ColPali: EFFICIENT DOCUMENT RETRIEVAL WITH VISION LANGUAGE MODELS

Manuel Faysse[*1,3] **Hugues Sibille**[*1,4] **Tony Wu**[*1] **Bilel Omrani**[1]
Gautier Viaud[1] **Céline Hudelot**[3] **Pierre Colombo**[2,3]
[1]Illuin Technology [2]Equall.ai [3]CentraleSupélec, Paris-Saclay [4]ETH Zürich
manuel.faysse@centralesupelec.fr

ABSTRACT

Documents are visually rich structures that convey information through text, but also figures, page layouts, tables, or even fonts. Since modern retrieval systems mainly rely on the textual information they extract from document pages to index documents -often through lengthy and brittle processes-, they struggle to exploit key visual cues efficiently. This limits their capabilities in many practical document retrieval applications such as Retrieval Augmented Generation (RAG). To benchmark current systems on visually rich document retrieval, we introduce the Visual Document Retrieval Benchmark *ViDoRe*, composed of various page-level retrieval tasks spanning multiple domains, languages, and practical settings. The inherent complexity and performance shortcomings of modern systems motivate a new concept; doing document retrieval by directly embedding the images of the document pages. We release *ColPali*, a Vision Language Model trained to produce high-quality multi-vector embeddings from images of document pages. Combined with a late interaction matching mechanism, *ColPali* largely outperforms modern document retrieval pipelines while being drastically simpler, faster and end-to-end trainable. We release models, data, code and benchmarks under open licenses at https://hf.co/vidore.

1 INTRODUCTION

Document Retrieval consists of matching a user query to relevant documents in a given corpus. It is central to many widespread industrial applications, either as a standalone ranking system (search engines) or as part of more complex information extraction or Retrieval Augmented Generation (RAG) pipelines.

Over recent years, pretrained language models have enabled large improvements in text embedding models. In practical industrial settings, however, the primary performance bottleneck for efficient document retrieval stems not from embedding model performance but from the prior data ingestion pipeline. Indexing a standard PDF document involves several steps. First, PDF parsers or Optical Character Recognition (OCR) systems are used to extract words from the pages. Document layout detection models can then be run to segment paragraphs, titles, and other page objects such as tables, figures, and headers. A chunking strategy is then defined to group text passages with some semantical coherence, and modern retrieval setups may even integrate a captioning step to describe visually rich elements in a natural language form, more suitable for embedding models. In our experiments (Table 2), we typically find that optimizing the ingestion pipeline yields much better performance on visually rich document retrieval than optimizing the text embedding model.

Contribution 1: *ViDoRe*. In this work, we argue that document retrieval systems should not be evaluated solely on the capabilities of text embedding models (Bajaj et al., 2016; Thakur et al., 2021; Muennighoff et al., 2022), but should also consider the context and visual elements of the documents to be retrieved. To this end, we create and openly release *ViDoRe*, a comprehensive benchmark to evaluate systems on page-level document retrieval with a wide coverage of domains, visual elements, and languages. *ViDoRe* addresses practical document retrieval scenarios, where

[*]Equal Contribution

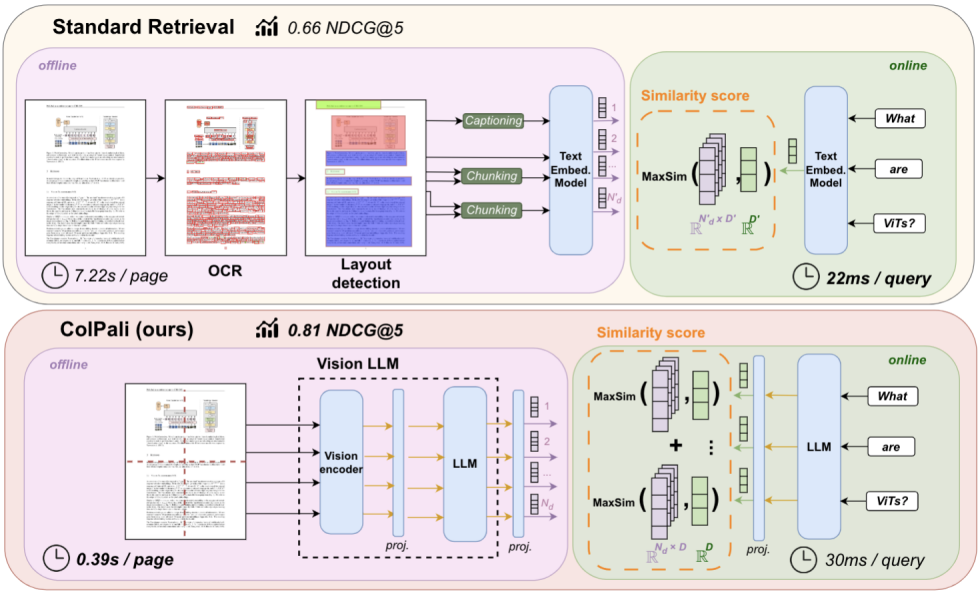

Figure 1: *ColPali* simplifies document retrieval w.r.t. standard retrieval methods while achieving stronger performances with better latencies. Latencies and results are detailed in section 5 and subsection B.4.

