Unlocking Multimodal Document Intelligence: From Current Triumphs to Future Frontiers of Visual Document Retrieval in the Era of Large Language Models

Yibo Yan *yanyibo70@gmail.com*
Hong Kong University of Science and Technology (Guangzhou)
Alibaba Cloud Computing
Hong Kong University of Science and Technology

Jiahao Huo
Hong Kong University of Science and Technology (Guangzhou)
Alibaba Cloud Computing
University of Illinois Chicago

Guanbo Feng
Hong Kong University of Science and Technology (Guangzhou)

Mingdong Ou*
Alibaba Cloud Computing

Yi Cao
Alibaba Cloud Computing

Xin Zou
Hong Kong University of Science and Technology (Guangzhou)
Hong Kong University of Science and Technology

Shuliang Liu
Hong Kong University of Science and Technology (Guangzhou)
Hong Kong University of Science and Technology

Yuanhuiyi Lyu
Hong Kong University of Science and Technology (Guangzhou)
Hong Kong University of Science and Technology

Yu Huang
Hong Kong University of Science and Technology (Guangzhou)
Alibaba Cloud Computing

Jungang Li
Hong Kong University of Science and Technology (Guangzhou)

Kening Zheng
University of Illinois Chicago

Xu Zheng
Hong Kong University of Science and Technology (Guangzhou)
Hong Kong University of Science and Technology

Philip S. Yu
University of Illinois Chicago

James Kwok
Hong Kong University of Science and Technology

Xuming Hu† *xuminghu@hkust-gz.edu.cn*
Hong Kong University of Science and Technology (Guangzhou)
Hong Kong University of Science and Technology

arXiv:2602.19961v3 [cs.CL] 7 Sep 2026

*Project Lead.
†Corresponding author.

Abstract

With the rapid proliferation of multimodal information, Visual Document Retrieval (VDR) has emerged as a critical frontier in bridging the gap between unstructured visually rich data and precise information acquisition. Unlike traditional natural image retrieval, visual documents exhibit unique characteristics defined by dense textual content, intricate layouts, and fine-grained semantic dependencies. This paper surveys the VDR landscape **as a retrieval problem in its own right**, rather than as the front end of a generation pipeline, and does so specifically through the lens of the Multimodal Large Language Model (MLLM) era. We begin by examining the benchmark landscape, including the recent turn toward reasoning-intensive evaluation, and then dive into the methodological evolution along two orthogonal axes: *what a retriever is*, spanning multimodal *embedding models* and *reranker models*, and *how it is deployed*, from single-stage retrieval through *Retrieval-Augmented Generation* (RAG) to *Agentic systems*. Cutting across both, we analyse the representation–efficiency trade-off that late interaction imposes. We ground these categories in leaderboard evidence on where the empirical frontier actually lies, and close by identifying persistent challenges and outlining promising future directions for multimodal document intelligence.

1 Introduction

Multimodal retrieval, the task of retrieving relevant multimodal information from a large-scale collection using queries that span multiple modalities like text and vision, has become a cornerstone of modern information retrieval (Mei et al., 2025; Zheng et al., 2025a). Historically, research in this domain has predominantly focused on natural image retrieval, targeting datasets of photographs and web images where the primary goal is to match objects, scenes, or holistic visual concepts (Wu et al., 2024a; Arslan et al., 2024). However, both academia and industry have begun to turn their attention to a distinct yet ubiquitous data type: **visual documents**[1]. These documents, ranging from scanned PDFs and business reports to invoices and academic papers, are characterized by a dense interplay of textual content, complex layouts, and graphical elements (Tang et al., 2023; Li et al., 2024d). This shift necessitates an exploration of retrieval techniques tailored to the unique structure and information density of visual documents, pushing the boundaries of what multimodal systems can achieve.

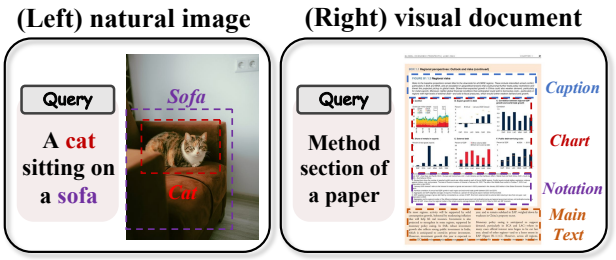
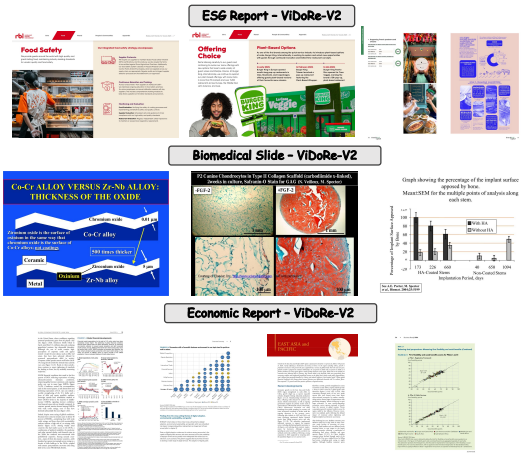


Figure 1: Natural-image retrieval (*left*) versus visual-document retrieval (*right*), the focus of this survey.

The pivot towards **Visual Document Retrieval** (VDR) is driven by three fundamental differences that distinguish visual documents from natural images, as illustrated in Figure 1. ❶ **Information modality and density**: unlike natural images which convey semantic meaning through holistic scenes, visual documents are hybrid entities where meaning is co-determined by rich textual information and a structured spatial layout. The information is dense, hierarchical, and multi-modal by nature. ❷ **Semantic granularity**: retrieval in natural images often targets high-level concepts (*e.g.,* "a cat sitting on a sofa"), whereas VDR demands a much finer-grained understanding. Users may query for specific facts embedded within a table, a particular sentence in a paragraph, or information contingent on its document-level position (*e.g.,* "the methodology section of a paper"). ❸ **User intent and task complexity**: VDR is geared towards precise information-seeking, question answering, and evidence-based reasoning, rather than conceptual or aesthetic matching.

[1]They are also commonly referred to as *visually rich documents*, *document images*, *etc.* We use "visual documents" as a unifying term. See the illustrative examples from representative VDR datasets in Figure 2.



(a) Illustrative examples of ViDoRe-V2 (Macé et al., 2025).



(b) Illustrative examples of JinaVDR (Günther et al., 2025).

Figure 2: Illustrative examples from two representative datasets, ViDoRe-V2 (a) and JinaVDR (b).

Furthermore, as the general capabilities of Multimodal Large Language Models (MLLMs) advance (Song et al., 2025; Yan et al., 2025b;a), the VDR field is increasingly focusing on their integration. This includes the development of MLLM-based embedding and reranker models to enhance semantic matching (Tao et al., 2024; Zhang et al., 2024b; Wang et al., 2024b). Beyond that, there is active exploration into leveraging these models within more sophisticated frameworks like Retrieval-Augmented Generation (RAG) pipelines (Gao et al., 2023; Cheng et al., 2025; Gan et al., 2025a) and Agentic systems (Singh et al., 2025) to tackle complex document-based settings.

Table 1: Coverage of prior surveys relative to ours. Instead of tagging each work with a single Retrieval/Understanding label, a dichotomy that misrepresents retrieval-oriented surveys that happen to be framed around RAG, we score four *checkable* axes of the retrieval stack: dedicated treatment of **Emb**edding models, of **Rer**ankers as a first-class stage, of index **Eff**iciency (pruning/merging/compression), and of **Rea**soning-intensive retrieval. ✔ = dedicated section or table; ✓ = discussed without dedicated treatment; blank = absent. Each judgement is reproducible from the cited work's table of contents.

Survey	Venue	Scope	Retrieval-Stack Coverage				Agent
			Emb	Rer	Eff	Rea	
Alaei et al. (2016)	IJCNN'16	IR4Doc					
Zhou et al. (2017)	arxiv'17	IR4Img					
Subramani et al. (2020)	NeurIPS Workshop'20	DL4Doc					
Cui et al. (2021)	ICDAR'21	IR4Doc					
Hameed et al. (2021)	CE'21	IR4Img					
Sassioui et al. (2023)	WINCOM'23	IR4Doc					
Zhao et al. (2023)	EMNLP Findings'23	LLM4Img	✓				
Ding et al. (2024)	arxiv'24	DL4Doc	✓				
Zhang et al. (2025c)	arxiv'25	LLM4Img	✔	✓		✓	
Ding et al. (2025)	IJCAI Tutorial'25	LLM4Doc	✓				
Zhang (2025)	AACL-IJCNLP'25	LLM4Doc	✔	✓	✓		
Abootorabi et al. (2025)	ACL Findings'25	LLM4MM	✔	✓	✓	✓	✔
Huang et al. (2025b)	IEEE TKDE'25	LLM4Doc	✓				✔
Gao et al. (2026)	ACL'26	LLM4Doc	✔	✓	✔	✓	✔
Ke et al. (2026)	ACM TOIS'26	LLM4Doc	✓			✓	✔
Rombach & Fettke (2026)	ACM Computing Survey'26	DL4Doc	✓				
Ours	-	**LLM4Doc**	✔	✔	✔	✔	✔

Scope. Prior reviews approach this material from adjacent vantage points (Table 1). Earlier work concentrated on classical document information retrieval (Alaei et al., 2016) and deep learning for document understanding (Subramani et al., 2020; Sassioui et al., 2023; Ding et al., 2024), or on retrieval for natural images (Zhou et al., 2017; Hameed et al., 2021). Surveys written after the rise of MLLMs largely retain a generation-centric or understanding-centric framing (Huang et al., 2025b; Rombach & Fettke, 2026; Ding et al., 2025; Ke et al., 2026) or target natural-image retrieval (Zhao et al., 2023; Zhang et al., 2025c).

Three recent works are close enough that we state the difference explicitly rather than rely on a checkmark. Abootorabi et al. (2025) survey multimodal RAG broadly; retrieval appears there as one component of a generation pipeline, and visually rich documents are one modality among many, so the document-specific retrieval stack (layout-aware multi-vector representations, index compression, document rerankers) is not the organising concern. Gao et al. (2026) is genuinely retrieval-oriented and covers embedding models, the ColPali family, and compression; its framing runs from multimodal RAG toward document understanding, and we differ in emphasis rather than in kind, treating retrieval as the object of study instead of the front end of a generator. Closest of all is Zhang (2025), a VDR retrieval survey with substantially overlapping scope; relative to it we add a dedicated treatment of rerankers as a first-class stage with its ranking-paradigm taxonomy, the

reasoning-intensive turn in benchmarks, an empirical reading of the current leaderboard frontier, and the efficiency literature that has appeared since.

Our contribution is therefore not to be first to mention VDR, but to organise it as a retrieval problem in its own right: we **treat embedding, reranking, and agentic deployment along explicit axes, and pair each with the benchmark and efficiency evidence needed to compare methods**[2].

Structure. We begin from a **benchmark perspective**, systematically organizing the field by examining task formulations, basic settings, and dataset characteristics such as multilingual support and the growing emphasis on reasoning-intensive queries. We then turn to a **methodology-centric analysis**.

Here we make our organising principle explicit, because a flat list of "paradigms" would conflate two questions that vary independently. We use **two orthogonal axes and one cross-cutting dimension**. The first axis asks *what a retriever is*, that is, its representation and scoring mechanism, and runs from single-vector bi-encoders through multi-vector late interaction (▷ Section 3.1) to cross-encoder joint scoring (▷ Section 3.2) and, most recently, generative formulations. The second axis asks *how a retriever is deployed*: single-stage retrieval, retrieve-then-rerank cascades, static RAG pipelines, or iterative agentic loops (▷ Section 3.3). These axes are genuinely independent: a multi-vector embedder can be deployed single-stage or inside an agent, and an agentic system can be assembled from any underlying retriever. We note explicitly that the second axis is *downstream* of the first, which is why Section 3.3 builds on rather than parallels the two preceding sections. Cutting across both is the **resource–quality trade-off**: the representation choice that wins on accuracy also determines index size and query-time cost, a tension we quantify in Section 3.1 and pursue in Section 4.3.

Finally, we conclude by discussing the challenges and outlining **future frontiers**, aiming to provide valuable insights and inspire subsequent research in the multimodal document intelligence community.

2 Benchmark Perspective

This section provides a systematic review of the VDR evaluation landscape. We first establish a formal mathematical definition (▷ Section 2.1), and then analyze the current trends in terms of data scale and metrics (▷ Section 2.2), followed by a discussion on the emerging frontiers of multilingual support (▷ Section 2.3) and reasoning-intensive retrieval (▷ Section 2.4), which reflect the shift from keyword matching to complex document intelligence.

2.1 Formulation

VDR aims to identify the most relevant document images from a large-scale corpus based on a given query. Formally, let $\mathcal{C} = \{d_n\}_{n=1}^{|\mathcal{C}|}$ be a corpus of document pages, where each $d_n \in \mathcal{I}$ is a visually rich document image. Given a query q, the task is to produce a ranked list of documents from $\mathcal{C}$ such that the top-ranked items maximize a relevance score $s(q, d)$.

Input and Modalities. In the standard VDR setting, the query is typically a natural language string $q \in \mathcal{T}$ (text-to-image retrieval). However, the generalized VDR setting extends the query space to multimodal inputs, including images or interleaved text-and-image sequences $q \in \{\mathcal{T} \cup \mathcal{I}\}^*$. Unlike traditional OCR-based retrieval that treats documents as plain text, VDR models process the document d directly in the visual domain, often representing it as a set of patch-level embeddings to preserve layout and graphical information.

Mathematical Representation. Following the late-interaction paradigm pioneered by ColPali (Faysse et al., 2024), let $\mathbf{U}(q) = \{\mathbf{u}_i\}_{i=1}^{n_q}$ denote the query-side embeddings (token-level for the standard text-query setting) and $\mathbf{V}(d) = \{\mathbf{v}_j\}_{j=1}^{n_d}$ the patch-level document embeddings, with $\mathbf{u}_i, \mathbf{v}_j \in \mathbb{R}^h$. Relevance can then be computed with a late-interaction operation such as `MaxSim`: $s(q, d) = \sum_{i=1}^{n_q} \max_{j=1}^{n_d} \mathbf{u}_i^\top \mathbf{v}_j$.

Evaluation Metrics. This section provides detailed mathematical formulations for the primary retrieval metrics used in the benchmarks discussed in this survey. For a given set of test queries $\mathcal{Q}$, the goal of a retrieval system is to return a ranked list of documents from a corpus $\mathcal{C}$ for each query $q \in \mathcal{Q}$.

[2]This version covers literature up to **25 August, 2026**, with updates scheduled every two months for the fast-growing research community.

❶ **Recall@k** Recall at k (Recall@k) measures the fraction of relevant documents that are successfully retrieved within the top-k results. It is a metric of coverage, evaluating how well the system is able to find all ground-truth documents. For a given query q, let $\mathcal{D}_q^+$ be the set of ground-truth relevant documents and $\pi_k(q)$ be the ranked list of the top-k retrieved documents. The overall Recall@k is the average score across all queries in $\mathcal{Q}$.

Recall@k is calculated as:

$$\text{Recall@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{|\mathcal{D}_q^+ \cap \pi_k(q)|}{|\mathcal{D}_q^+|} \tag{1}$$

where $|\mathcal{Q}|$ is the total number of queries; $|\mathcal{D}_q^+|$ is the total number of relevant documents for query q; and $|\mathcal{D}_q^+ \cap \pi_k(q)|$ is the number of relevant documents found in the top-k retrieved list.

❷ **Mean Reciprocal Rank (MRR)** Mean Reciprocal Rank (MRR) evaluates the ranking of the *first* correct document retrieved. It is particularly useful for tasks where finding a single relevant item quickly is the primary goal. For each query q, the reciprocal rank is the inverse of the rank of the first relevant document. If no relevant document is retrieved, the reciprocal rank is 0.

MRR is calculated as:

$$\text{MRR} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{\text{rank}_q} \tag{2}$$

where rank_q is the position of the first relevant document for query q, with $\text{rank}_q = \infty$ when no relevant document is retrieved.

❸ **Mean Average Precision (mAP)** Mean Average Precision (mAP) provides a comprehensive evaluation of a ranked list by considering both precision and recall. It rewards retrieving relevant documents at higher ranks. The Average Precision (AP) for a single query is first calculated by averaging the precision at each relevant document's position. mAP is then the mean of these AP scores over all queries.

The Average Precision for a single query q is:

$$\text{AP}_q = \frac{1}{|\mathcal{D}_q^+|} \sum_{r=1}^{|\mathcal{C}|} \text{Prec}_q(r)\ \text{rel}_q(r) \tag{3}$$

And mAP is the mean of all AP scores:

$$\text{mAP} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{AP}_q \tag{4}$$

where $|\mathcal{C}|$ is the total number of documents in the corpus; $\text{Prec}_q(r)$ is the precision for query q at rank r; and $\text{rel}_q(r)$ is 1 if the document at rank r is relevant to q, and 0 otherwise.

❹ **Normalized Discounted Cumulative Gain (nDCG@k)** Normalized Discounted Cumulative Gain (nDCG@k) is a measure of ranking quality that accounts for the graded relevance of documents. It assigns higher scores to more relevant documents placed at top ranks, with a logarithmic discount for documents at lower ranks. The score is normalized by the ideal ranking to fall between 0 and 1.

The DCG@k is first calculated as:

$$\text{DCG@}k(q) = \sum_{r=1}^{k} \frac{2^{g_{q,r}} - 1}{\log_2(r+1)} \tag{5}$$

The nDCG@k is then:

$$\text{nDCG@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{\text{DCG@}k(q)}{\text{IDCG@}k(q)} \tag{6}$$

where k is the number of results being evaluated; $g_{q,r}$ is the graded relevance of the document at rank r for query q; and $\text{IDCG@}k(q)$ is the DCG of the ideal ranking for q up to rank k.

❺ Hit Rate (HR@k) As used in benchmarks like `Double-Bench` (Shen et al., 2025), Hit Rate at k (HR@k) is a binary success metric that evaluates whether *at least one* relevant document is found within the top-k retrieved results. It is particularly useful for assessing whether the retrieval system can find any correct evidence, which is a prerequisite for downstream tasks. For multi-hop queries, a "hit" is only registered if evidence for *all* hops is found within the top-k results.

For a single-hop query q, the HR@k is:

$$\text{HR@}k(q) = \mathbb{I}(\pi_k(q) \cap \mathcal{D}_q^+ \neq \emptyset) \tag{7}$$

The overall HR@k is the average across all queries:

$$\text{HR@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \text{HR@}k(q) \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise; $\pi_k(q)$ is the top-k retrieved list and $\mathcal{D}_q^+$ is the set of ground-truth relevant documents.

❻ Averaged Normalized Longest Common Subsequence (ANLCS) Used in `VisDoMBench` (Suri et al., 2025) to evaluate evidence extraction, Averaged Normalized Longest Common Subsequence (ANLCS) measures the textual similarity between retrieved content and ground-truth evidence. Instead of evaluating ranking, it assesses the quality of the retrieved content itself. First, the Normalized Longest Common Subsequence (NLCS) is calculated for a pair of text strings, measuring their shared content normalized by their lengths.

The NLCS between a ground-truth evidence string x_{gt} and a retrieved chunk string x_{ret} is:

$$\text{NLCS}(x_{\text{gt}}, x_{\text{ret}}) = \frac{2\,|\text{LCS}(x_{\text{gt}}, x_{\text{ret}})|}{|x_{\text{gt}}| + |x_{\text{ret}}|} \tag{9}$$

The ANLCS score for a query is then computed by finding the best-matching retrieved chunk for each ground-truth evidence chunk and averaging these scores. The final metric is the average over all queries.

$$\text{ANLCS@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left(\frac{1}{|\mathcal{G}_q|} \sum_{g \in \mathcal{G}_q} \max_{x \in \mathcal{X}_k(q)} \text{NLCS}(g, x) \right) \tag{10}$$

where $\text{LCS}(x_{\text{gt}}, x_{\text{ret}})$ is the Longest Common Subsequence between the two strings; $|x|$ denotes the length of string x; $\mathcal{G}_q$ is the set of ground-truth evidence texts for query q; and $\mathcal{X}_k(q)$ is the set of retrieved text chunks in the top-k results for query q.

2.2 Basic Setting

Escalating Research Momentum. Over the past two years, VDR has transitioned from a niche task to a central focus in both industry and academia. As shown in Table 2, the majority of specialized VDR benchmarks, such as the ViDoRe series (Faysse et al., 2024) and Real-MM-RAG (Wasserman et al., 2025b), emerged in 2024 and 2025. This surge is driven by the realization that (OCR-based) text-only retrieval fails to capture the visual nuances of documents, such as tables, charts, and spatial hierarchies (Zhang et al., 2025b; Most et al., 2025).

Diverse Dataset Scales. Current benchmarks exhibit a wide range of data magnitudes (Cao et al., 2025). While expert-annotated sets like SeaDoc (Xiao et al., 2025b) and M4DocBench (Dong et al., 2025c) focus on specialized samples, recent large-scale efforts have pushed boundaries. NL-DIR (Guo et al., 2025) provides 205k queries, and Jina-VDR (Günther et al., 2025) utilizes over 70k documents, reflecting a trend toward scaling both query volume and corpus diversity.

Standardized Evaluation. nDCG and Recall remain the primary metrics for performance measurement. However, as VDR is increasingly integrated into RAG, some benchmarks (*e.g.,* EVisRAG (Sun et al., 2025c) and MMLongBench-Doc (Ma et al., 2024b)) incorporate downstream Accuracy and F1 scores to evaluate how retrieval quality directly impacts final multimodal generation.

