benchmark. Specifically, for both visual and textual channels, we retrieve an initial candidate set of $K' = 10$ pages. After LckBorda fusion, the final Top $K = 5$ pages are provided to the generator. The page description prompt used in the textual channel is: “Provide a comprehensive description of the document in the image in English. Structure your response into the following three sections: Summary: Provide a high-level overview of the document’s content, identifying its type (e.g., slide, report, chart) and main topic. Details: Describe the visual elements in detail. For any charts, graphs, or tables, explain the visual components (such as x/y axes, legends, and trends) and the relationships they depict. Extracted Data: Extract all visible text and numerical values from the document. List specific data points from charts separately from the main textual paragraphs to ensure clarity. The generation process uses a temperature of 0.2 and a maximum output limit of 512 tokens. Baseline results are sourced from official leaderboards as of [December 2025]. We acknowledge that minor version discrepancies may exist; our evaluation follows an apples-to-apples protocol for all retrieved evidence provided to the same generator backbone. Besides, our method represents a ‘system-level’ improvement rather than merely a backbone upgrade.” This is primarily attributed to differences in the evaluation focus of each dataset, leading to distinct top-performing models on their respective leaderboards.

5. Results

Table 1 presents a comparison of our architecture against several baseline models on two multi-page datasets. In MPDocVQA, ORA indicates the gold answer page is provided, MCO indicates Model’s Page Filtering and Answer Integration Capabilities, and CON indicates Model’s ability to process ultra-long sequences. Average ANLS results per question type are abbreviated as (Abs)tractive, (Ex)tractive, (N)ot-(A)nswerable, (Li)st.

Table 1. Comparison results of MPDocVQA and DUDE between our current pre-training model and the current pre-training model.

Modal	Modality	MPDocVQA				DUDE			
		ORA	MCO	CON	Abs	Ex	NA	Li	ANLS
BERT	T	59.04	53.47	41.83	7.28	42.23	5.88	11.13	25.48
BigBird	T	64.50	58.84	52.97	7.11	40.26	12.75	8.46	26.27
T5	T	68.14	40.28	50.50	47.62	50.49	63.72	7.56	41.80
Longformer	T	61.77	55.06	52.87	8.55	43.58	10.78	10.62	27.14
Hi-VT5	T+V+L	65.72	62.01	62.01	33.94	17.60	61.76	6.83	23.06
LayoutLMv3	T+V+L	67.29	55.13	45.38	8.10	32.60	8.82	7.82	20.31
(Ours)	T+V	75.28	66.07	37.56	50.39	54.77	58.51	15.47	49.38

In MPDocVQA, models are required to answer questions based on PDF documents containing up to 20 pages. The DUDE QA task evaluates capabilities in layout navigation, multi-hop reasoning, and includes assessment of “unanswerable” (12.74%) and “list-type” (6.62%) answers. As illustrated, models such as BigBird, BERT, and Longformer demonstrate strong performance in pure question-answering ability (ORA) and extractive answer tasks when no page interference is present, highlighting the substantial comprehension capacity of advanced architectures. However, they exhibit insufficient capability in handling long sequences and integrating list-based information. Furthermore, the T5 model shows notable results in abstractive and unanswerable answer types.

In both evaluations, model performances vary partly due to DUDE’s lack of page-specific answer indications, where differing input constructions can lead to suboptimal outcomes. Our method demonstrates robust performance across both benchmarks. The relatively modest results in the CON metric are a deliberate trade-off. By treating the page image as the minimal unit and employing token merging, we prioritize ‘retrieval precision’ and ‘visual structural integrity’ over ‘token-level sequence length’. While this slightly impacts ultra-long context serialization (CON), it significantly boosts the ANLS (accuracy) in visually complex QA tasks, which is the core focus of this study. Future work will explore adaptive token merging to balance CON and latency. Nevertheless, the overall results sufficiently illustrate the effectiveness and capability of our proposed architecture.

Table 2 presents an evaluation on the visual retrieval benchmark ViDoRe V2, primarily analyzed using the nDCG@5 metric. For clarity, the following abbreviations are used in the table: MIT Biomedical Multilingual (BioM), Economics Macro Multilingual (MacM), ESG Restaurant Human English (ResH), and ESG Restaurant Synthetic Multilingual (ResSM). In our experiments, we substituted the visual channel embedding model within our framework with cutting-edge embedding models such as Jina and EvoQwen2.5 for evaluation. As the results demonstrate, our dual-channel approach yields significant performance improvements compared to the VDR retrieval model trained solely on direct visual feature embedding. The state-of-the-art EvoQwen2.5-VL-Retriever-7B-v1 model achieves an average nDCG@5 score of 65 on the ViDoRe V2 benchmark. Notably, our framework elevates this score to 68.7, confirming the effectiveness of our dual-channel method without requiring additional training. Comparative experiments across diverse datasets demonstrate that our method achieves particularly significant results on the BioM dataset, showing notable improvement over the baseline models. These findings affirm the outstanding performance and effectiveness of our framework, which is specifically designed for the elderly care and rehabilitation domain.