queries often necessitate both textual and visual understanding for accurate document matching. We highlight the shortcomings of current text-centric systems in these settings.[1]

Contribution 2: *ColPali*. We propose a novel concept and model architecture based on Vision Language Models (VLMs) to efficiently index documents purely from their visual features, allowing for subsequent fast query matching with late interaction mechanisms (Khattab & Zaharia, 2020). Our method, *ColPali*, significantly outperforms all other retrieval systems on *ViDoRe* while being fast and end-to-end trainable. These results demonstrate the potential and the many benefits of this novel *Retrieval in Vision Space* concept, which could significantly alter the way document retrieval is approached in the industry moving forward. We release all resources at `https://hf.co/vidore`.

2 Problem Formulation & Related Work

Problem Setting. In our setting, a retrieval system scores how relevant a document d from corpus $\mathcal{D}$ is with respect to a query q. Computing the similarity score $s(q, d) \in \mathbb{R}$ for each of the $|\mathcal{D}|$ documents in the corpus creates a ranking we can use to extract the most relevant documents. In this work, we focus on page-level retrieval: *given a query, is the correct document page retrieved by the system?* For coherence with existing literature, we further use the term *document* to refer to individual pages, i.e. the atomic retrieved elements in our setting. As we focus on practical industrial retrieval applications (RAG, search engines) with potentially large corpora sizes, latency constraints are imposed on scoring systems. Most current retrieval systems can be decomposed into (1) an *offline* indexation phase in which a document index is built and (2) an *online* querying phase in which a query is matched to documents from the index and where low latency is vital to the user experience.

Under these industrial constraints, we identify three main properties an efficient document retrieval systems should exhibit: ***(R1) strong retrieval performance****, as measured by standard retrieval metrics;* ***(R2) fast online querying****, measured through average latencies;* ***(R3) high throughput corpus indexation****, ie. the number of pages that can be embedded in a given timeframe.*

[1]The *ViDoRe* benchmark leaderboard is hosted publicly at `https://huggingface.co/spaces/vidore/vidore-leaderboard` to encourage further developments.

2.1 TEXTUAL RETRIEVAL METHODS

Document Retrieval in Text Space.

Statistical methods based on word frequency like TF-IDF (Sparck Jones, 1972) and BM25 (Robertson et al., 1994) are still widely used due to their simplicity and efficiency. More recently, neural embedding models based on fine-tuned large language models display state-of-the-art performance on a variety of text embedding tasks and top the retrieval leaderboards (Muennighoff et al., 2022).

Neural Retrievers. In bi-encoder models (Reimers & Gurevych, 2019; Karpukhin et al., 2020; Wang et al., 2022), documents are independently mapped *offline* to a dense vector space. Queries are embedded *online* and matched to documents through a fast cosine distance computation. A slower, but slightly more performant alternative, cross-encoder systems (Wang et al., 2020; Cohere, 2024) concatenate query and document as a single input sequence and iteratively attribute matching scores to each possible combination. This enables full attention computation between query and document terms but comes at the cost of computational efficiency, as $|\mathcal{D}|$ encoding passes must be done online.

Multi-Vector retrieval via late interaction. In the late interaction paradigm introduced by ColBERT (Khattab & Zaharia, 2020), an embedding is pre-computed and indexed per document token. At runtime, similarity can be computed with individual query token embeddings. The idea is to benefit from the rich interaction between individual query and document terms while taking advantage of the offline computation and fast query matching enabled by bi-encoders. See section E for more details.

Retrieval Evaluation. Although benchmarks and leaderboards have been developed to evaluate text embedding models (Thakur et al., 2021; Muennighoff et al., 2022), much of the performance improvements in industrial use cases of embedding models stem from the prior data ingestion pipeline. While documents often rely on visual elements to more efficiently convey information to human readers, text-only systems barely tap into these visual cues. Other work has also independently studied table or chart retrieval systems through repurposed Question Answering datasets (Zhang et al., 2019; Nowak et al., 2024) but only assessing specialized methods for each task.