Category	Benchmark	Publication		Dataset					Resource
		Team	Venue	#Query*	#Corpus*	Multilingual	Reasoning	Retrieval Metric	
VDR	MRMR (Zhang et al., 2026b)	NTU	ICLR'26	1.50k	42.80k	en	✔	nDCG / HR	
VDR	Invoice Haystack (Kanagalingam et al., 2026)	Melbourne	ECCV'26	200	1.50k	en	-	Recall	
VDR	VisDocAgentBench (Hu et al., 2026b)	PKU	arxiv'26	120	2.38k	en	✔	MRR / Recall	
VDR	MMLongEmbed (Wang et al., 2026a)	Soochow & Ant Group	arxiv'26	8.46k	20.88k	en	-	Precision	
VDR	CHaystack (Li, 2026)	SCU	arxiv'26	4.54k	17.49k	zh	-	Recall	
VDR	KoViDoRe (Choi et al., 2026b)	Kyung Hee	arxiv'26	706	6.73k	ko	-	nDCG	
VDR	LëtzCross (El Bachyr et al., 2026)	Luxembourg	arxiv'26	579	908	4	-	nDCG	
VDR	ViDoRe-V3 (Loison et al., 2026)	Illuin Tech	arxiv'26	3.10k	26.00k	en, fr	-	nDCG	
VDR	IRPAPERS (Shorten et al., 2026)	Weaviate	arxiv'26	180	3.23k	en	-	Recall	
VDR	ViDoRe-V1 (Faysse et al., 2024)	Illuin Tech	ICLR'25	3.81k	8.31k	en, fr	-	nDCG / Recall / MRR	
VDR	VisRAG (Yu et al., 2024)	Tsinghua	ICLR'25	3.61k	20.64k	en	-	Recall / MRR	
VDR	SeaDoc (Xiao et al., 2025b)	Alibaba	NeurIPS'25	1.00k	5.06k	4	-	nDCG	
VDR	MMDocRAG (Dong et al., 2025b)	Huawei	NeurIPS'25	4.06k	14.87k	en	-	Recall	
VDR	Real-MM-RAG (Wasserman et al., 2025b)	IBM	ACL'25	4.55k	8.60k	en	-	nDCG / Recall	
VDR	NL-DIR (Guo et al., 2025)	CAS	CVPR'25	205.00k	41.80k	en	-	Recall / MRR	
VDR	OpenDocVQA (Tanaka et al., 2025)	NTT	CVPR'25	43.47k	206.27k	en	-	nDCG	
VDR	ViDoSeek (Wang et al., 2025b)	Alibaba	EMNLP'25	1.14k	5.39k	en	-	Recall / MRR	
VDR	MMDocIR (Dong et al., 2025a)	Huawei	EMNLP'25	1.66k	20.40k	en	-	Recall	
VDR	Jina-VDR (Günther et al., 2025)	Jina AI	EMNLP'25	57.33k	70.62k	20	-	nDCG	
VDR	VisDoMBench (Suri et al., 2025)	UMD	NAACL'25	2.27k	8.30k	en	-	ANLCS	
VDR	ViDoRe-V2 (Macé et al., 2025)	Illuin Tech	arxiv'25	1.15k	4.54k	en, fr	-	nDCG	
VDR	EVisRAG (Sun et al., 2025c)	PKU	arxiv'25	4.26k	-	en	-	Acc / F1	
VDR	Double-Bench (Shen et al., 2025)	SCUT	arxiv'25	5.17k	73.06k	6	-	HR	
VDR	VisR-Bench (Chen et al., 2025c)	UB	arxiv'25	35.57k	35.63k	16	-	Acc	
VDR	MR2-Bench (Zhou et al., 2025a)	BAAI	arxiv'25	1.31k	47.74k	en	✔	nDCG / Recall	
VDR	MIRACL-VISION (Osmulski et al., 2025)	NVIDIA	arxiv'25	7.90k	338.73k	18	-	nDCG	
VDR	UniDoc-Bench (Peng et al., 2025)	Salesforce	arxiv'25	1.74k	70.00k	en	-	Recall / Precision	
VDR	ViMDoc (Kim et al., 2025a)	KAIST	arxiv'25	10.90k	76.35k	en	-	Recall	
VDR	M4DocBench (Dong et al., 2025c)	Huawei	arxiv'25	158	6.18k	en, zh	✔	Recall	
VDR	SDS KoPub VDR (Lee et al., 2025a)	Samsung	arxiv'25	600	40.78k	en, ko	-	nDCG / Recall	
VDR	Nayana-IR (Kolavi & Jain, 2025)	CognitiveLab	arxiv'25	1.00k	5.40k	22	-	nDCG / Recall / mAP / MRR	
VDR	MMLongBench-Doc (Ma et al., 2024b)	NTU	NeurIPS'24	1.08k	6.41k	en	-	Acc / F1	
Gen.	MMEB-V2 (Meng et al., 2025)	Salesforce	TMLR'26	-	-	en	-	nDCG	
Gen.	MM-BRIGHT (Abdallah et al., 2026b)	University of Innsbruck	arxiv'26	2.80k	2.50M	en	✓	nDCG	
Gen.	ARK-Bench (Lin et al., 2026)	SCU	arxiv'26	1.55k	36.03k	en	✓	nDCG / Recall	
Gen.	MMEB (Jiang et al., 2024c)	Waterloo & Salesforce	ICLR'25	12.00k	12.00M	en	-	Precision	
Gen.	Visual Haystacks (Wu et al., 2024b)	UCB	ICLR'25	20.10k	97.00k	en	-	Acc	
Gen.	COCO-Facet (Li et al., 2025c)	Washington	NeurIPS'25	9.11k	911.20k	en	-	Recall	
Gen.	UMRB (Zhang et al., 2024c)	PolyU & Alibaba	CVPR'25	200.00k	40.00M	en	-	nDCG / Recall	
Gen.	MIEB (Xiao et al., 2025c)	Durham University	ICCV'25	1.25M	7.62M	38	-	nDCG / Recall / mAP	
Gen.	MMNeedle (Wang et al., 2025a)	Rutgers University	NAACL'25	280.00k	40.00k	en	-	Acc	
Gen.	MMMEB (Musacchio et al., 2025)	Uniba	arxiv'25	9.20k	9.20k	5	-	Precision	
Gen.	MM-NIAH (Wang et al., 2024c)	FDU & Shanghai AI Lab	NeurIPS'24	5.58k	-	en	-	Acc	
Gen.	M2KR (Lin et al., 2024b)	Cambridge	ACL'24	39.79k	514.85k	en	-	Recall	
Gen.	M-BEIR (Wei et al., 2024a)	Waterloo	ECCV'24	190k	5.6M	en	-	Recall	

Table 2: Comparison of VDR and general multimodal retrieval benchmarks. In the *Resource* column, we denote the corresponding GitHub codebase, Hugging Face repository, and paper (*e.g.,* arXiv paper, technical report, or blog) as / / . * indicates that #query and #corpus correspond to the retrieval-related #evaluation samples and #candidates, respectively, for general multimodal retrieval benchmarks.

2.3 Multilingual Support

The majority of early VDR benchmarks are predominantly English-centric. Following advances in multilingual text-embedding evaluation (Enevoldsen et al., 2025; Zhang et al., 2023), recent VDR work has begun to bridge this linguistic gap. Jina-VDR (Günther et al., 2025) and Nayana-IR (Kolavi & Jain, 2025) represent a significant shift, supporting 20 and 22 languages, respectively. Similarly, MIRACL-VISION (Osmulski et al., 2025) introduces a large-scale multilingual corpus covering 18 languages. This movement highlights the necessity for VDR models to move beyond English-specific representations and develop cross-lingual semantic alignment capabilities, particularly for visually rich documents. Figure 3 makes concrete what that demand involves. Across the six subsets the queries are not translations of one another but are grounded in locally specific artefacts, a Habsburg-era subscription notice, a Belarusian price list, a Niigata ramen catalogue, an Arabic-labelled chart, a Hindi house journal and an Italian letterhead, and the answer span is a different structural object in each. Two consequences follow for benchmark design. First, reporting a single averaged multilingual score conflates script recognition with layout understanding, since a model can fail panel (b) for want of table parsing while failing panel (e) for want of Devanagari reading. Second, right-to-left scripts such as the Arabic of panel (d) alter the reading order that patch-level models implicitly assume, an effect that no amount of additional training data in left-to-right languages will surface.

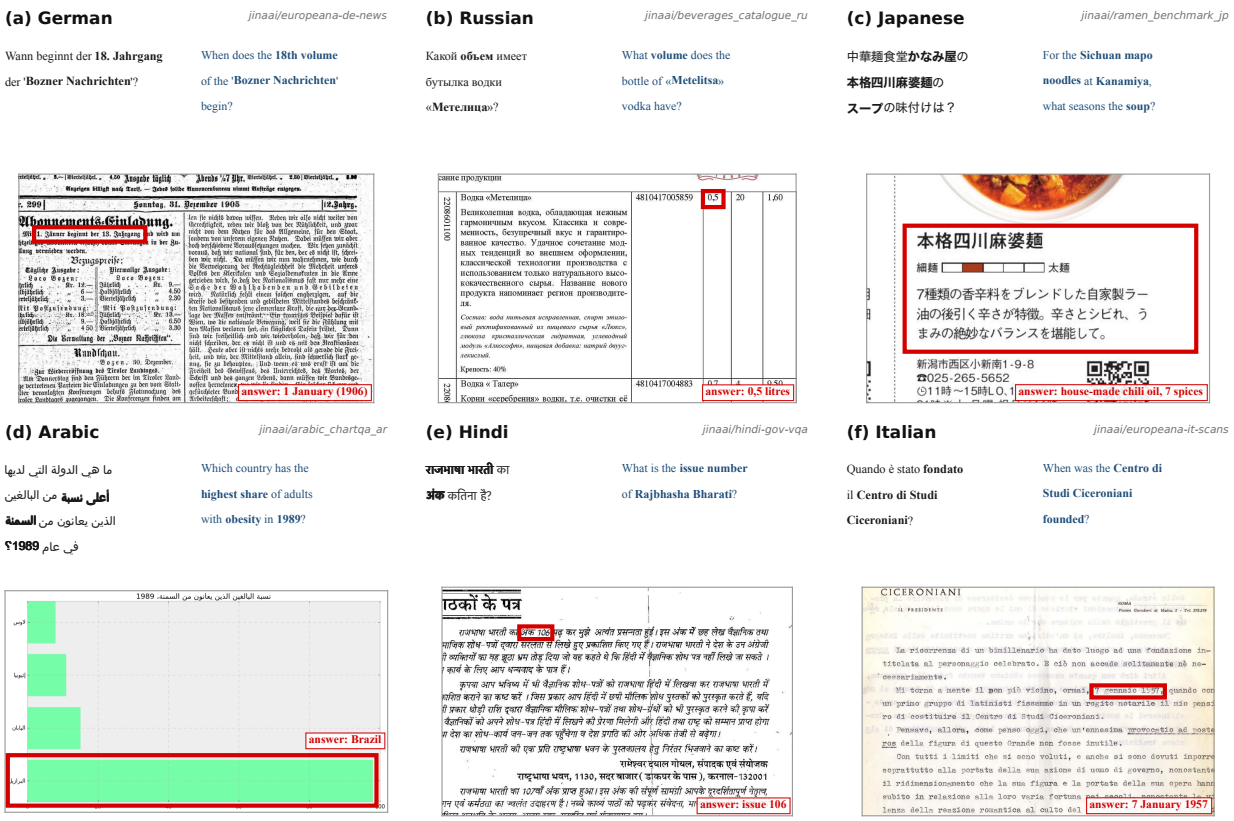


Figure 3: What multilingual VDR actually asks of a retriever. One case from each of six non-English JinaVDR subsets (Günther et al., 2025). Each panel pairs the query in its original script (*left*) with an English gloss (*right*) on the same line, and outlines in **red** the page region that answers it, annotated with the answer itself. Key entities, dates and quantities, the tokens that must survive cross-script alignment, are set in bold in *both* the original and the gloss, so the two readings can be compared term by term. The panels span four writing systems and both text directions, and the evidence takes a different form in each: a masthead notice (a), a single table cell (b), a shop card (c), a bar in a chart (d), an issue number in running prose (e), and a date inside a letter body (f). A retriever therefore needs script-level reading, layout parsing and numeric grounding at once, which is why the reranking gap discussed in Section 3.2 is not merely a matter of adding languages.

Results expose substantial task-level heterogeneity. Figure 4 compares Jina-VDR results on the same six subsets visualized qualitatively above. Jina v4 (multi) ranks first on every subset and reaches a six-task macro-average of 68.09, compared with 60.14 for its single-vector mode and 48.35 for DSE-Qwen2, the strongest non-Jina-v4 baseline on this slice. Because the two Jina v4 modes share a backbone, their comparison controls model capacity more closely than a cross-model contrast: late interaction adds 7.96 nDCG points on average. The gain is nevertheless far from uniform: it is largest on the two Europeana scan tasks, Italian (+15.28) and German (+14.87), but narrows to +2.05 on the Russian catalogue and +3.88 on the Japanese ramen catalogue. This pattern is consistent with, although it does not by itself prove, the hypothesis that patch-level matching is especially useful when a small answer-bearing region must be localized within degraded or layout-heavy pages. The older systems also reverse order across subsets: DSE-Qwen2 is the strongest non-Jina-v4 model for German, Russian and Arabic; Jina v3 + OCR leads that group for Japanese; and ColPali does so for Hindi and Italian. Nor should these scores be interpreted as an isolated ranking of language difficulty: the best score ranges from only 22.49 on Hindi government documents to 94.65 on the Japanese catalogue, while each column simultaneously changes the script, domain, corpus size, scan quality and evidence structure. A single multilingual average therefore hides both architectural sensitivity and task-specific failure modes, and the comparison supports a claim about heterogeneity across *language–document pairs*, rather than a causal claim that one language is intrinsically harder than another.



Figure 4: Model performance on the six subsets illustrated in Figure 3. Results from Jina-VDR (Günther et al., 2025): cells show nDCG@5 (%) on a fixed 0–100 scale; bold values mark the per-column maximum, and the outlined row wins every column. "Macro mean" is the unweighted average over only these six subsets, rather than JinaVDR's all-task average. Each language code is paired with its document type because language, domain and visual format co-vary.

2.4 Reasoning-Intensive Setting

The rapid advancement of MLLMs has expanded information-retrieval evaluation beyond surface-level semantic matching toward tasks that require explicit reasoning (Bi et al., 2025; Su et al., 2025b; Lin et al., 2025b; Yan et al., 2024). **Text-centric retrieval has already developed reasoning-aware capabilities** through benchmarks[3] (*e.g.,* BRIGHT (Su et al., 2024), RAR-b (Xiao et al., 2024), ATEB (Han et al., 2025a), R2MED (Zhang et al., 2025g)), embedding models (*e.g.,* ReasonIR (Shao et al., 2025), ReasonEmbed (Chen et al., 2025d), RaDeR (Das et al., 2025), DIVER-Retriever (Sun et al., 2025a), RITE (Liu et al., 2025f), REAPER (Joshi et al., 2024)), and rerankers (*e.g.,* Reason-to-Rank (Ji et al., 2024), InteRank (Samarinas & Zamani, 2025), Rank1 (Weller et al., 2025b), Rank-K (Yang et al., 2025a), TFRank (Fan et al., 2025), TS-SetRank (Huang et al., 2025a), InsertRank (Seetharaman et al., 2025), RGS (Xu & Chen, 2025)). VDR is now undergoing the same transition: dense textual information combined with intricate visual layouts (*e.g.,* charts, tables, and diagrams) demands logical deduction beyond traditional OCR-based matching (Duan et al., 2025; Liao et al., 2025; Nacson et al., 2025; Jiang et al., 2025).

Recent pioneering efforts have begun to formalize this reasoning-intensive frontier through three distinct perspectives:

❶ **Vision-Centric Logic and Abstract Reasoning.** As highlighted by **MR2-Bench** (Zhou et al., 2025a), current VDR models often suffer from "shallow matching," where they succeed by identifying object-text correlations but fail at logical, spatial, or causal inference. MR2-Bench introduces vision-centric tasks such as *Visual Puzzle* (solving Raven-style matrices) and *Visual Illustration Search* (*e.g.,* matching a mathematical formula to its corresponding geometric proof). Their findings reveal a massive "reasoning gap": models achieving high scores on general multimodal benchmarks (*e.g.,* MMEB) exhibit a significant performance drop when required to solve abstract visual analogies or multi-image relational scenarios.

[3]Instruction-following retrieval (*e.g.,* FollowIR (Weller et al., 2024) and mFollowIR (Weller et al., 2025a)), long-context retrieval (*e.g.,* LongEmbed (Zhu et al., 2024a)) and code retrieval benchmarks (*e.g.,* CoIR (Li et al., 2024b) and CodeRAG-Bench (Wang et al., 2024e)) are not within the scope of reasoning-intensive retrieval in this survey.

❷ Expert-Level Multidisciplinary and Contradiction Reasoning. Moving beyond general knowledge, **MRMR** (Zhang et al., 2026b) introduces *Contradiction Retrieval*, a novel task requiring models to identify rules or requirements that conflict with a given case description (*e.g.,* identifying a traffic violation in a visual scene based on a text-based rulebook). MRMR focuses on expert domains like medicine and engineering, where visually rich documents (*e.g.,* pathological slides) require specialized interpretation. Figure 5 makes the distinction among its three task families concrete: Knowledge grounds a visual symbol in expert facts, Theorem infers and instantiates the rule needed to solve a problem, whereas Contradiction verifies a polarity conflict between visible evidence and a candidate statement. Consequently, success requires an intermediate inference that is not literally shared by the query and relevant document, rather than surface-level image–text overlap alone.

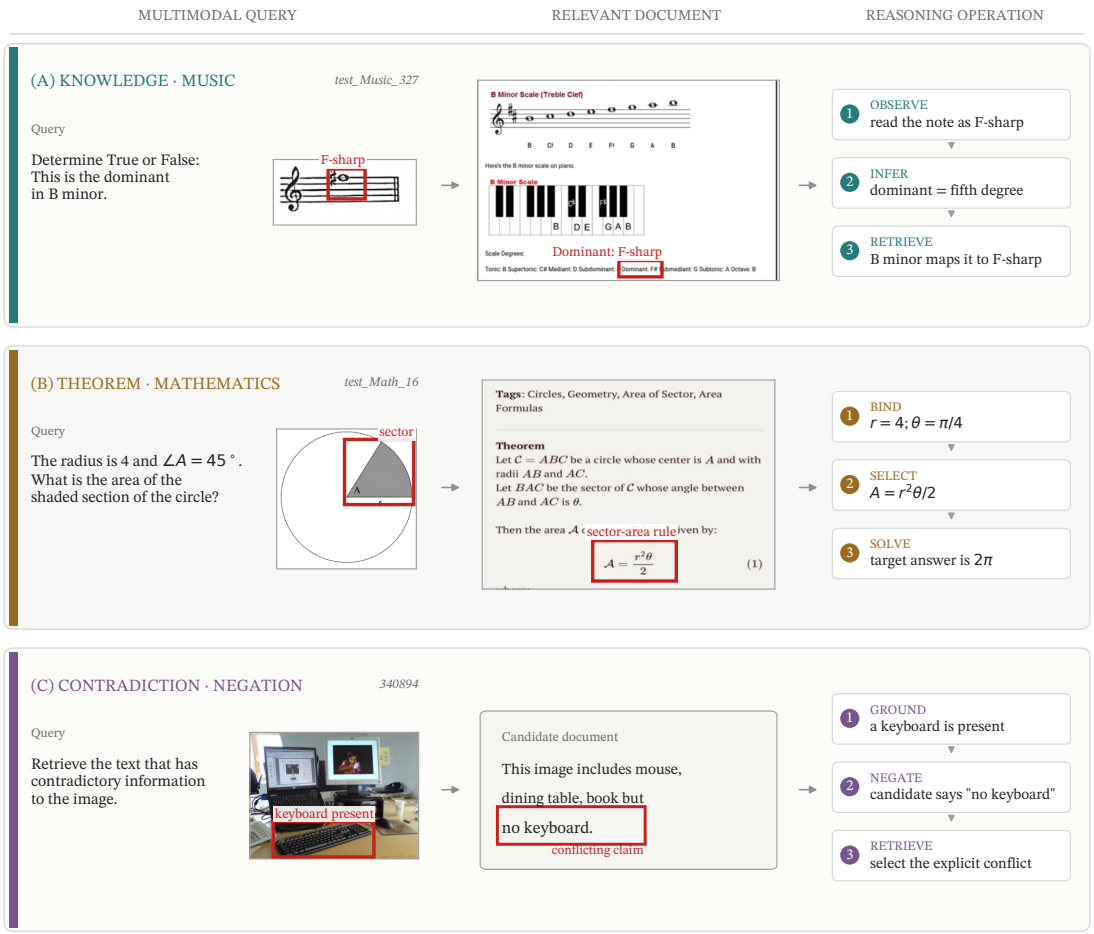


Figure 5: Three reasoning operations in MRMR. Each row links a multimodal query to its relevant document through an explicit inference; **red** boxes mark the decisive cues. Knowledge grounds F-sharp in B minor, Theorem applies the sector-area rule, and Contradiction detects the conflict between a visible keyboard and "no keyboard." Cases are from MRMR (Zhang et al., 2026b).

Figure 6 further shows that Qwen3-Embedding with image captions obtains the best overall score (54.1) and leads 9 of the 11 subtasks, indicating that language-model-mediated visual interpretation remains stronger than direct multimodal alignment on this benchmark. Even the strongest native multimodal configuration, Ops-MM-Embedding with merged images, falls from a 67.4 Knowledge mean to 37.4 on Theorem and 36.6 on Contradiction; moreover, 15 of the 16 configurations remain at or below 20 Hit@1 on Negation, below its 25% random baseline.



Figure 6: MRMR performance by task family and subtask. Results from MRMR (Zhang et al., 2026b): the matrix retains all 16 configurations and 11 subtask scores, with unweighted family means. Negation uses Hit@1 (25% chance), while the other columns use nDCG@10. Bold values mark maxima; the solid and dashed outlines identify the best overall and strongest native multimodal configurations, respectively.