Table 2. Evaluation results on ViDoRe V2 compared with colpali-v1.3.

Title 1	Average	BioM	MacM	ResH	ResSM
EvoQwen2.5-VL-v1	65	65.2	58.2	77	59.7
jina-e-v4	57.6	60.9	51.9	65.1	52.5
eager-e-v1	56.4	63.6	51	55.3	55.9
colpali-v1.3	54.2	54.6	48.6	58.5	54.9
Ours+colpali-v1.3	63.2	68.5	57.4	63.9	63.1
+eager-e-v1	64.9	66.8	65.3	67.3	60.3
+jina-e-v4	65.4	67.1	55.8	77.2	61.4
+EvoQwen2.5-VL-v1	68.7	67.1	63.6	81	63

To further investigate the contribution of each channel and the efficacy of the LckBorda fusion mechanism, we conducted an ablation study on the ViDoRe-v2 benchmark using ColPali v1.3 as the visual retriever. As shown in Table 3, the visual-only channel excels in graphic-heavy tasks, while the text-only channel provides essential semantic anchors. Simple rank concatenation (Simple Merge) provides marginal gains. However, our LckBorda mechanism, by dynamically weighting results based on page composition, achieves the highest nDCG@5, demonstrating that the improvement stems from both the dual-channel architecture and the adaptive fusion heuristic.

Table 3. Ablation control experiment with different channel contribution degrees.

nDCG@5 (Average)	ColPali v1.3	eager-e-v1	jina-e-v4	EvoQwen2.5-VL-v1
Visual-only (ColPali v1.3)	54.2	56.4	57.6	65
Text-only (Qwen2.5-VL)	56.1	-	-	-
Simple Merge	60	60.2	62.3	63
Weighted Average ($w = 0.5$)	58.3	60	61.6	63.8
LckBorda (Ours)	63.2	64.9	65.4	68.7

Table 4 presents the evaluation results on the MMLongBench-Doc dataset, where questions and their corresponding answers require processing individual PDFs containing up to 120 pages. This effectively assesses the model’s long-context comprehension capability. The table reports generalized accuracy (ACC) across five evidence source modalities—Text (TXT), Layout (LAY), Charts (CHA), Tables (TAB), and Images (IMG)—and three evidence location types: Single-page (SIN), Multi-page (MUL), and Unanswerable (UNA).

Table 4. Evaluation results on MMLongBench-Doc.

Modal	Evidence Locations			Evidence Modalities					ACC
	SIN	MUL	UNA	TXT	LAY	CHA	TAB	FIG	
ChatGLM-128k	18.8	11.5	18.1	23.4	12.7	9.7	10.2	12.2	16.3
Mistral-Instruct-v0.2	16.9	11.3	24.1	19.9	13.4	10.2	10.1	11.0	16.4
DeepSeek-VL-Chat	5.2	7.0	12.8	7.2	6.5	1.6	5.2	7.6	7.4
Idefics2	7.7	7.2	5.0	9.0	10.6	4.8	4.1	8.7	7.0
Qwen-VL-Chat	5.2	7.1	6.2	5.5	9.0	5.4	2.2	6.9	6.1
M3DOCRAAG	32.4	14.8	5.8	30.0	23.5	18.9	20.1	20.8	21.0
Ours	33.1	18.5	26.3	31.0	25.3	17.8	22.4	21.9	25.4

In Table 4, the first five entries are generative models without an RAG structure. The first two text-based models directly process the raw content of a single PDF document—comprising up to 120 pages in the MMLongBench-Doc dataset—to answer each question.