To our knowledge, no benchmark evaluates document retrieval systems in practical settings; in an end-to-end manner, across several document types and topics, and by evaluating the use of both textual and visual document features.

2.2 INTEGRATING VISUAL FEATURES

Contrastive Vision Language Models. Mapping latent representations of textual content to corresponding representations of visual content has been done by aligning disjoint visual and text encoders through contrastive losses (Radford et al., 2021; Zhai et al., 2023). While some OCR capabilities exist in these models, the visual component is often not optimized for text understanding.

The Fine-grained Interactive Language-Image Pre-training (Yao et al., 2021) framework extends the late interaction mechanism to cross-modal Vision Language Models, relying on max similarity operations between text tokens and image patches.

Visually Rich Document Understanding. To go beyond text, some document-focused models jointly encode text tokens alongside visual or document layout features (Appalaraju et al., 2021; Kim et al., 2021; Huang et al., 2022; Tang et al., 2022). Large Language transformer Models (LLMs) with strong reasoning capabilities have recently been combined with Vision Transformers (ViTs) (Dosovitskiy et al., 2020) to create VLMs (Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2023; Laurençon et al., 2024b) where image patch vectors from contrastively trained ViT models (Zhai et al., 2023) are fed as input embeddings to the LLM and concatenated with the text-token embeddings.

PaliGemma. The PaliGemma-3B model (Beyer et al., 2024) extends concepts from Pali3 (Chen et al., 2023), and projects `SigLIP-So400m/14` (Alabdulmohsin et al., 2023) patch embeddings into Gemma-2B's text vector space (Gemma Team et al., 2024). Along with its reasonable size w.r.t. other performant VLMs, an interesting property of PaliGemma's text model is that it is fine-tuned with full-block attention on the prefix (instruction text and image tokens). See Appendix E for more details.

VLMs display enhanced capabilities in Visual Question Answering, captioning, and document understanding (Yue et al., 2023), but are not optimized for retrieval tasks.

3 THE *ViDoRe* BENCHMARK

Existing benchmarks for contrastive vision-language models primarily evaluate retrieval for natural images (Lin et al., 2014; Borchmann et al., 2021; Thapliyal et al., 2022). On the other hand, textual retrieval benchmarks (Muennighoff et al., 2022) are evaluated at at textual passage level and are not tailored for document retrieval tasks. We fill the gap with *ViDoRe*, a comprehensive benchmark for document retrieval using visual features.

3.1 BENCHMARK DESIGN

ViDoRe is designed to comprehensively evaluate retrieval systems on their capacity to match queries to relevant documents at the page level. This benchmark encompasses multiple orthogonal subtasks, with focuses on various modalities - text, figures, infographics, tables; thematic domains - medical, business, scientific, administrative; or languages - English, French. Tasks also span varying levels of complexity, in order to capture signals from both weaker and stronger systems. As many systems require large amounts of time to index pages (captioning-based approaches can take dozens of seconds per page for instance), we limit the number of candidate documents for each retrieval task in order to evaluate even complex systems in a reasonable timeframe without sacrificing quality. For trainable retrieval systems, we provide a reference training set that can be used to facilitate comparisons.

Dataset	Language	# Queries	# Documents	Description
Academic Tasks				
DocVQA	English	500	500	Scanned documents from UCSF Industry
InfoVQA	English	500	500	Infographics scrapped from the web
TAT-DQA	English	1600	1600	High-quality financial reports
arXiVQA	English	500	500	Scientific Figures from arXiv
TabFQuAD	French	210	210	Tables scrapped from the web
Practical Tasks				
Energy	English	100	1000	Documents about energy
Government	English	100	1000	Administrative documents
Healthcare	English	100	1000	Medical documents
AI	English	100	1000	Scientific documents related to AI
Shift Project	French	100	1000	Environmental reports

Table 1: *ViDoRe* comprehensively evaluates multimodal retrieval methods.

Academic Tasks. We repurpose widely used visual question-answering benchmarks for retrieval tasks: for each page-question-answer triplet, we use the question as the query, and the associated page as the gold document (Table 1). These academic datasets either focus on single specific modalities (Mathew et al., 2020; 2021; Li et al., 2024) or target more varied visually rich documents (Zhu et al., 2022). Moreover, we consider TabFQuAD, a human-labeled dataset on tables extracted from French industrial PDF documents released with this work. Details can be found in subsection A.1.