❸ **Agentic, Multi-hop, and Process-Oriented Reasoning.** In complex enterprise scenarios, **M4DocBench** (Dong et al., 2025c) redefines reasoning as an iterative, agentic process. It introduces the *M4 framework* (Multi-modal, Multi-hop, Multi-document, and Multi-turn), focusing on deep research tasks where evidence is scattered across dozens of documents. Unlike single-shot retrieval, M4DocBench evaluates a system's ability to perform *Query Decomposition* and iterative refinement. It underscores that deep document intelligence requires "Strategic Planning", the ability to filter relevant documents from noisy collections and adaptively select the optimal retrieval granularity (*e.g.,* chunk, page, or summary) based on the evolving state of the research workflow.

Future Frontiers. Despite these early triumphs, several reasoning challenges remain largely unexplored. Future VDR research must address **implicit retrieval intents**, where queries involve fuzzy constraints that can only be resolved through world-knowledge synthesis (Zhang et al., 2025f; Wei et al., 2024b). Furthermore, the development of **active retrieval agents**, capable of self-correcting their search path when initial reasoning leads to dead ends, will be paramount to unlocking the full potential of multimodal document intelligence in open-domain, large-scale scenarios (Zhu et al., 2024b; Liu et al., 2025c; Zhang et al., 2025d).

3 Methodology Perspective

This section deconstructs the technical trends that underpin modern VDR, following the two axes declared in Section 1.

Section 3.1 and Section 3.2 traverse the first axis, *what a retriever is*: embedding models that score a query against a precomputed index (Table 3), and rerankers that jointly encode a query–candidate pair to trade throughput for precision. The two are separated by where interaction happens, at index time versus at query time, which is what fixes their asymmetric cost profiles. Section 3.3 then traverses the second axis, *how a retriever is deployed*, from static RAG pipelines to iterative agentic loops. This second axis presupposes the first: an agentic system is assembled *from* the components of Section 3.1 and Section 3.2 rather than competing with them, so the reader should expect composition here, not a third alternative.

3.1 Embedding Models

3.1.1 Formulation

In the context of VDR, an embedding model, denoted by an encoder $\mathcal{E}(\cdot)$, processes multimodal inputs to produce vector representations. Let $T_q = \{\tau_i^q\}_{i=1}^{n_q}$ be the token sequence for a query q and $P_d = \{p_j^d\}_{j=1}^{n_d}$ the image-patch sequence for a document page d. The objective is to learn an encoder that aligns their representations. While early models produced a single vector for each item (Zhang et al., 2024c; Liu et al., 2025e; Jiang et al., 2024c), recent VDR models adopt a multi-vector paradigm, yielding the embedding sets $\mathbf{U}(q) = \mathcal{E}(T_q)$ and $\mathbf{V}(d) = \mathcal{E}(P_d)$ introduced in Section 2.1. Their relevance score $s(q, d)$ is often computed using late interaction such as MaxSim (Faysse et al., 2024).

3.1.2 Common VDR Training Sets.

The primary distinction between specialized VDR models and general multimodal embedding models lies in their training data. VDR models are fine-tuned extensively on visual document datasets to master the fine-grained semantic and layout-aware understanding required for this domain. We discuss common VDR training sets as follows.

❶ **colpali-train-set.** This dataset serves as the training data for the ColPali model[4] and represents a hybrid approach to data collection (Faysse et al., 2024). It comprises approximately 127,000 query-image pairs, strategically combining established academic benchmarks (63%), such as DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TAT-DQA (Zhu et al., 2022) and arXivQA (Li et al., 2024a), with a custom synthetic dataset (37%). The synthetic component was created from a diverse collection of web-crawled PDF documents, with pseudo-questions generated by Claude-3 Sonnet. The dataset is intentionally English-only, designed to facilitate research into zero-shot cross-lingual generalization capabilities of VDR models.

❷ **VisRAG-Ret-Train-Synthetic-data.** As the synthetic training component for the VisRAG model[5] (Yu et al., 2024; Sun et al., 2025c), this dataset is composed entirely of VLM-generated data, totaling around 239k query-document pairs. The corpus was constructed from web-crawled PDFs spanning diverse domains, including college-level textbooks[6], academic papers from premier conferences like ICML'23 and NeurIPS'23, and product manuals[7]. A powerful VLM, GPT-4o, was leveraged to generate pseudo-queries for these document pages, creating a large-scale resource tailored for training retrieval models on a variety of document layouts and topics.

❸ **vdr-multilingual-train.** This dataset[8] marks a significant step towards multilingual VDR, containing nearly 500,000 query-image samples across five languages (English, Spanish, Italian, German, and French).

[4] https://huggingface.co/datasets/vidore/colpali_train_set
[5] https://huggingface.co/datasets/openbmb/VisRAG-Ret-Train-Synthetic-data
[6] https://openstax.org/
[7] https://www.manualslib.com/
[8] https://huggingface.co/datasets/llamaindex/vdr-multilingual-train

Category	Embedding	Publication		Model						Novelty	Resource
		Team	Venue	Params.	Backbone	Multilingual	Multi-Vec	Training-Free	Generative		
VDR	Nemotron ColEmbed V2 (Moreira et al., 2026)	NVIDIA	ECIR'26	4B/8B	Qwen3-VL/Llama3.2	en	✔	-	-	M D	
VDR	Argus-Retriever (Abdallah et al., 2026a)	Innsbruck & Chungbuk	arxiv'26	2B/4B/9B	Qwen3-VL	en	✔	-	-	M	
VDR	MM-Matryoshka (Xiang et al., 2026)	HKUST(GZ) & Alibaba	arxiv'26	2B/3B	PaliGemma/Qwen2-VL	en	✔	-	-	T E	
VDR	VaRS-Doc (Wang et al., 2026b)	SJTU	arxiv'26	3B	Qwen2.5-VL	en	✔	-	-	M	
VDR	NanoVDR (Liu et al., 2026c)	Aalto University	arxiv'26	69M/112M/151M	BERT	en	-	-	-	T E	
VDR	CausalEmbed (Huo et al., 2026)	HKUST(GZ) & Alibaba	arxiv'26	3B	Qwen2.5-VL/PaliGemma	en	✔	-	✔	T E	
VDR	ColParse (Yan et al., 2026c)	HKUST(GZ) & Alibaba	arxiv'26	2B/7B	Qwen2/2.5-VL	en	✔	✔	-	M E	
VDR	ColChunk (Yan et al., 2026a)	HKUST(GZ) & Alibaba	arxiv'26	2B/7B	Qwen2/2.5-VL	en	✔	✔	-	M E	
VDR	Ops-Colqwen3 (OpenSearch-AI, 2026)	Alibaba	blog'26	4B	Qwen3-VL	30	✔	-	-	M D	
VDR	ColPali/ColQwen (Faysse et al., 2024)	Illuin Tech	ICLR'25	3B	PaliGemma	en	✔	-	-	M D T	
VDR	Unveil (Sun et al., 2025b)	PKU	ACL'25	3B	MiniCPM-V	en	-	-	-	M	
VDR	ColMate (Masry et al., 2025)	ServiceNow	EMNLP'25	3B	PaliGemma	en/fr	✔	-	-	M	
VDR	ColFlor (Masry & Hoque, 2025)	York Uni.	arxiv'25	0.2B	Florence-2	en/fr	✔	-	-	E	
VDR	ColModernVBERT (Teiletche et al., 2025)	Illuin Tech	arxiv'25	0.3B	SigLIP	en	✔	-	-	M E	
VDR	Nemoretriever ColEmbed (Xu et al., 2025a)	NVIDIA	arxiv'25	1B/3B	Llama-3.2	18	✔	-	-	M	
VDR	jina-embeddings-v4 (Günther et al., 2025)	Jina AI	arxiv'25	4B	Qwen2.5-VL	20	✔	-	-	M T	
VDR	Tomoro-ColQwen (Huang & Tan, 2025)	Tomoro AI	blog'25	4B/8B	Qwen2.5/3	en/fr	✔	-	-	E	
VDR	ColNetraEmbed (Kolavi & Jain, 2025)	CognitiveLab	arxiv'25	4B	Gemma 3	22	✔	-	-	D	
VDR	Granite-vision-embedding (Team et al., 2025a)	IBM	arxiv'25	2B	Granite-3.1	en	-	-	-	D E	
VDR	SauerkrautLM-ColQwen (Golchinfar, 2025)	VAGO Solutions	blog'25	2B/4B/8B	Qwen3-VL	6	✔	-	-	D	
VDR	vdr-2b-multi-v1 (Cimolai & Markewich, 2025)	LlamaIndex	blog'25	2B	Qwen2VL	5	-	-	-	D E	
VDR	Eager Embed (Balarini, 2025)	Eagerworks	blog'25	4B	Qwen2.5-VL	en	-	-	-	-	
VDR	ColNomic Embed Multimodal (Nomic AI, 2025)	Nomic AI	blog'25	3B/7B	Qwen2.5-VL	5	✔	-	-	D T	
VDR	DSE (Ma et al., 2024a)	Waterloo	EMNLP'24	4B	Phi-3-Vision	en	-	-	-	D	
Gen.	MetaEmbed (Xiao et al., 2025d)	Meta	ICLR'26	3-32B	Qwen2.5/Llama-3.2	en/fr	✔	-	-	M E	
Gen.	UME-R1 (Lan et al., 2025b)	Tencent	ICLR'26	2B/7B	Qwen2-VL	en	-	-	✔	M T	
Gen.	TTE (Cui et al., 2025b)	Meta	ICLR'26	2B/7B	Qwen2-VL	en/fr	-	-	✔	M T	
Gen.	U-MARVEL (Li et al., 2025d)	NJU	ICLR'26	7B	Qwen2-VL	en	-	-	-	M T	
Gen.	VLM2Vec-V2 (Meng et al., 2025)	Salesforce	TMLR'26	2B	Qwen2-VL	en/fr	-	-	-	D	
Gen.	Evo-Retriever (Li et al., 2026c)	Alibaba	CVPR'26	3B/8B	Qwen2.5-VL	en	✔	-	-	T	
Gen.	MuCo (Gu et al., 2026)	NAVER AI	CVPR'26	2B/7B	Qwen2-VL	en	-	-	-	T	
Gen.	UniME-V2 (Gu et al., 2025b)	MiroMind	AAAI'26	2B/7B	Qwen2-VL/LLaVA	en	-	-	-	D T	
Gen.	Qwen3-VL-Embedding (Li et al., 2026a)	Alibaba	arxiv'26	2B/8B	Qwen3-VL	30	-	-	-	D T	
Gen.	V-Retrver (Chen et al., 2026a)	Tsinghua	arxiv'26	7B	Qwen2.5-VL	en	-	-	✔	M T	
Gen.	TTE-v2 (Cui et al., 2025a)	Meta	arxiv'26	2B/7B	Qwen2-VL	en	-	-	✔	M	
Gen.	Embed-RL (Jiang et al., 2026)	Tsinghua & Kuaishou	arxiv'26	2B/4B	Qwen3-VL	en	-	-	✔	M T	
Gen.	PLUME (He et al., 2026)	SEU & CAS	arxiv'26	2B	Qwen2-VL	en	-	-	✔	M T	
Gen.	RAR (Zhang et al., 2026a)	Wisconsin-Madison	arxiv'26	8B	Qwen3-VL	30	-	-	✔	M D	
Gen.	TRACE (Hao et al., 2026)	CAS	arxiv'26	7B	Qwen2.5-VL	en	-	-	✔	M D	
Gen.	MMEmb-R1 (Wang et al., 2026c)	CUHK & ByteDance	arxiv'26	2B/4B/7B	Qwen2/3-VL	en	-	-	✔	M T	
Gen.	Magic-MM-Embedding (Li et al., 2026b)	Huawei	arxiv'26	7B	InternVL3	en	-	-	-	M T E	
Gen.	e5-omni (Chen et al., 2026b)	RUC	arxiv'26	3B/8B	Qwen2.5-Omni	en	-	-	-	M T	
Gen.	AGFF-EMBED (Hu et al., 2026a)	Tencent	arxiv'26	7B	Qwen2-VL	en	-	-	-	M T	
Gen.	MM-Embed (Lin et al., 2024a)	NVIDIA	ICLR'25	7B	LLaVA-NeXT	en	-	-	-	T	
Gen.	VLM2Vec (Jiang et al., 2024c)	Waterloo & Salesforce	ICLR'25	8B	Phi-3.5-V	en	-	-	-	M D	
Gen.	B3 (Thirukovalluru et al., 2025)	Duke	NeurIPS'25	2B/7B/8B	Qwen2-VL/InternVL3	en	-	-	-	D T E	
Gen.	LCO-Embedding (Xiao et al., 2025b)	Alibaba	NeurIPS'25	3B/7B	LLaVA/Qwen2.5-VL	38	-	-	-	M	
Gen.	MMRet-MLLM (Zhou et al., 2025b)	BUPT	ACL'25	7B	LLaVA-1.6	en	-	-	-	D	
Gen.	UniSE (Liu et al., 2025b)	BAAI	ACL'25	0.4/2B	CLIP/Qwen2-VL	en	-	-	-	M D	
Gen.	GME (Zhang et al., 2024c)	PolyU & Alibaba	CVPR'25	7B	Qwen2-VL	en	-	-	-	D	
Gen.	LamRA (Liu et al., 2025e)	SJTU & Xiaohongshu	CVPR'25	7B	Qwen2-VL	en	-	-	-	M	
Gen.	VladVA (Ouali et al., 2025)	Samsung	CVPR'25	7B	LLaVA-1.5	en	-	-	✔	M	
Gen.	CAFe (Yu et al., 2025a)	Meta	ICCV'25	0.5B/7B	LLaVA-OV	en	-	-	✔	T	
Gen.	UniME-V1 (Gu et al., 2025a)	Sydney	MM'25	4B/7B	Phi-3.5/LLaVA	en	-	-	-	T	
Gen.	CoMa (Li et al., 2025a)	CAS & Kuaishou	arxiv'25	3B/7B	Qwen2.5-VL	en	-	-	-	M E	
Gen.	FreeRet (Zhu et al., 2025c)	NJU	arxiv'25	2-32B	Qwen2/2.5-VL/InternVL3	en	-	-	✔	M	
Gen.	LLaVE (Lan et al., 2025a)	Tencent	arxiv'25	0.5B/7B	LLaVA/Aquila	en	-	-	-	T	
Gen.	mmE5 (Chen et al., 2025b)	RUC	arxiv'25	11B	Llama-3.2	93	-	-	-	D	
Gen.	MoCa (Chen et al., 2025a)	RUC	arxiv'25	3B/7B	Qwen2.5-VL	en/fr	-	-	-	M T	
Gen.	PDF (Wang et al., 2025e)	Alibaba	arxiv'25	2B/7B	LLaVA/Qwen2-VL	en	-	-	-	M T	
Gen.	QQMM (Xue et al., 2025)	Tencent	arxiv'25	7B	LLaVA/Qwen2-VL	en	-	-	-	T	
Gen.	ReMatch (Liu et al., 2025d)	Xiaohongshu	arxiv'25	2B/7B	Qwen2/2.5-VL	en	-	-	-	M	
Gen.	RGE (Liu et al., 2025a)	SenseTime	arxiv'25	3B	Qwen2.5-VL	en	-	-	✔	M T	
Gen.	RzenEmbed (Jian et al., 2025)	360 AI	arxiv'25	2B/8B	Qwen2-VL	en	-	-	-	D T	
Gen.	Unite (Kong et al., 2025)	Kuaishou	arxiv'25	2B/7B	Qwen2-VL	en	-	-	-	D T	
Gen.	VIRTUE (Wang et al., 2025d)	Sony	arxiv'25	2B/7B	Qwen2-VL	en	-	-	-	M D	
Gen.	xVLM2Vec (Musacchio et al., 2025)	UBAM	arxiv'25	4B	Phi-3.5-V	5	-	-	-	D T	
Gen.	SAIL-Embedding (Lin et al., 2025a)	ByteDance	arxiv'25	-	SAIL-VL	en	-	-	-	D T	
Gen.	Ops-MM-embedding (OpenSearch-AI, 2025)	Alibaba	blog'25	2B/7B	Qwen2-VL	en	-	-	-	D	
Gen.	EvoQwen2.5-VL-Retriever (ApsaraStackMaaS, 2025)	Alibaba	blog'25	3B/7B	Qwen2.5-VL	en	✔	-	-	D	
Gen.	Seed1.6-Embedding (ByteDance, 2025)	ByteDance	blog'25	-	Seed1.6-flash	en	-	-	-	D	
Gen.	PreFLMR (Lin et al., 2024b)	Cambridge	ACL'24	2B	BERT	en	✔	-	-	M D	
Gen.	Vista (Zhou et al., 2024)	BUPT	ACL'24	0.2B	BGE-v1.5	en	-	-	-	M D	
Gen.	UniIR (Wei et al., 2024a)	Waterloo	ECCV'24	0.4B	CLIP	en	-	-	-	D	
Gen.	M-Solomon (Kim et al., 2025b)	NC AI	CIKM'24	7B	Qwen2-VL	en	-	-	-	M	
Gen.	Nomic Embed Vision (Nussbaum et al., 2024)	Nomic AI	arxiv'24	0.2B	BERT	en	-	-	-	-	
Gen.	E5-V (Jiang et al., 2024b)	Beihang	arxiv'24	8B	LLaVA-NeXT	en	-	-	-	M T	

Table 3: Comparison of VDR and general multimodal embedding models. In the *Novelty* column, we denote Model-/Data-/Training-/Efficiency-level contributions as M / D / T / E. In the *Resource* column, we denote the corresponding GitHub codebase, Hugging Face repository, and paper (*e.g.,* arXiv paper, technical report, or blog) as / / .

Its construction involved a highly sophisticated pipeline. First, a diverse corpus of ~50k documents was scraped from the internet using topic-based search queries for each language. A key innovation was the use of layout analysis to sample pages, ensuring an even distribution of text-only, visual-only, and mixed-modality pages. Synthetic queries were then generated using powerful VLMs (Gemini-1.5-Pro and Qwen2-VL-72B) with an advanced prompting technique that distinguished between general and specific questions to improve query quality. The dataset underwent rigorous cleaning, filtering, and hard-negative mining to enhance its utility for training robust retrieval models.

❹ **VDR_MEGA_MultiDomain_DocRetrieval.** This dataset[9] represents the largest and most comprehensive resource to date, amalgamating approximately 1.09 million examples across five languages. It functions as a meta-dataset, strategically fusing the three previously mentioned datasets (`colpali-train-set`, `VisRAG-Ret-Train-Synthetic-data`, and `vdr-multilingual-train`) with several new, domain-specific collections. These additions cover specialized fields such as military[10], energy[11], hydrogen technology[12], and geotechnical engineering[13]. By unifying multiple datasets, it provides unparalleled scale and diversity in both language and subject matter, making it an ideal resource for training highly generalized and robust VDR systems.

❺ **docmatix.** This is an interleaved multimodal pre-training dataset[14] created for the modality-aware continual pre-training of MoCa models (Chen et al., 2025a). It is adapted from the original Docmatix dataset (Laurençon et al., 2024), a massive-scale Document Visual Question Answering resource containing approximately 2.4 million images and 9.5 million question-answer pairs, which was initially used to fine-tune the Idefics3 model. The adaptation process involves transforming the original question-answer pairs by concatenating document screenshots with their corresponding texts, thereby creating an interleaved format suitable for continuous pre-training.

Common Characteristics. The current generation of VDR training sets reveals several unifying trends. Firstly, there is a clear paradigm shift towards **leveraging synthetic data generation at scale**. All major datasets heavily rely on powerful VLMs to create pseudo-queries for large, unannotated document corpora, effectively bypassing the bottleneck of manual annotation. Secondly, these corpora are primarily built from **web-crawled PDF documents**, which provides a rich diversity of layouts, domains, and styles that mirror real-world scenarios. Thirdly, there is a growing emphasis on **sophisticated data curation**, including techniques like layout-aware page sampling and automated hard-negative mining, to improve training efficiency and model performance. Finally, a clear trajectory exists towards **increasing scale and multilingualism**, evolving from English-only datasets to massive, multi-language compilations that enable the development of globally competent models.

Future Directions. Looking ahead, the optimization of VDR training sets can be advanced in several key directions. The most critical frontier is moving **beyond simple semantic matching towards reasoning-intensive data**. Future datasets should include query-document pairs that necessitate multi-hop, logical, or causal reasoning to find the correct answer, mirroring the challenges posed by benchmarks like MRMR and M4DocBench. Secondly, there is a need for **enhanced data authenticity and complexity**. While VLM-generated queries are scalable, they can lack the nuance and "messiness" of real user intent. Future work could explore mining queries from anonymized user logs or using agentic workflows to simulate more realistic information-seeking behaviors. Lastly, training sets could benefit from **finer-grained structural annotations**. Instead of just matching a query to a page, future datasets could provide explicit links to specific sub-page elements (*e.g.,* a single row in a table, a data point in a chart, or a specific paragraph), which would be invaluable for training models that can perform precise, element-level evidence retrieval.

3.1.3 Technical Trends

Model Choices. The field has witnessed a clear trend in both model scale and architecture. While early models were based on smaller backbones like BERT (Nussbaum et al., 2024), current VDR embeddings are predominantly built on powerful MLLMs, such as PaliGemma (Beyer et al., 2024) and Qwen-VL series (Bai et al., 2025). Parameter counts have also escalated, with SOTA models typically ranging from 2 to 8 billion parameters, a significant increase that enables more nuanced comprehension of complex document structures.