The subsequent three models are multimodal; however, due to context window constraints, they concatenate page images into five low-resolution inputs. Both M3DOCRA and our proposed DRAG utilize the top-5 most relevant retrieved images as input. The generator receives full-page images (resized to 1700×2200 pixels) instead of cropped segments, thereby preserving the holistic layout information. We adopt Qwen2.5-VL-32B as the multimodal generator, whose 128k-token context window design enables the simultaneous processing of five high-resolution images, ensuring no evidence is truncated during the reasoning phase. With the advancement of large language models, many multimodal models can now accept inputs such as document images, yet their context window size remains constrained. As indicated by the results, most current models exhibit limited capacity for processing lengthy documents, showing a significant accuracy drop on multi-page evidence compared to single-page scenarios. Furthermore, even for multimodal models, accuracy declines relative to text-only conditions when handling different evidence modalities.

M3DocRAG demonstrates some improvement by employing the retrieval methodology of the ColPali model. Our model further advances these results, not only enhancing performance in multi-page and other complex settings but also achieving robust outcomes in the unanswerable category. This underscores the advantages conferred by our dual-channel architecture.

The fusion approach proposed in our Retrieval-Augmented Generation (RAG) framework exerts a crucial influence on the generation performance. In Table 5, diverse fusion strategies determine the degree to which information from each modality is integrated. An excessive inclination towards the visual channel results in outcomes similar to those of a pure visual retriever, thereby relinquishing the advantages of textual semantic retrieval. Conversely, over-dependence on the textual channel leads to a comparable limitation. Significantly, the text-dominant strategy outperformed the vision-dominant strategy. We ascribe this to our preprocessing procedure, which retrieves the original page images for the results from the textual channel, guaranteeing that the final set C_Q encompasses both textual descriptions and their corresponding page images. Surprisingly, the simple merge strategy achieved competitive results. Our LckBorda method, which integrates the merits of both channels via adaptive weighting adjustments, exhibited the highest performance in our experimental assessments.

Table 5. Comparison of Different Fusion Methods.

Title 1	Average	BioM	MacM	ResH	ResSM
LckBorda	63.2	68.5	57.4	63.9	63.1
Single merge	60	64.2	55.3	61.5	59.1
Averaged Weighting	58.3	63	54.1	58.7	57.6
Visual-Dominant	56.9	57.4	52.3	61.5	56.4
Text-Dominant	57.4	59.3	52.6	62.1	55.8

Impact of Hallucination: While VLM descriptions enhance semantic recall, we observed minor instances of generative noise. Our ablation study on fusion methods (Table 4) indirectly supports our defense mechanism: the ‘LckBorda’ fusion outperforms ‘Text-Dominant’ strategies, indicating that balancing generated text with raw visual features effectively corrects potential semantic drifts.

6. Conclusions

This research delves into retrieval-augmented generation (RAG) as a strategy to alleviate the intrinsic problems in large language models, including hallucinations and

knowledge obsolescence. We put forward Dual-channel Retrieval-Augmented Generation for a Hybrid-Modal Document Understanding framework. It facilitates document question answering by amalgamating direct retrieval from a visual channel with conceptual understanding retrieval through generated descriptions in a textual channel. The visual channel, which embeds and retrieves page images, maps queries and raw document information into a unified semantic space, thus significantly improving retrieval accuracy. Simultaneously, the textual channel provides multi-source evidence for query answering by integrating the extracted raw text with generated descriptions.

To evaluate the performance of the proposed framework, experiments were conducted on four benchmark datasets: MPDocVQA, DUDE, ViDoRe-v2, and MMLongBench-Doc. The experimental results presented in Section 4 demonstrate the advancement and effectiveness of our approach, showing significant improvements over prior methods across various tasks, including multi-hop reasoning, layout and chart understanding, handling of unanswerable queries, and long-context comprehension. It is imperative to emphasize that the design objective of DRAG does not lie in the pursuit of extreme performance on open-domain benchmarks. Instead, it aims to develop an interpretable, transferable, and practically deployable Retrieval-Augmented Generation (RAG) system for vertical domains characterized by long documents, multi-modal information, and strong structural correlations, such as medical care and rehabilitation guidance. We anticipate that our work will serve as a reference for visual RAG (VRAG) research and facilitate the practical deployment of efficient multimodal information retrieval and generation in the industrial domain within the era of large models.

Author Contributions: Conceptualization, Z.X. and S.X.; methodology, Z.X., S.X. and X.G.; software, Z.X.; validation, X.G.; resources, Z.X., S.X. and X.G.; data curation, Z.X.; writing—original draft preparation, Z.X. and S.X.; visualization, X.G.; supervision, X.G.; project administration, X.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hebei Provincial High-level Scientific Research Project, grant number QN2025060; and the Tianjin Science and Technology Program Projects, grant number 25ZYCGGX00550.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The funders had no role in the design of the research; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682. [\[CrossRef\]](#)
2. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2025**, arXiv:2303.18223. [\[PubMed\]](#)
3. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [\[CrossRef\]](#)
4. Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*; Park, J.C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., Krisnadhi, A.A., Eds.; Association for Computational Linguistics: Nusa Dua, Bali, 2023; pp. 675–718.
5. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*; JMLR.org: Norfolk, MA, USA, 2020; Volume 119, pp. 3929–3938.

6. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474.
7. Yu, Z.; Xiong, C.; Yu, S.; Liu, Z. Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 2421–2436.
8. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the International Conference on Learning Representations 2024*, Vienna Austria, 7–11 May 2024.
9. Yan, S.-Q.; Gu, J.-C.; Zhu, Y.; Ling, Z.-H. Corrective Retrieval Augmented Generation. *arXiv* **2024**, arXiv:2401.15884. [[CrossRef](#)]
10. Zhang, J.; Zhang, Q.; Wang, B.; Ouyang, L.; Wen, Z.; Li, Y.; Chow, K.-H.; He, C.; Zhang, W. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. *arXiv* **2024**. [[CrossRef](#)]
11. Mishra, A.; Shekhar, S.; Singh, A.K.; Chakraborty, A. OCR-VQA: Visual Question Answering by Reading Text in Images. In *Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR)*, Sydney, Australia, 20–25 September 2019; pp. 947–952.
12. Zhou, J.; Liu, Z.; Xiao, S.; Zhao, B.; Xiong, Y. VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 3185–3200.
13. Zhou, T.; Mei, S.; Li, X.; Liu, Z.; Xiong, C.; Liu, Z.; Gu, Y.; Yu, G. MARVEL: Unlocking the Multi-Modal Capability of Dense Retrieval via Visual Module Plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Ku, L.-W., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 14608–14624.
14. Hayat, M.; Aramvith, S. Transformer’s Role in Brain MRI: A Scoping Review. *IEEE Access* **2024**, *12*, 108876–108896. [[CrossRef](#)]
15. Nguyen, T.; Lei, Y.; Ju, J.-H.; Yates, A. SERVAL: Surprisingly Effective Zero-Shot Visual Document Retrieval Powered by Large Vision and Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V., Eds.; Association for Computational Linguistics: Suzhou, China, 2025; pp. 30795–30810.
16. Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; Bansal, M. M3DocRAG: Multi-Modal Retrieval Is What You Need for Multi-Page Multi-Document Understanding. *arXiv* **2024**, arXiv:2411.04952.
17. Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; Colombo, P. ColPali: Efficient Document Retrieval with Vision Language Models. In *Proceedings of the International Conference on Learning Representations*, Singapore, 24–28 April 2025.
18. Semnani, S.J.; Yao, V.Z.; Zhang, H.C.; Lam, M.S. WikiChat: A Few-Shot LLM-Based Chatbot Grounded with Wikipedia. *arXiv* **2023**, arXiv:2305.14292.
19. Huang, Y.; Huang, J. *A Survey on Retrieval-Augmented Text Generation for Large Language Models*; ACM: New York, NY, USA, 2024; 37p.
20. Robertson, S.E.; Walker, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 232–241.
21. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text Embeddings by Weakly-Supervised Contrastive Pre-Training. *arXiv* **2024**, arXiv:2212.03533. [[CrossRef](#)]
22. Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; Wei, F. Multilingual E5 Text Embeddings: A Technical Report. *arXiv* **2024**, arXiv:2402.05672. [[CrossRef](#)]
23. Formal, T.; Piwowarski, B.; Clinchant, S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2288–2292.
24. Li, H.; Li, Y.; Tian, A.; Tang, T.; Xu, Z.; Chen, X.; Hu, N.; Dong, W.; Li, Q.; Chen, L. A Survey on Large Language Model Acceleration Based on KV Cache Management. *arXiv* **2025**, arXiv:2412.19442.
25. Nogueira, R.; Jiang, Z.; Pradeep, R.; Lin, J. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 708–718.
26. Lin, X.V.; Chen, X.; Chen, M.; Shi, W.; Lomeli, M.; James, R.; Rodriguez, P.; Kahn, J.; Szilvasy, G.; Lewis, M.; et al. RA-DIT: Retrieval-Augmented Dual Instruction Tuning. *arXiv* **2023**, arXiv:2310.01352.
27. Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Lin, X.V.; Smith, N.A.; Zettlemoyer, L.; Yih, W.; Lewis, M. In-Context Pretraining: Language Modeling Beyond Document Boundaries. *arXiv* **2023**, arXiv:2310.10638.
28. Tang, Y.; Zhang, R.; Guo, J.; de Rijke, M.; Chen, W.; Cheng, X. Listwise Generative Retrieval Models via a Sequential Learning Process. *ACM Trans. Inf. Syst.* **2024**, *42*, 1–31. [[CrossRef](#)]