Practical tasks. We construct topic-specific retrieval benchmarks spanning multiple domains to go beyond repurposed QA datasets and evaluate retrieval in more realistic industrial situations (e.g. RAG). To achieve this, we collect publicly accessible PDF documents and generate queries pertaining to document pages using Claude-3 Sonnet, a high-quality proprietary vision-language model (Anthropic, 2024). In total, we collect 1,000 document pages per topic, which we associate with 100 queries extensively filtered for quality and relevance by human annotators. The corpus topics are intentionally specific to maximize syntactic proximity between documents, creating more challenging retrieval tasks and covering an array of orthogonal domains (Table 1). [2]

Evaluation Metrics. We evaluate performance on our benchmark (Requirement R1) using standard metrics from the retrieval literature (nDCG, Recall@K, MRR). We report nDCG@5 values as the main performance metric in this work and release the complete sets of results along with the models[3].

[2]Answers are generated alongside queries to (1) ground queries and improve their quality and (2) provide resources to foster future work.

[3]https://huggingface.co/vidore

To validate compliance with practical industrial requirements (section 2), we also consider query latencies (R2) and indexing throughputs (R3).

3.2 Assessing Current Systems

Unstructured. We evaluate retrieval systems representative of those found in standard industrial RAG pipelines. As is common practice, we rely on the `Unstructured`[4] off-the-shelf tool in the highest resolution settings to construct high-quality text chunks from PDF documents. `Unstructured` orchestrates the document parsing pipeline, relying on deep learning vision models to detect titles and document layouts (Ge et al., 2021), OCR engines (Smith, 2007) to extract text in non-native PDFs, specialized methods or models to detect and reconstruct tables, and implements a chunking strategy (`by-title`) that leverages the detected document structure to preserve section boundaries when concatenating texts. As is common practice, in our simplest `Unstructured` configuration (*text-only*), only textual elements are kept and figures, images, and tables are considered noisy information and are filtered out.

Unstructured + X. While `Unstructured` is a strong baseline by itself, we further augment `Unstructured`'s output by integrating the visual elements. In (+ *OCR*), tables, charts, and images are run through an OCR engine, processed by Unstructured, and chunked independently. In (+ *Captioning*), we set up a fully-fledged captioning strategy (Zhao et al., 2023), in which we feed visual elements to a strong proprietary Vision Language Model (Claude-3 Sonnet (Anthropic, 2024)) to obtain highly detailed textual descriptions of the elements. Both strategies aim to integrate visual elements in the retrieval pipeline but incur significant latency and resource costs (subsection 5.2).

Embedding Model. To embed textual chunks, we evaluate Okapi BM25, the *de facto* standard sparse statistical retrieval method, and the dense encoder of BGE-M3 (Chen et al., 2024), a multilingual neural method with SOTA performance in its size category. Chunks are embedded and scored independently, and page-level scores are obtained by max-pooling over the page's chunk scores.[5]

Contrastive VLMs. We also evaluate the strongest available vision-language embedding models; Jina CLIP (Koukounas et al., 2024), Nomic Embed Vision (Nomic, 2024), and `SigLIP-So400m/14` (Alabdulmohsin et al., 2023).

Results. From a performance perspective, best results are obtained by combining the `Unstructured` parser with visual information, either from captioning strategies or by running OCR on the visual elements (Table 2). Little difference is seen between BM25 and BGE-M3 embeddings highlighting the visual information bottleneck. Contrastive VLMs lag behind. Beyond retrieval performance (R1), the indexing latencies (R2) reported in Figure 2 illustrate that PDF parsing pipelines can be very lengthy, especially when incorporating OCR or captioning strategies. Querying latencies at runtime (R3) are very good for all evaluated systems ($\leq$ 22 ms on a NVIDIA L4) due to fast query encoding and cosine similarity matching.

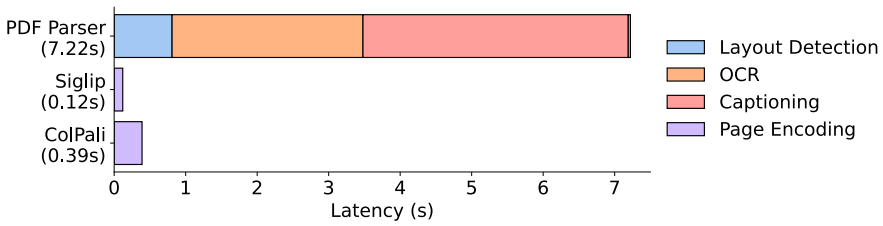

Figure 2: Offline document indexing with *ColPali* is much simpler and faster compared to standard retrieval methods. The PDF Parser results are obtained following the *Unstructured* settings with BGE-M3 detailed in subsection 3.2. All indexing speeds are averaged per-page latencies. More details in subsection B.4

.

[4] `www.unstructured.io`

[5] We empirically validated the max-pooling strategy over sub-page chunks to be more effective than concatenating all page chunks before embedding pagewise.