[9] https://huggingface.co/datasets/racineai/VDR_MEGA_MultiDomain_DocRetrieval
[10] https://huggingface.co/datasets/racineai/VDR_Military
[11] https://huggingface.co/datasets/racineai/VDR_Energy
[12] https://huggingface.co/datasets/racineai/VDR_Hydrogen
[13] https://huggingface.co/datasets/racineai/VDR_Geotechnie
[14] https://huggingface.co/datasets/moca-embed/docmatix

Multilingual Support. As document intelligence applications become increasingly global, multilingual capability has emerged as a critical focus in VDR. Unlike general-domain image retrieval where this aspect is less emphasized, the prevalence of multilingual business reports, academic papers, and official forms necessitates cross-lingual understanding. Recent models like jina-embeddings-v4 (Günther et al., 2025) and Nemoretriever (Xu et al., 2025a) reflect this shift, offering support for over 20 languages and demonstrating strong performance on multilingual VDR benchmarks such as Jina-VDR (Günther et al., 2025) and Nayana-IR (Kolavi & Jain, 2025). Complementing these broad-coverage models, a language-specialised strand is emerging: KoVRE (Choi et al., 2026a) trains a compact Korean VDR encoder, and is released alongside the KoViDoRe benchmark (Choi et al., 2026b), an instance of the co-design of model and evaluation that Section 3.2 argues is precisely what multilingual reranking still lacks.

Multi-Vector Representation. The multi-vector paradigm, popularized in the VDR domain by ColPali (Faysse et al., 2024), has become a dominant approach for fine-grained retrieval (Li et al., 2025b; Yang et al., 2026; Chaffin et al., 2026; Zhu et al., 2026), as shown in Figure 7. Instead of compressing an entire document or query into a single embedding, this method generates a set of token-level or patch-level embeddings. This granular representation is particularly effective for VDR because it enables "late interaction" matching (Khattab & Zaharia, 2020), where specific phrases in a query can be precisely aligned with corresponding visual or textual regions in a document, thus preserving local details often lost in single-vector representations. ColPali (Faysse et al., 2024) adapts PaliGemma to produce multi-vector outputs by treating the embeddings of individual image patches as distinct vectors for late interaction. PreFLMR (Lin et al., 2024b) generates its multi-vector representations by concatenating embeddings from text tokens with both global and patch-level visual features, which are further refined through cross-attention to be query-aware. Addressing efficiency, ColModernVBERT (Teiletche et al., 2025) demonstrates that a compact, bidirectional language encoder can be effectively aligned with a vision encoder to generate granular embeddings; while MetaEmbed (Xiao et al., 2025d) introduces a fixed set of learnable "Meta Tokens" whose final hidden states serve as a compact and scalable multi-vector representation, enabling flexible trade-offs between retrieval quality and efficiency at test-time.

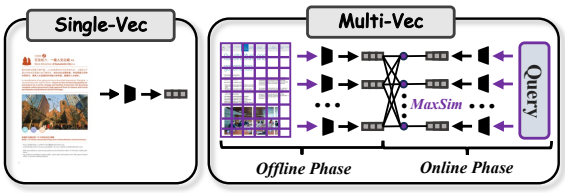


Figure 7: Single-vec (*left*) and multi-vec (*right*) VDR.

The empirical case for this paradigm is now unusually clear-cut, and Figure 8 makes it concrete. Of the top-20 systems on the public ViDoRe-v3 leaderboard, **18 are multi-vector late-interaction models**; the only single-vector entries are the two Qwen3-VL-Embedding checkpoints, and the stronger of them ranks only 14th. Three readings of the figure are worth drawing out. ❶ *The gap is not a scale artefact.* The best dense model (Qwen3-VL-Embedding-8B, 8.14B parameters) reaches 57.78, while a late-interaction model at *fewer* parameters (Nemotron ColEmbed-VL-4B-v2, 4.80B) reaches 61.54, a 3.76-point advantage at roughly 40% fewer parameters. ❷ *Late interaction buys parameter efficiency.* VultronRetriever-Flash at 0.85B lands within 1.6 points of the 8.14B dense model, a nearly ten-fold parameter reduction, indicating that fine-grained matching recovers signal that a single pooled vector discards regardless of encoder capacity. ❸ *Within-family scaling is real but secondary.* The dashed trajectories climb monotonically, yet their slopes are visibly shallower than the vertical offset between the dense and multi-vector frontiers: choosing the interaction mechanism matters more than adding parameters. We caution that these are leaderboard-reported numbers under each submitter's own inference configuration rather than a controlled study, so the magnitudes should be read as indicative; the ordering, however, is stable across the public and private splits, as panels (b) and (c) of Figure 8 confirm.

The split-level view. Separating the two splits matters for two reasons. First, it is a contamination check: the public split overlaps corpora that several submitters may have trained on, whereas the private split does not. Second, it tests whether the multi-vector advantage just claimed is an artefact of the public data. It is not, since the ordering of the dense and multi-vector frontiers is preserved on the private split, where the best single-vector model reaches 55.73 against 62.04 for a smaller late-interaction model. The comparison does, however, expose a per-model robustness signal that the aggregate hides. Most systems lose only a fraction

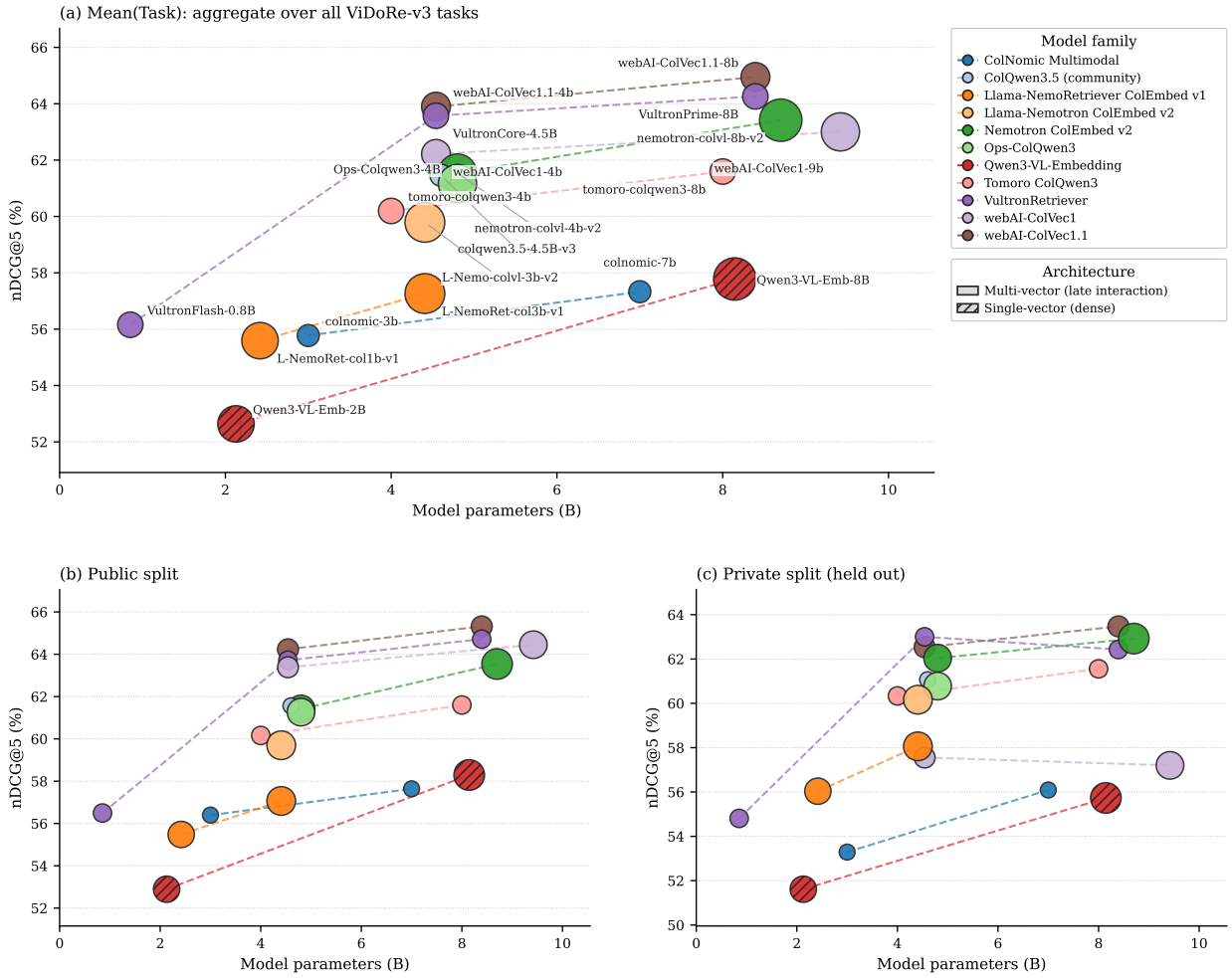


Figure 8: Why late interaction became the default VDR architecture. Parameter count against retrieval quality for the top-20 entries of the public ViDoRe-v3 leaderboard, using the latest results available as of 25 Aug, 2026. Panel (a) aggregates all tasks; panels (b) and (c) repeat the analysis on the public and held-out private splits, and share the legends of panel (a). Colour groups a model family, with dashed lines joining its released scales; bubble area is proportional to embedding dimension, the per-vector cost that late interaction multiplies by the number of retained patches; hatching marks the only single-vector (dense) models in the top-20. The dense frontier sits below the multi-vector frontier at every scale in all three panels, and the gap does not close as parameters grow.

of a point moving from public to private, but a few degrade sharply: webAI-ColVec1-9b falls from 64.45 to 57.20, a 7.25-point drop, while its successor webAI-ColVec1.1-8b loses only 1.85 points. Such asymmetric drops are a useful diagnostic for generalisation, and we suggest that future VDR leaderboards report the public–private delta as a first-class metric rather than a derived one. This gap is precisely what motivates the compression literature of Section 4.3: late interaction is winning on quality while carrying a storage cost that single-vector models do not pay.

A parallel line of work attacks the same limitation from the supervision side rather than the architecture side: Tilli & Mesgar (2026) argue that patch-level late interaction degenerates into a "bag of patches" that discards global page layout, and recover that structure by distilling a textual layout descriptor into the retriever during training. Hybrid designs go further and reintroduce explicit lexical signal alongside the dense one: DocRetriever (Hu et al., 2026c) pairs a layout-aware *sparse* embedding with a generalisable few-shot reranker, retaining OCR-free operation while restoring the exact-match behaviour that dense vectors blur.

More recently, ColParse (Yan et al., 2026c) introduces a layout-informed paradigm that leverages document parsing models to extract a compact set of semantically meaningful sub-images. By fusing these local layout-aware embeddings with a global page-level vector, ColParse achieves superior retrieval performance while significantly reducing storage requirements. Meanwhile, CausalEmbed (Huo et al., 2026) reframes the problem through an auto-regressive generative lens. It treats multi-vector embedding construction as a sequential latent space generation task, enabling flexible test-time scaling, where retrieval precision can be dynamically adjusted by modifying the token budget at inference, thereby achieving high-fidelity representations with only a fraction of the conventional token count.

Training Paradigm Exploration. The predominant training method for VDR embedding models is end-to-end supervised fine-tuning using a contrastive loss, such as InfoNCE (Oord et al., 2018). This objective pulls positive query-document pairs closer together in the embedding space while pushing negative pairs apart. Beyond this standard approach, two novel paradigms are gaining traction.

The first paradigm reduces the dependence on annotated *multimodal* supervision. Here it is worth separating two claims that are frequently conflated, since they imply different deployment costs.

E5-V (Jiang et al., 2024b) is best described as *single-modality training*, not as a training-free method. Its mechanism is prompt-based: carefully designed prompts drive the MLLM to compress either an image or a sentence into one representation, which bridges the modality gap and, as the authors report, already yields strong embeddings *even without fine-tuning*. Its actual proposal, however, is to then train the model **exclusively on text pairs**. This is what delivers the headline result: outperforming conventional training on image–text pairs while cutting training cost by roughly 95% and removing the need to collect multimodal training data altogether. The lesson generalises beyond E5-V: because the prompt already aligns the modalities, the supervision needed to sharpen the representation need not itself be multimodal.

Genuinely training-free retrieval is a stronger and rarer proposition, exemplified by FreeRet (Zhu et al., 2025c): a plug-and-play framework that drives off-the-shelf MLLMs for both embedding-based search and multiple-choice reranking with *no* parameter updates at all. The practical distinction is that single-modality training still requires a fine-tuning budget and a text corpus but yields a fixed, cheap-to-serve encoder, whereas a training-free pipeline eliminates that budget entirely and pays for it with higher inference-time compute.

The second paradigm explores how to harness the **generative capabilities** of MLLMs to enhance retrieval. Early works such as CAFe (Yu et al., 2025a) and VladVA (Ouali et al., 2025) introduced hybrid training frameworks; they jointly optimize a contrastive loss for discriminative power with an autoregressive, next-token prediction loss to preserve and leverage the model's inherent generative abilities. More recent approaches explicitly integrate reasoning as a precursor to embedding. For example, RGE (Liu et al., 2025a) and TTE (Cui et al., 2025b) both propose a "think-then-embed" process, where the model first generates an explicit rationale or reasoning trace, and the final embedding is conditioned on this generated context to capture more nuanced semantics. Moving one step beyond embedding construction, Retrv-R1 (Zhu et al., 2025b) employs reinforcement learning to optimize a step-by-step reasoning process that directly selects among candidates; we therefore treat it as an adjacent generative-retrieval formulation rather than include it in the embedding-model taxonomy below. Similarly, TTE-v2 (Cui et al., 2025a) extends this paradigm with a cascaded retrieval-reranking framework, using joint query-candidate reasoning to refine the initial results and leveraging the reranker as a teacher for hard negative mining. Embed-RL (Jiang et al., 2026) proposes an Embedder-Guided Reinforcement Learning (EG-RL) framework to optimize a reasoner to produce evidential traceability CoTs that better align with the embedding task. MMEmb-R1 (Wang et al., 2026c) introduces pair-aware reasoning selection to identify beneficial rationales and uses reinforcement learning to adaptively invoke reasoning only when necessary. In a different direction, PLUME (He et al., 2026) internalizes reasoning into a compact latent rollout of continuous hidden states, replacing expensive explicit CoT generation while preserving the benefits of multi-step computation for efficiency.

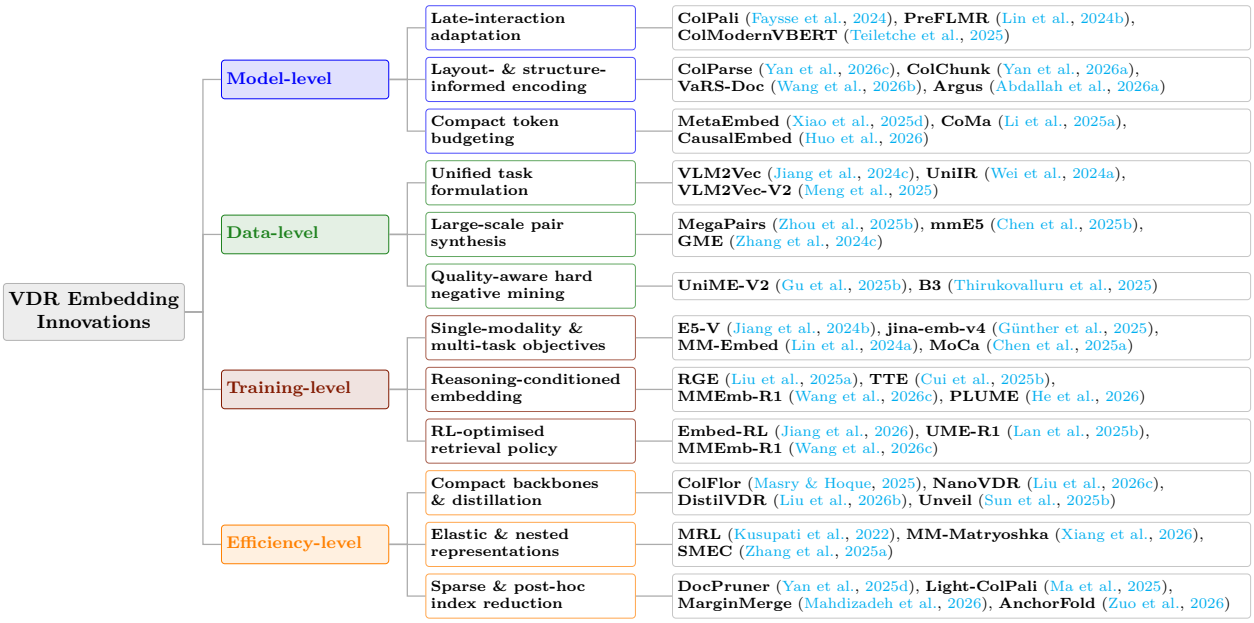


Figure 9: Taxonomy of VDR embedding-model innovations. Contributions are grouped by *where* they intervene (**model**, **data**, **training**, or **efficiency**), matching the M / D / T / E annotations in the *Novelty* column of Table 3. Each level is refined into the strategies the field has actually converged on, with representative work at the leaves. This axis is orthogonal to the representation axis of Section 3.1: it classifies *contribution type*, so a model may legitimately appear under more than one level.

3.1.4 Technical Innovations

Having placed embedding models on the representation axis, we now ask a different question: *where does a given paper actually intervene?* Two models may share the late-interaction mechanism yet contribute nothing comparable: one may add an architectural component, another only a data recipe. We therefore group contributions into four levels (**model**, **data**, **training** and **efficiency**) and subdivide each into the concrete strategies the literature has converged on. Figure 9 lays out the resulting taxonomy with representative work at the leaves; these are the same four levels annotated per row in Table 3, so the figure and the table are two views of one scheme. We stress that this is a taxonomy of *contribution type*, not a third representation axis: a single model can and often does appear under several levels, and the assignment below reflects each paper's primary claim.

❶ **Model-level—changing the inference-time representation path.** This level covers changes to the encoder, the units it emits, or the mechanism by which query and page interact; the defining test is that the model computes a different representation even if its data and loss are held fixed. *Late-interaction adaptation* first made visual patches directly retrievable: ColPali (Faysse et al., 2024) projects PaliGemma patch states and applies token-to-patch MaxSim, preserving local evidence without an OCR bottleneck. *Layout- and structure-informed encoding* instead changes the unit being matched. ColParse (Yan et al., 2026c), for example, replaces an undifferentiated patch grid with parsed semantic sub-images and combines their local vectors with a global page vector, so sections, tables and figures remain distinct retrieval units. *Compact token budgeting* controls how many such units reach the index: MetaEmbed (Xiao et al., 2025d) summarizes a page into a fixed set of learnable meta tokens, whereas CausalEmbed (Huo et al., 2026) autoregressively produces a variable number of latent vectors, turning representation fidelity into a test-time token-budget decision.

❷ **Data-level—changing which relevance relations are learned.** Data-level innovations leave the serving architecture largely intact but reshape the tasks, positive pairs and confusing negatives from which its geometry is learned. Under *unified task formulation*, VLM2Vec (Jiang et al., 2024c) casts heterogeneous

Reranker	Publication		Model				Benchmark*	Resource
	Team	Venue	Para.	Backbone	Ranking	Multilingual		
ZipRerank (Sun et al., 2026)	Magellan Tech. RI	ICML'26	8B	Qwen3-VL	LI	en	MMDocIR	
miniReranker (Fan et al., 2026)	EIT	arxiv'26	2B	Qwen3-VL	PO	en	MMDocIR	
Lychee-rerank-mm (Dai et al., 2025)	HIT	ICLR'26	7B	Qwen2.5-VL	PO	en	MRB/MRMR	
UniME-V2-Reranker (Gu et al., 2025b)	MiroMind AI	AAAI'26	7B	Qwen2.5-VL	PA LI	en	MMEB	
Qwen3-VL-Reranker (Li et al., 2026a)	Alibaba	arxiv'26	2B/8B	Qwen3-VL	PO	>30	MMEB-v2/JinaVDR/ViDoRe-v3	
llama-nemotron-rerank-vl (Moreira et al., 2026)	NVIDIA	ECIR'26	3B	Llama 3.2	LI	en	MIRACL-Vision/ViDoRe-v1/v2/v3	
Rank-Nexus (Cai, 2026)	Malaya	arxiv'26	2B	Qwen3-VL/InternVL3	LI	en	MMDocIR	
LamRA-Rank (Liu et al., 2025e)	SJTU & Xiaohongshu	CVPR'25	7B	Qwen2.5-VL	PO LI	en	M-BEIR	
DocReRank (Wasserman et al., 2025a)	WIS	EMNLP'25	2B	Qwen2-VL	PO	en	ViDoRe-v2/Real-MM-RAG	
RagVL (Chen et al., 2024)	IDEA	EMNLP Findings'25	1B/2B/4B/13B	Qwen-VL/InternVL/LLaVA-v1.5	PO	en	-	
MM-R5 (Xu et al., 2025b)	DP Tech	arxiv'25	7B	Qwen2.5-VL	LI	en	MMDocIR	
jina-reranker-m0 (Jina AI, 2025)	Jina AI	blog'25	2.4B	Qwen2-VL	PO	29	ViDoRe-v1/M-BEIR	
MonoQwen2-VL-v0.1 (Chaffin & Lac, 2024)	LightOn	blog'24	2B	Qwen2-VL	PO	en	ViDoRe-v1	

Table 4: Comparison of multimodal document rerankers. * indicates evaluation only on VDR-related benchmarks. In the *Ranking* column, we denote POintwise/PAirwise/LIstwise designs as PO / PA / LI. In the *Resource* column, we denote the corresponding GitHub codebase, Hugging Face repository, and paper (*e.g.,* arXiv paper, technical report, or blog) as / / .

multimodal tasks as instruction-conditioned retrieval in MMEB, while UniIR (Wei et al., 2024a) places image–text and cross-modal search in one retrieval interface; both broaden transfer by making formerly separate tasks share a ranking objective. *Large-scale pair synthesis* expands coverage beyond manually labelled corpora: MegaPairs (Zhou et al., 2025b) mines heterogeneous visual-semantic and visual-pattern relations before synthesising query–candidate pairs, so scale is coupled to relationship diversity rather than simple duplication. Finally, *quality-aware hard-negative mining* determines which near-misses should shape the boundary. UniME-V2 (Gu et al., 2025b) uses an MLLM judge to assign graded semantic supervision to retrieved candidates, whereas B3 (Thirukovalluru et al., 2025) builds a graph over the corpus and constructs batches rich in mutual hard negatives; the former improves label quality and the latter improves negative density, so they belong to the same data-level scope but are not the same mechanism.