29. Pradeep, R.; Hui, K.; Gupta, J.; Lelkes, A.; Zhuang, H.; Lin, J.; Metzler, D.; Tran, V. How Does Generative Retrieval Scale to Millions of Passages? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 1305–1321.
30. Zeng, H.; Luo, C.; Zamani, H. Planning Ahead in Generative Retrieval: Guiding Autoregressive Generation through Simultaneous Decoding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 469–480.
31. Wu, S.; Wei, W.; Zhang, M.; Chen, Z.; Ma, J.; Ren, Z.; de Rijke, M.; Ren, P. Generative Retrieval as Multi-Vector Dense Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1828–1838.
32. Ye, L.; Lei, Z.; Yin, J.; Chen, Q.; Zhou, J.; He, L. Boosting Conversational Question Answering with Fine-Grained Retrieval-Augmentation and Self-Check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 2301–2305.
33. Lai, Y.; Wu, J.; Zhang, C.; Sun, H.; Zhou, D. AdaCQR: Enhancing Query Reformulation for Conversational Search via Sparse and Dense Retrieval Alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*; Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2025; pp. 7698–7720.
34. Li, X.; Song, L.; Jin, L.; Mi, H.; Ouyang, J.; Yu, D. A Knowledge Plug-and-Play Test Bed for Open-Domain Dialogue Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., Xue, N., Eds.; ELRA and ICCL: Torino, Italy, 2024; pp. 666–676.
35. Alonso, N.; Figliolia, T.; Ndirango, A.; Millidge, B. Toward Conversational Agents with Context and Time Sensitive Long-Term Memory. *arXiv* **2024**, arXiv:2406.00057.
36. He, Y.; Huang, C.; Li, Z.; Huang, J.; Yang, Y. VisPlay: Self-Evolving Vision-Language Models from Images. *arXiv* **2025**, arXiv:2511.15661.
37. Wei, C.; Chen, Y.; Chen, H.; Hu, H.; Zhang, G.; Fu, J.; Ritter, A.; Chen, W. UniIR: Training and Benchmarking Universal Multimodal Information Retrievers. In *Proceedings of the Computer Vision—ECCV 2024: 18th European Conference, Milan, Italy, 29 September–4 October 2024, Proceedings, Part LXXXVII*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 387–404.
38. Chen, W.; Hu, H.; Chen, X.; Verga, P.; Cohen, W. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Abu Dhabi, United Arab Emirates, 2022; pp. 5558–5570.
39. Hu, X.; Liu, G.; Chen, C.; Zhao, Y.; Zhang, H.; Liu, X. TransDiffSBDD: Causality-Aware Multi-Modal Structure-Based Drug Design. *arXiv* **2025**. [[CrossRef](#)]
40. Ma, Y.; Li, J.; Zang, Y.; Wu, X.; Dong, X.; Zhang, P.; Cao, Y.; Duan, H.; Wang, J.; Cao, Y.; et al. Towards Storage-Efficient Visual Document Retrieval: An Empirical Study on Reducing Patch-Level Embeddings. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*; Che, W., Nabende, J., Shutova, E., Pilehvar, M.T., Eds.; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 19568–19580.
41. Tito, R.; Karatzas, D.; Valveny, E. Hierarchical Multimodal Transformers for Multipage DocVQA. *Pattern Recognit.* **2023**, *144*, 109834. [[CrossRef](#)]
42. Landeghem, J.V.; Powalski, R.; Tito, R.; Jurkiewicz, D.; Blaschko, M.; Borchmann, Ł.; Coustaty, M.; Moens, S.; Pietruszka, M.; Ackaert, B.; et al. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Paris, France, 2023; pp. 19471–19483.
43. Macé, Q.; Loison, A.; Faysse, M. ViDoRe Benchmark V2: Raising the Bar for Visual Retrieval. *arXiv* **2025**, arXiv:2505.17166.
44. Ma, Y.; Zang, Y.; Chen, L.; Chen, M.; Jiao, Y.; Li, X.; Lu, X.; Liu, Z.; Ma, Y.; Dong, X.; et al. MMLONGBENCH-DOC: Benchmarking Long-Context Document Understanding with Visualizations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 95963–96010.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.