❸ **Training-level—changing how supervision is converted into an encoder.** This level concerns objectives, schedules and optimisation policies rather than the examples themselves. *Single-modality and multi-task objectives* relax the standard one-loss/one-output recipe: E5-V (Jiang et al., 2024b) uses prompting to bridge modalities and then fine-tunes only on text pairs, while jina-embeddings-v4 (Günther et al., 2025) jointly learns single-vector and late-interaction outputs with task-specific adapters. *Reasoning-conditioned embedding* inserts computation before compression: RGE (Liu et al., 2025a) generates an explicit rationale and conditions the embedding on it, whereas PLUME (He et al., 2026) replaces readable chains of thought with compact continuous latent rollouts. *RL-optimised embedding policies* supervise the reasoning or control decision rather than only the final similarity: Embed-RL (Jiang et al., 2026) rewards evidential traces that improve a downstream embedder, UME-R1 (Lan et al., 2025b) optimizes reasoning-driven embedding generation, and MMEmb-R1 (Wang et al., 2026c) learns when reasoning is beneficial. These methods therefore remain within the embedding scope while optimizing when and how retrieval reasoning is performed.

❹ **Efficiency-level—changing the deployable cost envelope.** Efficiency-level work treats model depth, vector width and vector count as explicit resources, seeking lower indexing or query-time cost without redefining relevance. *Compact backbones and distillation* reduce encoder cost: NanoVDR (Liu et al., 2026c) distils a 2B visual retriever into a roughly 70M text encoder, while DistilVDR (Liu et al., 2026b) uses dual students to retain end-to-end visual retrieval in a smaller model. *Elastic and nested representations* train one checkpoint for several budgets; MM-Matryoshka (Xiang et al., 2026) nests both vector prefixes and intermediate encoder layers, whereas SMEC (Zhang et al., 2025a) stabilises dimension-wise Matryoshka learning through a sequential schedule. *Sparse and post-hoc index reduction* operates after encoding: DocPruner (Yan et al., 2025d) removes low-importance patch vectors, while AnchorFold (Zuo et al., 2026) selects attention-central anchors and folds non-anchor evidence into them. The last distinction matters operationally: pruning saves count by discarding vectors, whereas folding tries to preserve their contribution inside a smaller index.

3.2 Reranker Models

Table 4 summarizes the architectures, ranking paradigms and language coverage of representative multimodal document rerankers before we examine their formulation and emerging trade-offs.

3.2.1 Formulation

A reranker model, denoted by $R(\cdot,\cdot)$, performs query-conditioned joint encoding after first-stage retrieval. At the basic pointwise granularity, it takes a query q and a single candidate document d_c as a combined input. By jointly processing the query-token sequence $T_q = \{\tau_i^q\}_{i=1}^{n_q}$ and document-patch sequence $P_{d_c} = \{p_j^{d_c}\}_{j=1}^{n_{d_c}}$, the model can leverage deep cross-attention to capture fine-grained inter-modal dependencies, yielding the scalar score $\rho(q, d_c) = R(T_q, P_{d_c})$. Pairwise and listwise variants extend the jointly encoded input from one candidate to a pair or an entire candidate list, as detailed in Section 3.2.3. This deep interaction allows the model to make more accurate relevance judgments than the dot-product-based similarity used in a bi-encoder.

3.2.2 Technical Trends

Model Choices. Similar to embedding models, the trend in rerankers is towards larger, MLLM-based architectures. Models such as UniME-V2-Reranker (Gu et al., 2025b) and LamRA-Rank (Liu et al., 2025e) utilize powerful Qwen2.5-VL with parameter counts reaching 7 billion. The strong multimodal understanding capabilities of MLLMs make them suited for reranking.

Multilingual Support. In stark contrast to the rapid adoption of multilingualism in VDR embedding models, this capability remains largely underdeveloped in the reranker space. Of the eleven rerankers in Table 4, only two report broad language coverage, namely jina-reranker-m0 (Jina AI, 2025) with 29 languages and Qwen3-VL-Reranker (Li et al., 2026a) with over 30, while the remainder, including DocReRank (Wasserman et al., 2025a), MM-R5 (Xu et al., 2025b) and MonoQwen2-VL-v0.1 (Chaffin & Lac, 2024), are English-only. On the embedding side of Table 3, by contrast, multilingual coverage is routine: Nemoretriever ColEmbed spans 18 languages (Xu et al., 2025a), jina-embeddings-v4 20 (Günther et al., 2025), ColNetraEmbed 22 (Kolavi & Jain, 2025), and mmE5 93 (Chen et al., 2025b).

Why the asymmetry? Prior work notes the gap but rarely asks what sustains it. We attribute it to four structural causes rather than to a simple lack of effort, and each suggests a different remedy.

❶ *Supervision cost scales differently.* A multilingual embedder can be trained from (query, positive) pairs alone, with in-batch sampling supplying negatives for free; such pairs are cheap to mine or synthesise per language, which is exactly how the datasets of Section 3.1 were built. A reranker needs *graded, comparative* judgements (which of these k candidates is more relevant), and for listwise objectives an ordering over the whole list. Annotation cost therefore grows with languages $\times$ candidates rather than with languages alone, and the resulting labels are far harder to synthesise reliably.

❷ *Inference cost multiplies the research overhead.* As formalised in Section 3.2, a cross-encoder requires a forward pass per query–candidate pair, whereas an embedder amortises document encoding into a reusable index. Every additional language thus multiplies not just training but also validation and ablation cost. Faced with a fixed budget, the higher marginal return lies in improving a multilingual first-stage retriever, since first-stage recall upper-bounds what any reranker can recover.

❸ *Cross-lingual alignment is harder in pixel space.* An embedder can inherit cross-lingual transfer from the backbone's shared semantic space under a single contrastive objective. A document reranker must jointly (i) read text rendered as pixels, possibly in a script different from the query's, and (ii) judge relevance. Both must succeed for the score to be meaningful, so errors compound multiplicatively; and step (i) degrades precisely on low-resource and non-Latin scripts (connected Arabic forms, Indic conjuncts, dense CJK glyphs at low resolution), where the backbone's visual text recognition is weakest. This also explains why the two multilingual exceptions are built on backbones with strong multilingual pre-training rather than on document-specific architectures.

❹ *A missing-evaluation feedback loop.* The multilingual benchmarks that do exist, namely Jina-VDR (Günther et al., 2025), Nayana-IR (Kolavi & Jain, 2025), MIRACL-VISION (Osmulski et al., 2025) and SDS KoPub VDR (Lee et al., 2025a), report first-stage retrieval metrics and define no standard reranking protocol over their candidate pools. Without a leaderboard slot there is little incentive to train multilingual rerankers, and without such models there is little pressure to add the protocol: the gap is self-reinforcing. The cheapest

intervention is consequently not a new model but a reranking track over existing multilingual candidate pools, which would convert a latent gap into a measurable one.

3.2.3 Reranking Paradigms

Reranker models are typically trained using one or a combination of three main paradigms:

❶ **Pointwise:** This approach evaluates each query-document pair (q, d_i) independently, predicting an absolute relevance score. The model is trained using a loss function like Binary Cross-Entropy (BCE) on the predicted score $\rho_i = \rho(q, d_i)$ against a ground-truth label $y_i \in \{0, 1\}$. It can be formulated as $\mathcal{L}_{\text{pointwise}} = \sum_i \text{BCE}(\rho_i, y_i)$.

❷ **Pairwise:** This method learns relative relevance by comparing pairs. Given a query q and a pair of documents (d_i, d_j) where d_i is more relevant than d_j, the model is trained with a margin $m > 0$ to ensure $\rho(q, d_i) > \rho(q, d_j)$, formulated as $\mathcal{L}_{\text{pairwise}} = \sum_{(i,j):\, y_i > y_j} \max(0, m - (\rho_i - \rho_j))$.

❸ **Listwise:** This paradigm considers the entire list of candidate documents for a query simultaneously. It optimizes a loss function, such as ListNet's softmax cross-entropy, that directly corresponds to ranking metrics like nDCG, learning to predict the optimal ordering of the full list, formulated as $\mathcal{L}_{\text{listwise}} = -\sum_i \eta_i \log \hat{\eta}_i$, where η_i and $\hat{\eta}_i$ are the ground-truth and predicted probabilities assigned to candidate d_i.

Recent rerankers have shown that combining multiple loss functions often yields superior performance. LamRA-Rank (Liu et al., 2025e) and UniME-V2-Reranker (Gu et al., 2025b) adopt hybrid training strategies that combine listwise optimization, where the model predicts the correct item's position, with pointwise or pairwise objectives that classify the relevance of individual or paired candidates. Other approaches refine a single paradigm to great effect; models like Lychee-rerank-mm (Dai et al., 2025) and DocReRank (Wasserman et al., 2025a) utilize supervised fine-tuning to frame reranking as a pointwise classification task, predicting "yes" or "no" to align with the generative nature of MLLMs. Besides, MM-R5 (Xu et al., 2025b) advances the listwise paradigm by incorporating CoT reasoning and leveraging reinforcement learning with a task-specific reward to optimize the ranked output.

The three paradigms are usually presented as a modelling choice, but they differ just as much in *cost*, and that is what governs which is deployable. Pointwise scoring is embarrassingly parallel, since each pair is independent and a batch of k candidates is one batched forward pass, but it must produce a *calibrated* absolute score, and calibration drifts across domains, which is why pointwise rerankers often need per-corpus threshold tuning. Listwise scoring sidesteps calibration entirely, since only the induced order matters, but it must place all k candidates in a single context window. For visual documents this is punishing: each page costs hundreds to thousands of visual tokens, so a list of ten pages can exceed the context budget outright and grows the attention cost super-linearly. Pairwise scoring sits between the two, trading $O(k^2)$ comparisons for supervision that is easier to elicit than a full ordering.

This cost asymmetry explains the direction of recent work. ZipRerank (Sun et al., 2026) keeps the listwise formulation but compresses the visual evidence entering the context; Figure 10 shows that it reaches 89.0/89.3 macro Recall@5 on the $\text{DSE}_{\text{wiki}-\text{ss}}$/ColQwen candidate pools, 5.2/2.7 points above MM-R5 while reducing cached reranking time from 3.82/3.74 seconds to 0.36 seconds. It remains within 1.9/2.6 points of GPT-5-mini yet is 64.9×/68.6× faster, demonstrating that visual-context compression can retain most of the listwise quality without API-model latency. miniReranker (Fan et al., 2026) instead accepts the pointwise formulation and attacks its per-pair overhead through visual cache reuse and interaction sparsity. DocRetriever (Hu et al., 2026c) takes a third route, targeting the *generalisation* weakness of supervised rerankers with reasoning-augmented demonstrations so that a single reranker transfers across domains without per-corpus retraining. The common thread is that the paradigm is increasingly chosen for its cost profile and then repaired, rather than chosen for its statistical elegance.

3.3 RAG Pipelines & Agentic Systems

3.3.1 Formulation

The integration of a VDR embedding model $\mathcal{E}(\cdot)$ and a reranker $R(\cdot, \cdot)$ into a RAG pipeline can be formulated as a multi-stage process. Given a query q and a corpus $\mathcal{C}$, the process unfolds as follows:

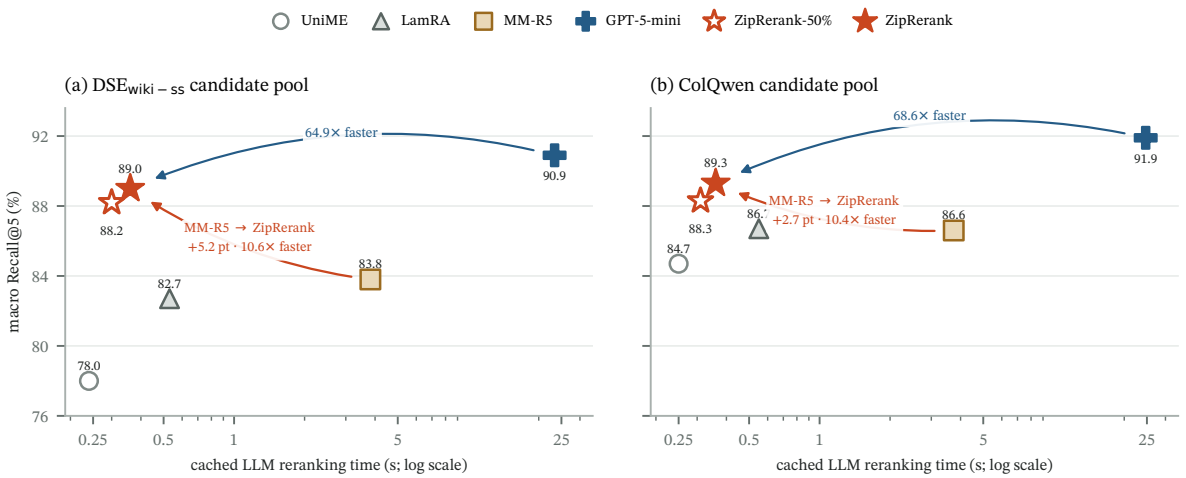


Figure 10: The quality–latency frontier of listwise visual-document reranking. Results from ZipRerank (Sun et al., 2026); each panel fixes the first-stage candidate pool and compares macro Recall@5 with cached-LLM reranking wall time on a log scale, excluding visual encoding and preprocessing. ZipRerank improves over MM-R5 by 5.2/2.7 points while being 10.6×/10.4× faster on the DSE$_{wiki-ss}$/ColQwen pools, respectively; it approaches the stronger GPT-5-mini listwise judge at a small fraction of its latency.

❶ **First-Stage Retrieval:** The embedding model efficiently retrieves an initial set of k candidate documents $\mathcal{C}_k(q)$ from the corpus according to the relevance function defined in Section 2.1: $\mathcal{C}_k(q) = \operatorname{Top}_{d \in \mathcal{C}}{}_k\ s(q,d)$.

❷ **Second-Stage Reranking:** The reranker jointly scores or orders the candidate set and returns its top $j \le k$ documents: $\widetilde{\mathcal{C}}_j(q) = \operatorname{Top}_j\big(R\big(q, \mathcal{C}_k(q)\big)\big)$. This notation covers pointwise score sorting as well as pairwise or listwise ordering.

❸ **Augmented Generation:** Finally, a generative MLLM $G(\cdot)$ synthesizes an answer a by conditioning on both the original query and the reranked documents: $a = G\big(q, \widetilde{\mathcal{C}}_j(q)\big)$.

In an **agentic system**, this process becomes dynamic and iterative. An agent A uses the VDR model as a tool. At each step t, based on the query q and internal state (or history) σ_t, the agent generates an action $\alpha_t = A(q, \sigma_t)$. If α_t issues a retrieval query q_t, the VDR system returns an evidence set $\mathcal{Z}_t = \text{Retrieve}(q_t)$, after which the agent updates its state as $\sigma_{t+1} = \text{Update}(\sigma_t, q_t, \mathcal{Z}_t)$ and decides on the next action.

3.3.2 Current Paradigms and Key Trends

The integration of VDR into RAG and agentic systems has moved beyond simple document fetching, evolving into sophisticated frameworks that emulate human-like reasoning and interaction (Li et al., 2025g; Arya & Gaud, 2025). This evolution is characterized by several key trends, progressing from foundational end-to-end pipelines to complex, iterative, and deeply aware reasoning workflows.

❶ **Foundational Multimodal RAG Pipelines.** The most fundamental shift has been the move from brittle OCR-based textual RAG to robust, end-to-end multimodal pipelines that process documents as visual inputs. This paradigm preserves critical layout and graphical information often lost in text extraction. A prime example is M3DocRAG (Cho et al., 2024), which establishes a flexible framework for multi-page and multi-document question answering by directly retrieving relevant page images for a multimodal generator, forming a foundational approach for subsequent innovations.

❷ **Expansion of Interaction Modalities.** Building on the visual-centric pipeline, researchers are expanding the interaction modalities beyond traditional text-based queries to create more natural and accessible interfaces. A pioneering work in this direction is TextlessRAG (Xie et al., 2025), which introduces a fully

"textless" pipeline that directly processes speech queries and generates spoken answers without any explicit Automatic Speech Recognition (ASR) or Text-to-Speech (TTS) steps, showcasing the potential to significantly broaden the application scenarios of VDR.

❸ **Emergence of Agentic and Iterative Reasoning Workflows.** A dominant trend is the replacement of static, linear RAG pipelines with dynamic, agentic systems that perform multi-step, iterative reasoning. These systems decompose complex queries and progressively refine evidence, mimicking human research processes.

- **Task Decomposition and Collaboration:** Many frameworks now employ a "society of agents," where specialized agents collaborate to solve a problem. For instance, MDocAgent (Han et al., 2025b) utilizes a team of five agents (e.g., General, Critical, Text, and Image agents) to synthesize insights from different modalities. Similarly, ViDoRAG (Wang et al., 2025b) introduces a coarse-to-fine workflow where a "Seeker" agent hunts for relevant images and an "Inspector" agent provides detailed feedback, enabling iterative evidence refinement.
- **Iterative Deep Research Workflows:** This agentic concept is scaled further in systems designed for "deep research." Doc-Researcher (Dong et al., 2025c) implements a comprehensive multi-agent framework with a "Planner" for query decomposition and a "Searcher-Refiner" loop that iteratively gathers and filters evidence across multiple documents and granularities (*e.g.,* chunks, pages, or summaries).

❹ **Enhancing Core RAG Components with Advanced Mechanisms.** Beyond structuring the workflow, significant innovation is occurring within the core retrieval and generation components themselves to make them more intelligent and aware.

- **Holistic Knowledge Retrieval and Fusion:** To address the challenge that standard retrieval often misses nuanced information, new methods are designed for more comprehensive knowledge extraction. HKRAG (Tong et al., 2025) explicitly models and retrieves both "salient" and "fine-print" knowledge using a hybrid masking retriever and an uncertainty-guided generator. In a similar vein, VisDoMRAG (Suri et al., 2025) runs parallel textual and visual RAG pipelines and then employs a consistency-constrained fusion mechanism to intelligently integrate their outputs.
- **Active Visual Perception and Learning:** The most advanced systems empower agents with the ability to actively interact with retrieved visual content. VRAG-RL (Wang et al., 2025c) pioneers this by defining a visual perception action space that allows an agent to perform actions like "crop" and "zoom" on retrieved images. This interactive process is optimized using reinforcement learning, enabling the agent to actively seek out fine-grained details in a coarse-to-fine manner, much like a human analyst.

What does the agentic loop actually buy? VisDocAgentBench (Hu et al., 2026b) makes the answer measurable rather than purely architectural. On the shared Qwen3 visual index in Figure 11(a), static Recall@1 falls from 40.0 on direct queries to 2.5 on two-bridge queries, whereas a GPT-5.6-sol agent retains 40.0; the separate Nemotron baseline shows that even 97.5 direct-match recall can still collapse to 2.5 when bridge evidence is required. This is a diagnostic contrast rather than a strict model-controlled comparison, but panel (c) supplies controlled component evidence: within the same GPT-5.6-sol agent, removing iterative search costs 8.34/9.17 Recall@1 points on the visual/OCR routes, and removing page inspection costs 15.84/21.67.

The representation and planner remain coupled. Every one of the ten planners in Figure 11(b) performs better with visual than OCR-text retrieval, so the modality advantage established in Section 3.1 survives inside an agentic loop rather than being washed out by it. At the same time, the wide spread among planners shows that the agent is not a commodity layer over an interchangeable retriever, strengthening the co-design argument of Section 4.4.

In summary, the application of VDR in RAG and agentic systems is rapidly maturing from a simple retrieval-and-generation process to highly dynamic, interactive, and collaborative workflows. The field is pushing

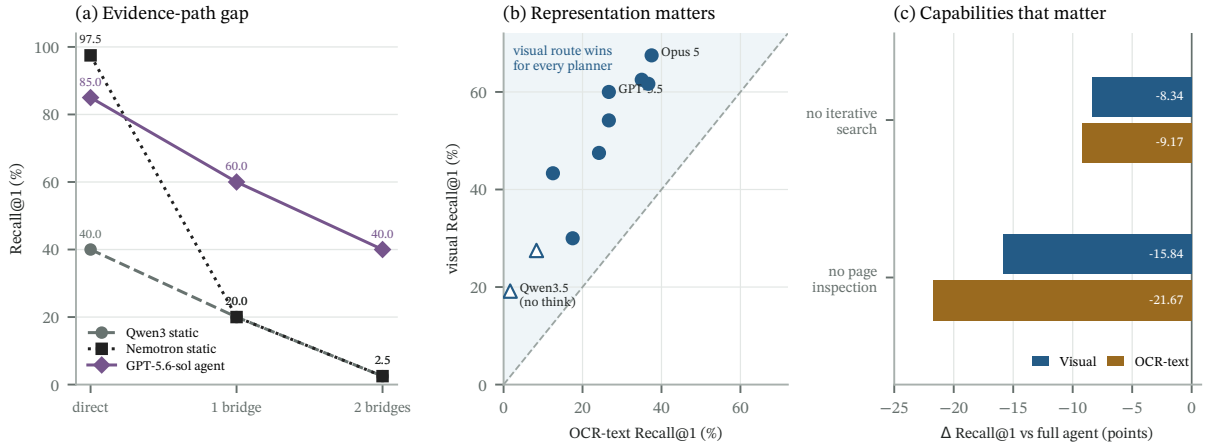


Figure 11: What interaction contributes to visual-document retrieval. Results from VisDocAgentBench (Hu et al., 2026b). **(a)** Recall@1 by evidence-path depth; the GPT-5.6-sol agent and Qwen3 static baseline use the same Qwen3 visual index, whereas Nemotron is a stronger but different static retriever. **(b)** Visual versus OCR-text Recall@1 across ten planners; marker shape distinguishes open-weight Qwen3.5 planners. **(c)** Controlled GPT-5.6-sol ablations, expressed as the point loss relative to the full agent. Together, the panels separate the need for iterative evidence acquisition, the value of native visual retrieval, and the contribution of page inspection.

towards systems that not only find relevant documents but also intelligently reason, synthesize, and interact with multimodal information to solve complex problems (Ashraf et al., 2025; Zhang et al., 2025i; Yan et al., 2025c; Su et al., 2025a).

3.3.3 The Evolutionary Synergy of VDU and VDR

The boundary between Visual Document Understanding (VDU) and VDR is rapidly dissolving in the MLLM era. Modern RAG pipelines are evolving into autonomous agentic systems capable of perception, strategic planning, and iterative refinement.

VDU Trends in RAG and Agentic Systems. Recent breakthroughs emphasize moving from single-turn OCR-based retrieval to multi-step reasoning-driven navigation. (Sourati et al., 2025) introduces LAD-RAG, which leverages an LLM agent to dynamically interact with a symbolic document graph and neural indices, capturing structural dependencies missed by dense embeddings. (Yu et al., 2025b) proposes MACT, a collaborative framework that decomposes document intelligence into planning, execution, and judgment agents, implementing a self-correction loop via procedural scaling. To address evidence sparsity, MHier-RAG (Gong et al., 2025) facilitates multi-granularity reasoning by retrieving parent pages and document summaries, while MoLoRAG (Wu et al., 2025) constructs page graphs to navigate logical connections beyond surface-level semantic similarity. For fine-grained localization, DocLens (Zhu et al., 2025a) employs a tool-augmented "zoom-in" strategy where agents locate specific visual elements like tables or charts within retrieved pages. Finally, SLEUTH (Liu et al., 2025b) optimizes the input quality through context engineering, utilizing page-screening agents to filter visual noise and construct evidence-dense multimodal contexts.

Comparison between VDR and VDU. Table 5 illustrates the multi-dimensional differences and the emerging convergence between retrieval and understanding tasks. While VDR focuses on efficient large-scale candidate localization, VDU emphasizes deep reasoning. The integration of the two, as seen in recent agentic frameworks, enables systems to perform "Retrieval as Understanding."

Reflections on VDR: Toward Cognitive Discovery. Future VDR systems should evolve from passive semantic matchers to active cognitive navigators. The integration of reinforcement learning (as in VRAG-RL

Dimension	Visual Document Retrieval (VDR)	Visual Document Understanding (VDU)	Convergence Trend
Core Objective	Candidate page localization	Information extraction & reasoning	R → U Integrated
Granularity	Coarse (Document/Page-level)	Fine (Element/Token-level)	Hierarchical Indexing
Compute Pattern	Broad, shallow (whole corpus)	Narrow, deep (top-k)	Adaptive Resource Scaling
Key Technique	Late Interaction, Multi-vector Indexing	Visual CoT, Multimodal Fusion	Agent-driven Seek-then-Verify
Main Metric	nDCG, Recall@k	Accuracy, F1-score	Source Attribution Accuracy

Table 5: Multi-dimensional comparisons between VDR and VDU. Recent agentic systems bridge the gap through iterative reasoning and dynamic filtering.

(Wang et al., 2025c)) and symbolic-neural fusion (as in LAD-RAG (Sourati et al., 2025)) suggests that the next frontier of multimodal document intelligence lies in the ability to understand the "logical topology" of a document. Rather than just returning snippets, future retrievers will actively discover implicit contradictions and synthesize cross-document insights, transforming VDR into a tool for complex decision support in expert domains.

It is worth being precise about what "convergence" does and does not mean here, since the claim is easy to overstate. VDR and VDU are not becoming the same task: their cost structures remain opposed, as Table 5 records: retrieval must touch the whole corpus cheaply, understanding must touch a handful of pages deeply. What is dissolving is the *interface* between them. In a one-shot pipeline that interface is a ranked list, a lossy summary of everything the retriever knew. In an agentic loop the interface becomes a dialogue, so partial understanding can reshape the next retrieval and retrieval can supply the evidence that understanding lacked. The practical consequence is an evaluation gap: neither nDCG, which ignores whether the evidence sufficed, nor answer accuracy, which cannot say whether a failure was retrieval's or the reader's, measures the quality of that interface. Source-attribution accuracy is the most promising candidate precisely because it is the one metric that both stages can be blamed for.

4 Challenges & Outlook

As shown in Figure 12, building a truly *effective*, *efficient*, and *interactive* VDR system entails persistent challenges. The ultimate goal extends beyond simple document matching to enabling nuanced, real-time document intelligence that is scalable and trustworthy. Achieving this vision requires addressing critical bottlenecks across the entire VDR ecosystem. These challenges span from the foundational level of data (▷ Section 4.1) and model architectures (▷ Section 4.2) to the practical imperatives of performance efficiency (▷ Section 4.3) and system interactivity (▷ Section 4.4), and culminate in the higher-order needs for understanding model scaling behavior (▷ Section 4.5). The following sections delineate five core challenges and outline promising future directions to navigate these frontiers.

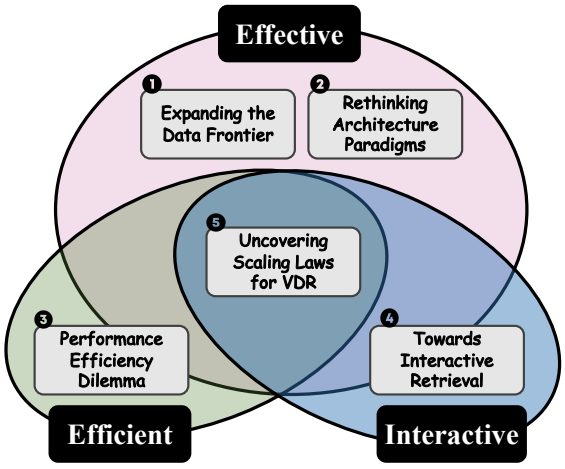


Figure 12: Overview of five future challenges for VDR.

4.1 Expanding the Data Frontier

The foundation of robust VDR systems is high-quality, diverse, and challenging data, yet the current landscape of benchmarks and training sets presents a significant bottleneck to progress. A primary limitation is the lack of diversity in language, domain, and document structure. While recent benchmarks like Jina-VDR (Günther et al., 2025) and MIRACL-VISION (Osmulski et al., 2025) have made commendable strides in multilingual support, the majority of available resources remain predominantly English-centric and confined to general-domain web documents. Furthermore, most benchmarks focus on retrieving single, self-contained pages, failing to capture the complex, multi-hop, and cross-document reasoning required for genuine information synthesis, a gap that pioneering benchmarks like MR2-Bench (Zhou et al., 2025a) and MRMR (Zhang et al., 2026b) are only beginning to explore. A more insidious issue lies in data provenance; the heavy reliance on VLM-generated queries for training data, a common practice in datasets, risks creating a hermetic feedback loop where models are evaluated on the same kind of logic they were trained on, inflating performance metrics without guaranteeing real-world generalization. This is compounded by the potential for data leakage between large, web-crawled training sets and evaluation benchmarks, which casts a shadow over the true robustness of current models (Apicella et al., 2025; Lilja et al., 2024).

Addressing these data frontiers requires a multi-pronged, forward-looking strategy that moves beyond mere data scaling. A primary imperative is the creation of large-scale benchmarks that are ***not only multilingual but also multi-domain***, extending into specialized corpora like legal contracts, financial reports, and medical records, which demand expert-level nuance (Tang & Yang, 2024; Cao, 2024; Chung et al., 2025). To cultivate ***reasoning-intensive capabilities***, future data curation can leverage agentic frameworks to automatically generate complex, multi-hop query-evidence chains that span multiple documents, simulating the realistic research workflows envisioned by systems like Doc-Researcher (Dong et al., 2025c). To break the cycle of synthetic data bias and enhance authenticity, a pivot towards ***incorporating real human queries*** is essential; this could be achieved by mining anonymized search logs or utilizing human-in-the-loop annotation platforms to capture the ambiguity and complexity of genuine user intent. Finally, to ensure ***robust and fair evaluation***, the community must adopt stricter data governance practices. This includes establishing "firewalled" test sets derived from entirely held-out web crawls or proprietary document collections and developing standardized protocols for data contamination detection, thereby fostering a more reliable assessment of model generalization and pushing the field towards true document intelligence (Song et al., 2024; Xu et al., 2024; Samuel et al., 2025).

4.2 Rethinking Architectural Paradigms

The multimodal retrieval field has largely converged on a paradigm centered around contrastively-trained discriminative models, typically bi-encoder architectures that learn to maximize the similarity between positive query-document pairs while minimizing it for negative ones. While highly effective for semantic matching, this approach has inherent limitations. It primarily treats powerful MLLMs as static feature extractors (Mischler et al., 2024; Jeong et al., 2024), underutilizing their sophisticated generative and reasoning capabilities, which can lead to information bottlenecks where fine-grained details or implicit document semantics are lost during compression into a fixed-budget representation. Moreover, this "one-size-fits-all" strategy forces a single, monolithic model to act as a generalist for an incredibly diverse array of document structures, from text-heavy reports and dense tables to complex charts and forms, limiting both performance specialization and computational efficiency (Gan et al., 2025b; Fan et al., 2024; Lo et al., 2025).

To overcome these limitations, future research is pivoting towards two promising architectural shifts: autoregressive retrieval and the integration of Mixture of Experts (MoE). The first paradigm ***reframes retrieval from a purely discriminative task to a generative one*** (Deng et al., 2025a; Xiong et al., 2024). For instance, future models could adopt an "embed-via-answering" framework, where embeddings are derived from generating a concise answer to an instruction, as pioneered by InBedder (Peng et al., 2024). More advanced paradigms may even involve generating sequences of non-human-readable "soft tokens" that iteratively refine the embedding, a concept introduced by GIRCSE (Tsai et al., 2025) which learns to "speak an embedding language" and shows improved quality with more generation steps at test time. This generative-representational synergy is also validated by frameworks like GritLM (Muennighoff et al., 2025), which successfully unifies

generative and embedding tasks within a single model, demonstrating that these capabilities can coexist without performance degradation. We summarize and compare the representative works from text-only and multimodal generative retrieval in Table 6.

Representative Works	Modality	Core Paradigm	Generated Content	Embedding Source	Training Objective
InBedder (Peng et al., 2024)	Text	Embed-via-Answering	Concise Answer	Hidden state of the 1st generated token	Instruction Tuning (QA)
GIRCSE (Tsai et al., 2025)	Text	Iterative Refinement	Soft Tokens (Non-readable)	Pooled hidden states of generated soft tokens	Iterative Contrastive Refinement (ICR)
GritLM (Muennighoff et al., 2025)	Text	Task Unification	Text Response or Embedding	Mean pooling over input with bidirectional attention	Joint Contrastive + LM Loss
RGE (Liu et al., 2025a) / **TTE** (Cui et al., 2025b)	Multimodal	Think-then-Embed	Reasoning Trace (CoT)	Hidden state conditioned on generated reasoning	Joint LM (on trace) + Contrastive Loss
CAFe (Yu et al., 2025a) / **VladVA** (Ouali et al., 2025)	Multimodal	Joint Training	N/A (for retrieval)	Last token's hidden state	Joint Contrastive + Autoregressive Loss
Retrv-R1 (Zhu et al., 2025b)	Multimodal	Reasoning-driven Selection	Reasoning trace leading to a selection	Not an embedding model; selects best candidate	SFT + Reinforcement Learning (GRPO)

Table 6: A multi-dimensional comparison of representative works in **generative retrieval**. These approaches move beyond static encoding by leveraging the generative capabilities of (M)LLMs. They differ in their core paradigm, the nature of the generated content (from concise answers to reasoning traces), and their training objectives, which range from joint contrastive-generative losses to reinforcement learning.

Concurrently, ***the MoE architecture offers a path to efficient specialization***. For instance, different experts can be trained to handle distinct document modalities (*e.g.,* text, tables, audio), a direction explored by Uni-MoE (Li et al., 2025i;h) and M3-JEPA (Lei et al., 2024), which use multi-gate MoE to disentangle modality-specific and shared information. This architecture also unlocks novel embedding sources, as demonstrated by MOEE (Li & Zhou, 2024), which shows that the expert routing weights themselves can serve as a potent, complementary embedding signal. Together, these architectural innovations promise to evolve VDR from static matching towards more dynamic, specialized, and generative systems.

4.3 Performance-Efficiency Dilemma

The high accuracy of state-of-the-art multi-vector VDR models stems directly from their fine-grained representation paradigm, where each document page is encoded into hundreds or even thousands of patch-level embeddings (Faysse et al., 2024). While this enables precise query-to-patch matching, it creates a significant performance-efficiency dilemma. The primary issue is the prohibitive storage overhead; for instance, a medium-sized document of just 50 pages can require approximately 10 MB of storage for its embeddings alone (Ma et al., 2025), rendering the large-scale deployment of such models economically and practically challenging. Furthermore, this abundance of vectors increases online computational costs during the late-interaction scoring stage. Figure 13 previews how far a recent structure-aware compressor can move this frontier, motivating embedding reduction techniques that retain retrieval evidence in a much smaller index.

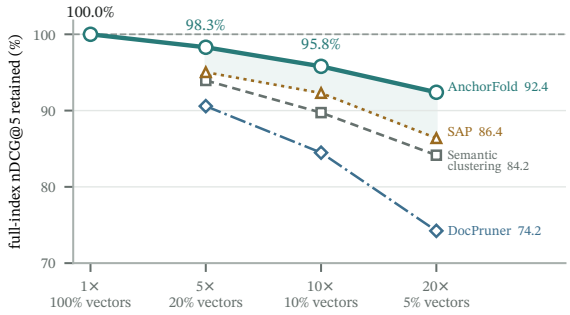


Figure 13: Quality retention under index compression. Results from AnchorFold (Zuo et al., 2026), averaged across three backbones and the ViDoRe v1/v2 suites. AnchorFold remains above three representative pruning or clustering baselines at every evaluated compression ratio.

Future solutions to this dilemma are evolving from post-hoc compression to built-in adaptability. The two primary post-hoc strategies are ***embedding pruning and merging*** (Yan et al., 2026b; Cha et al., 2026; Qin et al., 2026; Pony et al., 2026; Liu et al., 2026a). Pruning aims to discard redundant embeddings. While early heuristic-based methods proved unstable (Liu et al., 2024; Lassance et al., 2021; 2022), more sophisticated approaches like DocPruner (Yan et al., 2025d) use attention scores to adaptively prune 50-60% of patches per document with minimal performance loss. Alternatively, merging or clustering aggregates similar patches, a method that some argue is more suitable as it retains information rather than discarding it. For example, Light-ColPali (Ma et al., 2025) employs hierarchical clustering to merge embeddings, while HEAVEN (Kim et al., 2025a) uses visually-summarized pages to create a compact first-stage index. Recent work has pushed this in three directions that are worth distinguishing, because they trade different resources. *Learned pooling* keeps the number of vectors fixed but makes the reduction itself trainable, as in the lightweight fine-tuning of Josef (2026); *coverage-aware merging* such as MarginMerge (Mahdizadeh et al., 2026) selects

which patches to fuse so that the retained set still spans the page rather than clustering redundantly; and *quantisation* such as CrossQ (Salla et al., 2026) leaves the vector count untouched and instead shrinks each vector's width, which composes with either of the former. AnchorFold (Zuo et al., 2026) illustrates that these are not mutually exclusive, folding attention onto anchor tokens recursively to cut both count and cost. Quantitatively, Figure 13 retains 98.3% of full-index nDCG@5 at 5× compression and 92.4% at 20×, while storing only 5% of the vectors; at the latter budget it remains 6.0 points above the strongest baseline shown. This makes fold depth a promising corpus- or query-adaptive budget that future systems can combine with vector quantisation rather than treating compression methods as mutually exclusive.

A more forward-looking paradigm is ***Matryoshka Representation Learning (MRL)*** (Kusupati et al., 2022), which trains embeddings to be inherently flexible. MRL ensures that a single high-dimensional embedding contains a hierarchy of smaller, nested embeddings that are also effective. Flexible Matryoshka embeddings have been adopted by top-performing text embedding models (*e.g.,* Qwen3 Embedding (Zhang et al., 2025h), Gemini Embedding (Lee et al., 2025b), KaLM-Embedding-V2 (Zhao et al., 2025), EmbeddingGemma (Vera et al., 2025), jina-embeddings-v3 (Sturua et al., 2024), jina-colbert-v2 (Jha et al., 2024)) across diverse real-world tasks (Lai et al., 2024; Wang et al., 2024d; Nacar & Koubaa, 2025; Hanley & Durumeric, 2025; Fu et al., 2025). MM-Matryoshka (Xiang et al., 2026) transfers this idea to multi-vector VDR by nesting both vector prefixes and encoder layers. At the joint middle-layer-plus-32d budget, Figure 14 shows 74.3–85.8% retention across three backbones, versus only 14.4–31.4% under direct truncation—a 54.4–59.9-point gain. This makes a single two-dimensionally elastic checkpoint plausible; future systems could select depth and width jointly for each corpus, device, or query instead of maintaining separate retrievers. Formally, at full width $h_{\max}$, the encoder $\mathcal{E}$ produces the multi-vector representations $\mathbf{U}(q)$ and $\mathbf{V}(d)$ introduced in Section 2.1, with every constituent vector in $\mathbb{R}^{h_{\max}}$. MRL trains the model on a set of nested widths $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$, where $h_1 < h_2 < \ldots < h_K = h_{\max}$. For each h_i, $\mathbf{U}^{[h_i]}(q)$ and $\mathbf{V}^{[h_i]}(d)$ retain the first h_i components of every query and document vector, respectively. The training objective $\mathcal{L}_{\text{MRL}}$ sums a standard representation loss $\ell(\cdot,\cdot)$ over all nested widths:

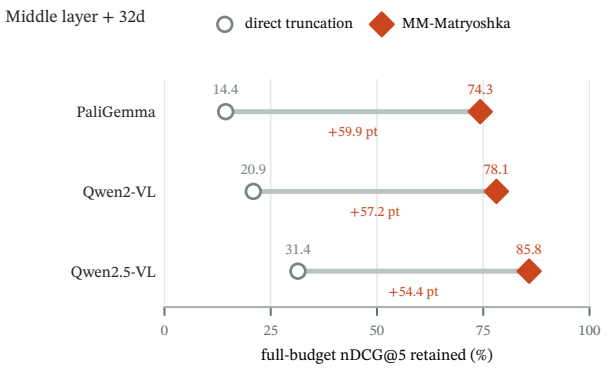


Figure 14: Joint depth–width elasticity in MM-Matryoshka. Results from MM-Matryoshka (Xiang et al., 2026): ViDoRe-v2 nDCG@5 retained at each backbone's middle layer and 32d, relative to its final-layer 128d reference.

$$\mathcal{L}_{\text{MRL}}(q, d^+) = \sum_{h_i \in \mathcal{H}} \lambda_i \, \ell\left(\mathbf{U}^{[h_i]}(q), \mathbf{V}^{[h_i]}(d^+)\right) \tag{11}$$

where q is a query, d^+ is a positive document, and λ_i are loss weights. This allows practitioners to truncate stored embeddings at inference time, providing a flexible dial to balance performance and efficiency. Building on this, 2DMSE (Li et al., 2024c) extends this elasticity to model depth, while frameworks like Starbucks (Zhuang et al., 2024) and MatTA (Verma et al., 2025) refine the training process to bridge the performance gap with individually trained models. The Matryoshka principle has also been generalized to new granularities and architectures: M3 (Cai et al., 2024) and MQT (Hu et al., 2024) apply it to the number of visual tokens in LVLMs; Matryoshka Re-Ranker (Liu et al., 2025g) enables flexible depth and sequence length for re-ranking; and MatMamba (Shukla et al., 2024) integrates it into State Space Models. The application of MRL has further expanded to diverse tasks such as improving feature hierarchy in Sparse Autoencoders (Bussmann et al., 2025) and enabling multi-precision models via Matryoshka Quantization (Nair et al., 2025). Finally, innovations in the training paradigm itself, such as the sequential learning in SMEC (Zhang et al., 2025a) and the post-hoc tuning of Matryoshka-Adaptor (Yoon et al., 2024), are making the creation of these flexible embeddings more efficient and accessible. Besides, recent works also focus on sparse coding (Wen et al., 2025; Guo et al., 2026) and isolation kernel (Zhang et al., 2026c) for more lightweight representation learning. We compare the representative works above in Table 7.

Table 7: A multi-dimensional comparison of representative works in **Matryoshka Representation Learning (MRL)**. The principle of nested, adaptable representations has been extended from embedding dimensions to model depth, token counts, and even quantization bit-widths, addressing a wide range of tasks from efficient retrieval to model interpretability.

Representative Works	Modality	Target Task	Matryoshka On	Core Innovation
MRL (Kusupati et al., 2022)	Multimodal	Classification / Retrieval	Embedding Dimension	Foundational concept of nested embeddings.
2DMSE (Li et al., 2024c)	Text	STS / Retrieval	Embedding Dimension & Model Depth	Extends MRL to two dimensions for greater flexibility.
Starbucks (Zhuang et al., 2024)	Text	STS / Retrieval	Embedding Dimension & Model Depth	Improves 2DMSE to match individually trained models.
MatTA (Verma et al., 2025)	Text	General Tasks	Model Depth & Width (FFN)	Teacher-TA-Student distillation for elastic student models.
M3 (Cai et al., 2024) / **MQT** (Hu et al., 2024)	Multimodal	VQA / Reasoning	Number of Visual Tokens	Enables adaptive visual granularity in LVLMs.
Matryoshka Re-Ranker (Liu et al., 2025g)	Text	Re-ranking	Model Depth & Sequence Length	Creates flexible re-rankers with configurable architecture.
SMEC (Zhang et al., 2025a)	Multimodal	Retrieval / Classification	Embedding Dimension	Sequential training to mitigate gradient variance.
Matryoshka-Adaptor (Yoon et al., 2024)	Multimodal	Retrieval	Embedding Dimension	Post-hoc tuning to impart Matryoshka properties.
MM-Matryoshka (Xiang et al., 2026)	Multimodal	Retrieval	Embedding Dimension & Encoder Depth	One multi-vector checkpoint with selectable 2D budgets.
MatMamba (Shukla et al., 2024)	Multimodal	LM / Classification	Hidden Dimension & Heads	Generalizes MRL to State Space Model (SSM) architectures.
Matryoshka SAE (Bussmann et al., 2025)	Model Activations	Interpretability	SAE Dictionary Size	Mitigates feature splitting in Sparse Autoencoders.
Matryoshka Quantization (Nair et al., 2025)	Model Weights	Model Compression	Quantization Bit-width	Leverages nested structure of integer data types.

4.4 Towards Interactive Retrieval

Integrating VDR with agentic systems holds immense potential, particularly for complex scenarios like *Deep Research*, where a query requires iterative evidence gathering from a vast candidate pool (Java et al., 2025; Chen et al., 2025e; Huang et al., 2025c; Zhang et al., 2025e). The current challenge, however, is that most agentic systems treat VDR models as passive, reactive tools rather than active, strategic partners. These agents typically operate within a predefined, reactive loop of reasoning and acting, which limits their ability to dynamically formulate or adapt a high-level retrieval strategy. For instance, while frameworks like WebThinker (Li et al., 2025f) and DeepResearcher (Zheng et al., 2025b) enable agents to interleave thought with web search actions, they often function as sophisticated *tool-executors* that reactively process information rather than as *research strategists* that proactively plan a multi-step information-seeking campaign. This limitation becomes particularly evident in complex research tasks, such as generating structured reports with dynamic outlines as demonstrated by WebWeaver (Li et al., 2025j) or creating text-chart interleaved content as in Multimodal DeepResearcher (Yang et al., 2025b), where the retrieval process itself must be strategically guided to gather diverse and structured evidence.

Future work must focus on the ***co-design of agents and VDR tools to foster a more organic and strategic synergy***. A primary direction is to empower agents with explicit retrieval planning capabilities, enabling them to decompose a high-level query into a multi-step, multi-granularity retrieval plan. This moves beyond simple tool use towards strategic orchestration. For example, a planner agent, inspired by the hierarchical structure of Cognitive Kernel-Pro (Fang et al., 2025), could adaptively select *retrieval granularity* (choosing between document summaries, full pages, or specific visual elements) and delegate these sub-tasks to specialized agents. To handle the inevitable "dead ends" in complex research, agents must also develop self-correction mechanisms. Frameworks like DeepAgent (Li et al., 2025e), with its autonomous memory folding, offer a powerful blueprint by allowing an agent to "take a breath," discard a failed exploration path, and restart from a consolidated memory state. To effectively learn such complex strategies, agents require more fine-grained feedback. This can be achieved through mechanisms like the Atomic Thought Rewards (ATR) proposed in Atom-Searcher (Deng et al., 2025b), which uses a reasoning reward model to provide process-level supervision on the agent's retrieval strategy. Finally, for massive-scale research, parallel exploration frameworks like the Research-Synthesis approach in WebResearcher (Qiao et al., 2025) and Tongyi DeepResearch (Team et al., 2025b), where multiple agents conduct concurrent searches, can transform interactive retrieval into a scalable, collaborative process. The ultimate goal is to evolve the VDR system from a passive tool into an active, *strategic partner* in the knowledge discovery process.

4.5 Uncovering Scaling Laws for VDR

While scaling laws (predictable power-law relationships between model performance, size, and data volume) are well-documented for general-purpose LLMs (Alabdulmohsin et al., 2022; Zhang et al., 2024a; Xiao et al., 2025a; Pearce et al., 2024), their application to the specialized domain of VDR remains a largely unexplored and complex frontier. The core challenge is that the scaling behavior in VDR is not a straightforward extrapolation of model size or data quantity. On the model axis, naively increasing parameter count does not

guarantee proportional performance gains. Larger models, without proper fine-tuning, can exhibit greater embedding anisotropy, which paradoxically harms retrieval performance in zero-shot settings by compressing the effective embedding space, even as they capture richer features for transfer tasks (Jiang et al., 2024a; Wu et al., 2026). This complexity is mirrored on the data axis, where simply increasing the volume of raw documents is insufficient. The effectiveness of VDR models is deeply tied to the quality and diversity of the training pairs. Recent studies powerfully illustrate this, showing that models trained on smaller, high-quality synthetic datasets can significantly outperform those trained on orders-of-magnitude more, but less curated, data. For example, both mmE5 (Chen et al., 2025b) and MegaPairs (Zhou et al., 2025b) demonstrated that superior performance can be achieved with a fraction of the data if it is diverse and of high quality, underscoring that scaling in VDR is a nuanced interplay between model capacity and data sophistication, rather than a simple numbers game.

To systematically navigate this challenge, future research must pivot from ad-hoc scaling to ***establishing formal, VDR-specific scaling laws, adapting methodologies from the dense retrieval domain*** (Fang et al., 2024). A foundational step is to adopt continuous and sensitive evaluation metrics, such as contrastive entropy, which can more accurately capture subtle performance changes compared to discrete ranking metrics like nDCG, thus enabling precise power-law curve fitting for parameter count (N_θ) and training-set size (N_{train}). Concurrently, advancing the "data" axis of the scaling law requires moving beyond simple data augmentation to sophisticated, document-aware synthetic data generation. This involves leveraging MLLMs not just for query generation but for creating entire ecosystems of diverse training tasks (Wang et al., 2024a), mining heterogeneous relationships between documents using multiple similarity models (*e.g.,* visual-semantic and visual-pattern correlations) (Zhou et al., 2025b), and implementing rigorous quality control through mechanisms like single-pass multi-aspect interpretation and self-refinement (Chen et al., 2025b). By combining a formal scaling law framework with advanced data synthesis, the VDR community can quantitatively model the trade-offs between investing in larger models versus more extensive data annotation. This will enable more efficient resource allocation and pave the way for building maximally effective VDR systems under practical budget constraints.

5 Conclusion

This survey systematically discusses the VDR landscape, categorizing the evolution of benchmarks and methodologies, including embedding models, rerankers, and their integration with RAG and agentic systems. Current triumphs push the boundaries of fine-grained matching, but the field's trajectory is shifting towards more complex document intelligence. Navigating the future frontiers of data complexity, architectural innovation, efficiency, and interactivity will be critical for realizing the full potential of VDR as a cornerstone of multimodal document intelligence.

6 Broader Impact and Societal Considerations

While this survey focuses on the technical advancements of VDR in the MLLM era, the deployment of such systems carries significant broader societal implications. As VDR transitions from academic benchmarks to real-world deployment in legal, medical, and enterprise workflows, we must critically evaluate its impact on accountability, equity, and environmental sustainability.

Transparency, Accountability, and Trust. A primary concern in deploying VDR systems, particularly in high-stakes domains, is the "black-box" nature of multimodal reasoning. When retrieval decisions are based on opaque, latent-space alignments, the inability to provide clear, human-interpretable rationales (or citations to evidence) complicates auditability. To ensure responsible deployment, the development of VDR should not only prioritize retrieval accuracy but also emphasize traceable reasoning. Systems that integrate "retrieval-as-reasoning", where the model provides an explicit chain-of-thought or highlights the specific visual/textual evidence, are essential for establishing trust and accountability, especially in regulatory contexts where document decisions must be legally defensible.

Mitigating Bias and Promoting Multilingual Equity. The VDR landscape remains heavily skewed towards English-centric datasets and models. If left unchecked, this "linguistic and cultural bottleneck"

threatens to marginalize non-English speaking communities and industries, potentially reinforcing existing global inequities in information access. The surge in multilingual benchmark development is a necessary step, but technical efforts must also address inherent biases within foundation models. Retrieval systems that unknowingly favor dominant cultural or Western-centric document formats risk systematic information omission. We advocate for research that explicitly models fairness across linguistic and structural diversities to ensure that VDR is a globally inclusive technology.

Environmental Impact of Large-Scale Intelligence. The transition from traditional, lightweight OCR-based search to massive-scale VDR models based on multi-vector late-interaction paradigms introduces a significant performance-efficiency dilemma. While these models offer unprecedented semantic depth, they impose substantial computational costs in terms of memory storage, index maintenance, and GPU-intensive inference. In an era where AI-related energy consumption is under intense scrutiny, the pursuit of performance must be balanced with environmental responsibility. Developing sparse, efficient, and Matryoshka-style representations is not merely a technical optimization; it is a critical requirement to ensure that the widespread adoption of multimodal intelligence does not come at the expense of an unsustainable carbon footprint.

Societal Safeguards and Data Integrity. The ability of VDR models to extract fine-grained, structured information from unstructured PDFs, including sensitive personal data, private business intelligence, and restricted intellectual property, poses risks of unintended information disclosure. The shift toward agentic RAG systems that "browse" and synthesize evidence across vast document stores exacerbates the danger of bypassing existing security controls. Therefore, we underscore the necessity of robust data governance, including fine-grained access control and privacy-preserving retrieval architectures, as foundational requirements for future VDR systems.

In summary, the evolution of VDR toward more sophisticated, autonomous, and reasoning-intensive workflows necessitates a proactive approach to societal safeguards. We hope this perspective encourages the community to view technical robustness, interpretability, and ecological sustainability not as peripheral concerns, but as integral pillars of responsible multimodal document intelligence.

References

Abdelrahman Abdallah, Mahmoud Abdalla, Mohammed Ali, and Adam Jatowt. Argus-retriever: Vision-LLM late-interaction retrieval with region-aware query-conditioned MoE for visual document retrieval. *arXiv preprint arXiv:2606.04300*, 2026a. URL https://arxiv.org/abs/2606.04300.

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun-Soo Kang. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09562*, 2026b.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.861. URL https://aclanthology.org/2025.findings-acl.861/.

Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022. doi: 10.52202/068431-1620. URL https://doi.org/10.52202/068431-1620.

Fahimeh Alaei, Alireza Alaei, Michael Blumenstein, and Umapada Pal. A brief review of document image retrieval methods: Recent advances. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3500–3507. IEEE, 2016. doi: 10.1109/IJCNN.2016.7727648. URL https://doi.org/10.1109/IJCNN.2016.7727648.

Andrea Apicella, Francesco Isgrò, and Roberto Prevete. Don't push the button! exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*, 58(11):339, 2025. doi: 10.1007/s10462-025-11326-3. URL https://doi.org/10.1007/s10462-025-11326-3.

ApsaraStackMaaS. EvoQwen2.5-VL-Retriever-3B-v1. Hugging Face model card, 2025. URL https://huggingface.co/ApsaraStackMaaS/EvoQwen2.5-VL-Retriever-3B-v1.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia computer science*, 246:3781–3790, 2024. doi: 10.1016/j.procs.2024.09.178. URL https://doi.org/10.1016/j.procs.2024.09.178.

Sumedha Arya and Nirmal Gaud. Advances in retrieval-augmented generation (rag) and related frameworks. *IJSAT-International Journal on Science and Technology*, 16(3), 2025. doi: 10.71097/IJSAT.V16.I3.7595. URL https://doi.org/10.71097/IJSAT.V16.I3.7595.

Tajamul Ashraf, Amal Saqib, Hanan Ghani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, et al. Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks. *arXiv preprint arXiv:2505.24876*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Juan Pablo Balarini. Eager embed v1: Multimodal dense embeddings for retrieval. GitHub repository, 2025. URL https://github.com/eagerworks/eager-embed.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Jing Bi, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Ali Vosoughi, Guangyu Sun, et al. Why reasoning matters? a survey of advancements in multimodal reasoning (v1). *arXiv preprint arXiv:2504.03151*, 2025.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

ByteDance. Seed1.6-Embedding. Project page, 2025. URL https://seed1-6-embedding.github.io/.

Hongyi Cai. When vision meets texts in listwise reranking. *arXiv preprint arXiv:2601.20623*, 2026.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.

Hongliu Cao. Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark. *arXiv preprint arXiv:2406.01607*, 2024.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, et al. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*, 2025.

Sungguk Cha, DongWook Kim, Mintae Kim, Youngsub Han, Byoung-Ki Jeon, and Sangyeob Lee. Reinpool: Reinforcement learning pooling multi-vector embeddings for retrieval system. *arXiv preprint arXiv:2601.07125*, 2026.

Antoine Chaffin and Aurélien Lac. Monoqwen: Visual document reranking. Hugging Face model card, 2024. URL https://huggingface.co/lightonai/MonoQwen2-VL-v0.1.

Antoine Chaffin, Luca Arnaboldi, Amélie Chatelain, and Florent Krzakala. Colbert-zero: To pre-train or not to pre-train colbert models. *arXiv preprint arXiv:2602.16609*, 2026. URL https://arxiv.org/abs/2602.16609.

Dongyang Chen, Chaoyang Wang, Dezhao Su, Xi Xiao, Zeyu Zhang, Jing Xiong, Qing Li, Yuzhang Shang, and Shichao Kan. V-retrver: Evidence-driven agentic reasoning for universal multimodal retrieval. *arXiv preprint arXiv:2602.06034*, 2026a.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. *arXiv preprint arXiv:2506.23115*, 2025a.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. *arXiv preprint arXiv:2502.08468*, 2025b.

Haonan Chen, Sicheng Gao, Radu Timofte, Tetsuya Sakai, and Zhicheng Dou. e5-omni: Explicit cross-modal alignment for omni-modal embeddings. *arXiv preprint arXiv:2601.03666*, 2026b.

Jian Chen, Ming Li, Jihyung Kil, Chenguang Wang, Tong Yu, Ryan Rossi, Tianyi Zhou, Changyou Chen, and Ruiyi Zhang. Visr-bench: An empirical study on visual retrieval-augmented generation for multilingual long document understanding. *arXiv preprint arXiv:2508.07493*, 2025c.

Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian, and Zheng Liu. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *arXiv preprint arXiv:2510.08252*, 2025d.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025e.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*, 2025.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Yongbin Choi, Gyuho Shim, and Youngjoon Jang. KoVRE: Training an efficient embedding model for korean visual document retrieval. *arXiv preprint arXiv:2608.01389*, 2026a. URL https://arxiv.org/abs/2608.01389.

Yongbin Choi, Yongwoo Song, and Mujeen Sung. KoViDoRe: Korean visual document retrieval. *arXiv preprint arXiv:2608.20840*, 2026b. URL https://arxiv.org/abs/2608.20840.

Isaac Chung, Imene Kerboua, Marton Kardos, Roman Solomatin, and Kenneth Enevoldsen. Maintaining mteb: Towards long term usability and reproducibility of embedding benchmarks. *arXiv preprint arXiv:2506.21182*, 2025.

Marco Cimolai and Logan Markewich. Visual Document Retrieval Goes Multilingual. Hugging Face Blog, jan 2025. URL https://huggingface.co/blog/vdr-2b-multilingual.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*, 2021.

Xuanming Cui, Hong-You Chen, Hao Yu, Hao Yuan, Zihao Wang, Shlok Kumar Mishra, Hanchao Yu, Yonghuan Yang, Jun Xiao, Ser-Nam Lim, et al. Reason to contrast: A cascaded multimodal retrieval framework. *arXiv preprint arXiv:2602.23369*, 2025a.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, et al. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025b.

Ziqi Dai, Xin Zhang, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Supervised fine-tuning or contrastive learning? towards better multimodal llm reranking. *arXiv preprint arXiv:2510.14824*, 2025.

Debrup Das, Sam O'Nuallain, and Razieh Rahimi. RaDeR: Reasoning-aware dense retrieval models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19970–19997, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1011. URL https://aclanthology.org/2025.emnlp-main.1011/.

Jingcheng Deng, Zhongtao Jiang, Liang Pang, Zihao Wei, Liwei Chen, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. Following the autoregressive nature of LLM embeddings via compression and alignment. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12661–12677, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.639. URL https://aclanthology.org/2025.emnlp-main.639/.

Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, et al. Atom-searcher: Enhancing agentic deep research via fine-grained atomic thought reward. *arXiv preprint arXiv:2508.12800*, 2025b.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. Deep learning based visually rich document content understanding: A survey. *arXiv preprint arXiv:2408.01287*, 2024.

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Qiang Sun, Geoffrey Martin, Wei Liu, and Yifan Peng. A survey on mllm-based visually rich document understanding: Methods, challenges, and emerging trends. *arXiv preprint arXiv:2507.09861*, 2025.

Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. *arXiv preprint arXiv:2501.08828*, 2025a.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. *arXiv preprint arXiv:2505.16470*, 2025b.

Kuicai Dong, Shurui Huang, Fangda Ye, Wei Han, Zhi Zhang, Dexun Li, Wenjun Li, Qu Yang, Gang Wang, Yichao Wang, et al. Doc-researcher: A unified system for multimodal document parsing and deep research. *arXiv preprint arXiv:2510.21603*, 2025c.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, et al. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4026–4037, 2025. doi: 10.1109/CVPR52734.2025.00381. URL https://doi.org/10.1109/CVPR52734.2025.00381.

Omar El Bachyr, Fred Philippy, Laura Maria Bernardy, Saad Ezzini, Jacques Klein, and Tegawende Bissyande. LëtzCross: A cross-lingual page-level benchmark for multimodal retrieval over luxembourgish documents. *arXiv preprint arXiv:2608.21714*, 2026. URL https://arxiv.org/abs/2608.21714.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025.

Dongyang Fan, Bettina Messmer, and Martin Jaggi. Towards an empirical understanding of moe design choices. *arXiv preprint arXiv:2402.13089*, 2024.

Yingqi Fan, Xuan Lu, Anhao Zhao, Junlong Tong, Ping Nie, Kai Zou, Yunpu Ma, Wei Zhang, and Xiaoyu Shen. minireranker: Efficient multimodal reranking through visual cache reuse and interaction sparsity. *arXiv preprint arXiv:2606.10759*, 2026. URL https://arxiv.org/abs/2606.10759.

Yongqi Fan, Xiaoyang Chen, Dezhi Ye, Jie Liu, Haijin Liang, Jin Ma, Ben He, Yingfei Sun, and Tong Ruan. Tfrank: Think-free reasoning enables practical pointwise llm ranking. *arXiv preprint arXiv:2508.09539*, 2025.

Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, et al. Cognitive kernel-pro: A framework for deep research agents and agent foundation models training. *arXiv preprint arXiv:2508.00414*, 2025.

Yan Fang, Jingtao Zhan, Qingyao Ai, Jiaxin Mao, Weihang Su, Jia Chen, and Yiqun Liu. Scaling laws for dense retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1339–1349, 2024. doi: 10.1145/3626772.3657743. URL https://doi.org/10.1145/3626772.3657743.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Chenghan Fu, Daoze Zhang, Yukang Lin, Zhanheng Nie, Xiang Zhang, Jianyu Liu, Yueran Liu, Wanxian Guan, Pengjie Wang, Jian Xu, et al. Moon embedding: Multimodal representation learning for e-commerce search advertising. *arXiv preprint arXiv:2511.11305*, 2025.

Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*, 2025a.

Wensheng Gan, Zhenyao Ning, Zhenlian Qi, and Philip S Yu. Mixture of experts (moe): A big data perspective. *arXiv preprint arXiv:2501.16352*, 2025b.

Sensen Gao, Shanshan Zhao, Xu Jiang, Lunhao Duan, Yong Xien Chng, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, Jia-Wang Bian, and Mingming Gong. Scaling beyond context: A survey of multimodal retrieval-augmented generation for document understanding. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pp. 4458–4489, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.204. URL https://aclanthology.org/2026.acl-long.204/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

David Golchinfar. Sauerkrautlm-colpali: Multi-vector vision retrieval models, 2025. URL https://github.com/VAGOsolutions/sauerkrautlm-colpali.

Ziyu Gong, Chengcheng Mai, and Yihua Huang. Mhier-rag: Multi-modal rag for visual-rich document question-answering via hierarchical and multi-granularity reasoning. *arXiv preprint arXiv:2508.00579*, 2025.

Geonmo Gu, Byeongho Heo, Jaemyung Yu, Jaehui Hwang, Taekyung Kim, Sangmin Lee, HeeJae Jun, Yoohoon Kang, Sangdoo Yun, and Dongyoon Han. Muco: Multi-turn contrastive learning for multimodal embedding model. *arXiv preprint arXiv:2602.06393*, 2026.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2860–2869, 2025a. doi: 10.1145/3746027.3754845. URL https://doi.org/10.1145/3746027.3754845.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning, 2025b. URL https://arxiv.org/abs/2510.13515.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, et al. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. *arXiv preprint arXiv:2506.18902*, 2025.

Hao Guo, Xugong Qin, Jun Jie Ou Yang, Peng Zhang, Gangyan Zeng, Yubo Li, and Hailun Lin. Towards natural language-based document image retrieval: New dataset and benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29722–29732, 2025. doi: 10.1109/CVPR52734.2025.02767. URL https://doi.org/10.1109/CVPR52734.2025.02767.

Lixuan Guo, Yifei Wang, Tiansheng Wen, Yifan Wang, Aosong Feng, Bo Chen, Stefanie Jegelka, and Chenyu You. Csrv2: Unlocking ultra-sparse embeddings. *arXiv preprint arXiv:2602.05735*, 2026.

Ibtihaal M Hameed, Sadiq H Abdulhussain, and Basheera M Mahmmod. Content-based image retrieval: A review of recent trends. *Cogent Engineering*, 8(1):1927469, 2021. doi: 10.1080/23311916.2021.1927469. URL https://doi.org/10.1080/23311916.2021.1927469.

Simeng Han, Frank Palma Gomez, Tu Vu, Zefei Li, Daniel Cer, Hansi Zeng, Chris Tar, Arman Cohan, and Gustavo Hernandez Abrego. Ateb: Evaluating and improving advanced nlp tasks for text embedding models. *arXiv preprint arXiv:2502.16766*, 2025a.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025b.

Hans William Alexander Hanley and Zakir Durumeric. Hierarchical level-wise news article clustering via multilingual matryoshka embeddings. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2476–2492, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.124. URL https://aclanthology.org/2025.acl-long.124/.

Xiangzhao Hao, Shijie Wang, Tianyu Yang, Tianyue Wang, Haiyun Guo, and Jinqiao Wang. Trace: Task-adaptive reasoning and representation learning for universal multimodal retrieval. *arXiv preprint arXiv:2603.02929*, 2026.

Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. Plume: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026.

Lexiang Hu, Youze Xue, Dian Li, Gang Liu, and Zhouchen Lin. Adaptive global and fine-grained perceptual fusion for mllm embeddings compatible with hard negative amplification. *arXiv preprint arXiv:2602.05729*, 2026a.

Lexiang Hu, Yanzhao Zhang, Mingxin Li, Dingkun Long, Yikang Li, Fuwei Zhang, Yisen Wang, and Zhouchen Lin. VisDocAgentBench: Benchmarking agents for visually rich document retrieval. *arXiv preprint arXiv:2608.17889*, 2026b. URL https://arxiv.org/abs/2608.17889.

Ruofan Hu, Menghui Zhu, Jieming Zhu, Bo Chen, Shengyang Xu, Minjie Hong, Xiaoda Yang, Sashuai Zhou, Li Tang, Tao Jin, and Zhou Zhao. DocRetriever: A plug-and-play framework for multimodal document retrieval with comprehensive benchmark. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026c. URL https://arxiv.org/abs/2605.30027. arXiv:2605.30027.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. *Advances in Neural Information Processing Systems*, 37:50168–50188, 2024. doi: 10.52202/079017-1588. URL https://doi.org/10.52202/079017-1588.

Jerry Huang, Siddarth Madala, Cheng Niu, Julia Hockenmaier, and Tong Zhang. Contextual relevance and adaptive sampling for llm-based document reranking. *arXiv preprint arXiv:2511.01208*, 2025a.

Kung-Hsiang Huang, Hou Pong Chan, May Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2550–2568, 2025b. doi: 10.1109/TKDE.2024.3513320. URL https://doi.org/10.1109/TKDE.2024.3513320.

Xin Huang and Kye Min Tan. Beyond text: Unlocking true multimodal, end-to-end rag with tomoro colqwen3. Hugging Face model card, 2025. URL https://huggingface.co/TomoroAI/tomoro-colqwen3-embed-4b.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025c.

Jiahao Huo, Yu Huang, Yibo Yan, Ye Pan, Kening Zheng, Wei-Chieh Huang, Yi Cao, Mingdong Ou, Philip S. Yu, and Xuming Hu. Causalembed: Auto-regressive multi-vector generation in latent space for visual document embedding. *arXiv preprint arXiv:2601.21262*, 2026.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. Characterizing deep research: A benchmark and formal definition. *arXiv preprint arXiv:2508.04183*, 2025.

Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. Llm-select: Feature selection with large language models. *arXiv preprint arXiv:2407.02694*, 2024.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *arXiv preprint arXiv:2408.16672*, 2024.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. Reasoningrank: Teaching student models to rank through reasoning-based knowledge distillation. *arXiv preprint arXiv:2410.05168*, 2024.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.181. URL https://aclanthology.org/2024.findings-emnlp.181/.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024b.

Zi-Han Jiang, Chien-Wei Lin, Wei-Hua Li, Hsuan-Tung Liu, Yi-Ren Yeh, and Chu-Song Chen. Relation-rich visual document generator for visual information extraction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14449–14459, 2025. doi: 10.1109/CVPR52734.2025.01347. URL https://doi.org/10.1109/CVPR52734.2025.01347.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024c.

Jina AI. jina-reranker-m0: Multilingual Multimodal Document Reranker. Hugging Face model card, 2025. URL https://huggingface.co/jinaai/jina-reranker-m0.

Stefan Josef. Learn to pool: Lightweight fine-tuning for flexible multi-vector compression. *arXiv preprint arXiv:2607.06036*, 2026. URL https://arxiv.org/abs/2607.06036.

Ashutosh Joshi, Sheikh Muhammad Sarwar, Samarth Varshney, Sreyashi Nag, Shrivats Agrawal, and Juhi Naik. Reaper: Reasoning based retrieval planning for complex rag systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4621–4628, 2024. doi: 10.1145/3627673.3680087. URL https://doi.org/10.1145/3627673.3680087.

Heethanjan Kanagalingam, Thenukan Pathmanathan, Mokeeshan Vathanakumar, Basim Azam, Sarah Monazam Erfani, and Naveed Akhtar. Invoice haystack: Benchmarking document retrieval and visual question answering under strong visual homogeneity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2606.25343. arXiv:2606.25343.

Wenjun Ke, Yifan Zheng, Yining Li, Hengyuan Xu, Dong Nie, Peng Wang, and Yao He. Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. *ACM Trans. Inf. Syst.*, 44(1), November 2026. ISSN 1046-8188. doi: 10.1145/3768156. URL https://doi.org/10.1145/3768156.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Juyeon Kim, Geon Lee, Dongwon Choi, Taeuk Kim, and Kijung Shin. Hybrid-vector retrieval for visually rich documents: Combining single-vector efficiency and multi-vector accuracy. *arXiv preprint arXiv:2510.22215*, 2025a.

Wongyu Kim, Hochang Lee, Sanghak Lee, Yoonsung Kim, and Jaehyun Park. Let multimodal embedders learn when to augment query via adaptive query augmentation. *arXiv preprint arXiv:2511.02358*, 2025b.

Adithya S Kolavi and Vyoman Jain. M3dr: Towards universal multilingual multimodal document retrieval. *arXiv preprint arXiv:2512.03514*, 2025.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W., Fuzheng Zhang, and Guorui Zhou. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*, 2025.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022. doi: 10.52202/068431-2192. URL https://doi.org/10.52202/068431-2192.

Riwei Lai, Li Chen, Weixin Chen, and Rui Chen. Matryoshka representation learning for recommendation. *arXiv preprint arXiv:2406.07432*, 2024.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*, 2025a.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*, 2025b.

Carlos Lassance, Maroua Maachou, Joohee Park, and Stéphane Clinchant. A study on token pruning for colbert. *arXiv preprint arXiv:2112.06540*, 2021.

Carlos Lassance, Maroua Maachou, Joohee Park, and Stéphane Clinchant. Learned token pruning in contextualized late interaction over bert (colbert). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2232–2236, 2022. doi: 10.1145/3477495.3531835. URL https://doi.org/10.1145/3477495.3531835.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.

Jaehoon Lee, Sohyun Kim, Wanggeun Park, Geon Lee, Seungkyung Kim, and Minyoung Lee. Sds kopub vdr: A benchmark dataset for visual document retrieval in korean public documents. *arXiv preprint arXiv:2511.04910*, 2025a.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025b.

Hongyang Lei, Xiaolong Cheng, Qi Qin, Dan Wang, Kun Fan, Huazhen Huang, Qingqing Gu, Yetao Wu, Zhonglin Jiang, Yong Chen, and Luo Ji. M3-JEPA: Multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture. *arXiv preprint arXiv:2409.05929*, 2024.

Da Li, Yuxiao Luo, Keping Bi, Jiafeng Guo, Wei Yuan, Biao Yang, Yan Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Compressing then matching: An efficient pre-training paradigm for multimodal embedding. *arXiv preprint arXiv:2511.08480*, 2025a.

Hanxi Li. CHaystack: Benchmarking chinese document retrieval and VQA. *arXiv preprint arXiv:2607.24760*, 2026. URL https://arxiv.org/abs/2607.24760.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL https://aclanthology.org/2024.acl-long.775/.

Maoliang Li, Ke Li, Yaoyang Liu, Jiayu Chen, Zihao Zheng, Yinjun Wu, Chenchen Liu, and Xiang Chen. Mirage: Runtime scheduling for multi-vector image retrieval with hierarchical decomposition. *arXiv preprint arXiv:2510.08976*, 2025b. URL https://arxiv.org/abs/2510.08976.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

Qi Li, Yanzhe Zhao, Yongxin Zhou, Yameng Wang, Yandong Yang, Yuanjia Zhou, and Jinxiang Liu. Magic-mm-embedding: Towards visual-token-efficient universal multimodal embedding with mllms. *arXiv preprint arXiv:2602.05275*, 2026b.

Siting Li, Xiang Gao, and Simon Shaolei Du. Highlighting what matters: Promptable embeddings for attribute-focused image retrieval. *arXiv preprint arXiv:2505.15877*, 2025c.

Weiqing Li, Jinyue Guo, Yaqi Wang, Haiyang Xiao, Yuewei Zhang, Guohua Liu, and Hao Henry Wang. Evo-retriever: Llm-guided curriculum evolution with viewpoint-pathway collaboration for multimodal document retrieval. *arXiv preprint arXiv:2603.16455*, 2026c.

Xiangyang Li, Kuicai Dong, Yi Quan Lee, Wei Xia, Hao Zhang, Xinyi Dai, Yasheng Wang, and Ruiming Tang. Coir: A comprehensive benchmark for code information retrieval models, 2024b. URL https://arxiv.org/abs/2407.02883.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2d matryoshka sentence embeddings. *arXiv preprint arXiv:2402.14776*, 2024c.

Xiaojie Li, Chu Li, Shi-Zhe Chen, and Xi Chen. U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. *arXiv preprint arXiv:2507.14902*, 2025d.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Guanting Dong, Jiajie Jin, Yinuo Wang, Hao Wang, Yutao Zhu, Ji-Rong Wen, Yuan Lu, et al. Deepagent: A general reasoning agent with scalable toolsets. *arXiv preprint arXiv:2510.21618*, 2025e.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025f.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15546–15555, 2024d. doi: 10.1109/CVPR52733.2024.01472. URL https://doi.org/10.1109/CVPR52733.2024.01472.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *arXiv preprint arXiv:2507.09477*, 2025g.

Yunxin Li, Xinyu Chen, Shenyuan Jiang, Haoyuan Shi, Zhenyu Liu, Xuanyu Zhang, Nanhao Deng, Zhenran Xu, Yicheng Ma, Meishan Zhang, et al. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*, 2025h.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025i. doi: 10.1109/TPAMI.2025.3532688. URL https://doi.org/10.1109/TPAMI.2025.3532688.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, et al. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*, 2025j.

Ziyue Li and Tianyi Zhou. Your mixture-of-experts llm is secretly an embedding model for free. *arXiv preprint arXiv:2410.10814*, 2024.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4038–4049, 2025. doi: 10.1109/CVPR52734.2025.00382. URL https://doi.org/10.1109/CVPR52734.2025.00382.

Adam Lilja, Junsheng Fu, Erik Stenborg, and Lars Hammarstrand. Localization is all you evaluate: Data leakage in online mapping datasets and how to fix it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22150–22159, 2024. doi: 10.1109/CVPR52733.2024.02091. URL https://doi.org/10.1109/CVPR52733.2024.02091.

Lin Lin, Jiefeng Long, Zhihe Wan, Yuchi Wang, Dingkang Yang, Shuang Yang, Yueyang Yao, Xu Chen, Zirui Guo, Shengqiang Li, et al. Sail-embedding technical report: Omni-modal embedding foundation model. *arXiv preprint arXiv:2510.12709*, 2025a.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024a.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*, 2024b.

Yijie Lin, Guofeng Ding, Haochen Zhou, Haobin Li, Mouxing Yang, and Xi Peng. Ark: A dual-axis multimodal retrieval benchmark along reasoning and knowledge. *arXiv preprint arXiv:2602.09839*, 2026.

Zhiyu Lin, Yifei Gao, Xian Zhao, Yunfan Yang, and Jitao Sang. Mind with eyes: from language reasoning to multimodal reasoning. *arXiv preprint arXiv:2503.18071*, 2025b.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging mllm reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025a.

Keliang Liu, Zizhi Chen, Mingcheng Li, Jingqun Tang, Dingkang Yang, and Lihua Zhang. Resolving evidence sparsity: Agentic context engineering for long-document understanding. *arXiv preprint arXiv:2511.22850*, 2025b.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2781–2790, 2025c. doi: 10.1145/3746027.3754761. URL https://doi.org/10.1145/3746027.3754761.

Qi Liu, Gang Guo, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, and Zhao Cao. An analysis on matching mechanisms and token pruning for late-interaction models. *ACM Transactions on Information Systems*, 42(5):1–28, 2024. doi: 10.1145/3639818. URL https://doi.org/10.1145/3639818.

Qianying Liu, Xiao Liang, Zhiqiang Zhang, Zhongfei Qing, Fengfan Zhou, Yibo Chen, Xu Tang, Yao Hu, and Paul Henderson. Rematch: Boosting representation through matching for multimodal retrieval. *arXiv preprint arXiv:2511.19278*, 2025d.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4015–4025, 2025e. doi: 10.1109/CVPR52734.2025.00380. URL https://doi.org/10.1109/CVPR52734.2025.00380.

Yuxiang Liu, Tian Wang, Gourab Kundu, Tianyu Cao, Guang Cheng, Zhen Ge, Jianshu Chen, Qingjun Cui, and Trishul Chilimbi. Exploring reasoning-infused text embedding with large language models for zero-shot dense retrieval. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4981–4985, 2025f. doi: 10.1145/3746252.3760855. URL https://doi.org/10.1145/3746252.3760855.

Zheng Liu, Chaofan Li, Shitao Xiao, Chaozhuo Li, Defu Lian, and Yingxia Shao. Matryoshka re-ranker: A flexible re-ranking architecture with configurable depth and width. *arXiv preprint arXiv:2501.16302*, 2025g.

Zheng Liu, Ze Liu, Zhengyang Liang, Junjie Zhou, Shitao Xiao, Chao Gao, Chen Jason Zhang, and Defu Lian. Any information is just worth one single screenshot: Unifying search with visualized information retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19238–19261, Vienna, Austria, July 2025h. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.943. URL https://aclanthology.org/2025.acl-long.943/.

Zhuchenyang Liu, Ziyu Hu, Yao Zhang, and Yu Xiao. Structural anchor pruning: Training-free multi-vector compression for visual document retrieval. *arXiv preprint arXiv:2601.20107*, 2026a. URL https://arxiv.org/abs/2601.20107.

Zhuchenyang Liu, Ziyi Wang, Yao Zhang, and Yu Xiao. DistilVDR: A compact end-to-end visual document retriever via dual-student distillation. *arXiv preprint arXiv:2608.10636*, 2026b. URL https://arxiv.org/abs/2608.10636.

Zhuchenyang Liu, Yao Zhang, and Yu Xiao. Nanovdr: Distilling a 2b vision-language retriever into a 70m text-only encoder for visual document retrieval. *arXiv preprint arXiv:2603.12824*, 2026c.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4427–4447, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.251. URL https://aclanthology.org/2025.findings-naacl.251/.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. Vidore v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*, 2026.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.373. URL https://aclanthology.org/2024.emnlp-main.373/.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024b. doi: 10.52202/079017-3041. URL https://doi.org/10.52202/079017-3041.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, et al. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. *arXiv preprint arXiv:2506.04997*, 2025.

Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval. *arXiv preprint arXiv:2505.17166*, 2025.

Ailar Mahdizadeh, Aria Salari, Sohail Rajabi, Shahriar Mirabbasi, Panos Nasiopoulos, and Alireza Morsali. Coverage matters: MarginMerge for compressing multi-vector visual document retrievers. *arXiv preprint arXiv:2608.02969*, 2026. URL https://arxiv.org/abs/2608.02969.

Ahmed Masry and Enamul Hoque. Colflor: Towards bert-size vision-language document retrieval models. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/MLSP62443.2025.11204231. URL https://doi.org/10.1109/MLSP62443.2025.11204231.

Ahmed Masry, Megh Thakkar, Patrice Bechard, Sathwik Tejaswi Madhusudhan, Rabiul Awal, Shambhavi Mishra, Akshay Kalkunte Suresh, Srivatsava Daruru, Enamul Hoque, Spandana Gella, Torsten Scholak, and Sai Rajeswar. ColMate: Contrastive late interaction and masked text for multimodal document retrieval. In Saloni Potdar, Lina Rojas-Barahona, and Sebastien Montella (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2071–2080, Suzhou (China), November 2025. Association for Computational Linguistics. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.145. URL https://aclanthology.org/2025.emnlp-industry.145/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225. URL https://doi.org/10.1109/WACV48630.2021.00225.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2582–2591, 2022. doi: 10.1109/WACV51458.2022.00264. URL https://doi.org/10.1109/WACV51458.2022.00264.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 6(12):1467–1477, 2024. doi: 10.1038/s42256-024-00925-4. URL https://doi.org/10.1038/s42256-024-00925-4.

Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, et al. Nemotron colembed v2: Top-performing late interaction embedding models for visual document retrieval. *arXiv preprint arXiv:2602.03992*, 2026.

Alexander Most, Joseph Winjum, Manish Bhattarai, Shawn M. Jones, Nishath Rajiv Ranasinghe, Ayan Biswas, and Dan O'Malley. Lost in ocr translation? vision-based approaches to robust document retrieval. In *Proceedings of the 2025 ACM Symposium on Document Engineering*, pp. 13:1–13:10, 2025. doi: 10.1145/3704268.3742698. URL https://doi.org/10.1145/3704268.3742698.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, and Giovanni Semeraro. xvlm2vec: Adapting lvlm-based embedding models to multilinguality using self-knowledge distillation. *arXiv preprint arXiv:2503.09313*, 2025.

Omer Nacar and Anis Koubaa. Enhancing semantic similarity understanding in arabic nlp with nested embedding learning. In *Generative AI and Large Language Models: Opportunities, Challenges, and Applications: Volume 1*, pp. 179–216. Springer, 2025. doi: 10.1007/978-3-031-90573-5_6. URL https://doi.org/10.1007/978-3-031-90573-5_6.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben-Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your vlm an efficient reader. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29005–29015, 2025. doi: 10.1109/CVPR52734.2025.02701. URL https://doi.org/10.1109/CVPR52734.2025.02701.

Pranav Nair, Puranjay Datta, Jeff Dean, Prateek Jain, and Aditya Kusupati. Matryoshka quantization. *arXiv preprint arXiv:2502.06786*, 2025.

Nomic AI. Nomic embed multimodal: Interleaved text, image, and screenshots for visual document retrieval. Hugging Face model card, 2025. URL https://huggingface.co/nomic-ai/colnomic-embed-multimodal-3b.

Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. Nomic embed vision: Expanding the latent space. *arXiv preprint arXiv:2406.18587*, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenSearch-AI. Ops-MM-embedding-v1-2B. Hugging Face model card, 2025. URL https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B.

OpenSearch-AI. Ops-Colqwen3: State-of-the-Art Multimodal Embedding Model for Visual Document Retrieval. Hugging Face model card, 2026. URL https://huggingface.co/OpenSearch-AI/Ops-Colqwen3-4B.

Radek Osmulski, Gabriel de Souza P Moreira, Ronay Ak, Mengyao Xu, Benedikt Schifferer, and Even Oldridge. Miracl-vision: A large, multilingual, visual document retrieval benchmark. *arXiv preprint arXiv:2505.11651*, 2025.

Yassine Ouali, Adrian Bulat, Alexandros Xenos, Anestis Zaganidis, Ioannis Maniadis Metaxas, Brais Martinez, and Georgios Tzimiropoulos. Vladva: Discriminative fine-tuning of lvlms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4101–4111, 2025. doi: 10.1109/CVPR52734.2025.00388. URL https://doi.org/10.1109/CVPR52734.2025.00388.

Tim Pearce, Tabish Rashid, Dave Bignell, Raluca Georgescu, Sam Devlin, and Katja Hofmann. Scaling laws for pre-training agents and world models. *arXiv preprint arXiv:2411.04434*, 2024.

Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. Answer is all you need: Instruction-following text embedding via answering the question. *arXiv preprint arXiv:2402.09642*, 2024.

Xiangyu Peng, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and Chien-Sheng Wu. Unidoc-bench: A unified benchmark for document-centric multimodal rag. *arXiv preprint arXiv:2510.03663*, 2025.

Roi Pony, Adi Raz Goldfarb, Oshri Naparstek, Idan Friedman, Udi Barzelay, and Eli Schwartz. Col-bandit: Query-time top-k estimation for late-interaction retrieval. *arXiv preprint arXiv:2602.02827*, 2026.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, et al. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*, 2025.

Hanxiang Qin, Alexander Martin, Rohan Jha, Chunsheng Zuo, Reno Kriz, and Benjamin Van Durme. Multi-vector index compression in any modality. *arXiv preprint arXiv:2602.21202*, 2026.

Alexander Rombach and Peter Fettke. Deep learning based key information extraction from business documents: Systematic literature review. *ACM Computing Surveys*, 58(2):45:1–45:37, 2026. doi: 10.1145/3749369. URL https://doi.org/10.1145/3749369.

Rohit Kumar Salla, Manoj Saravanan, and Ramya Manasa Amancherla. CrossQ: Task-aligned cross-token conditional quantization for late interaction retrieval. *arXiv preprint arXiv:2608.19204*, 2026. URL https://arxiv.org/abs/2608.19204.

Chris Samarinas and Hamed Zamani. Distillation and refinement of reasoning in small language models for document re-ranking. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 430–435, 2025. doi: 10.1145/3731120.3744613. URL https://doi.org/10.1145/3731120.3744613.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5058–5070, 2025. URL https://aclanthology.org/2025.coling-main.338/.

Abdellatif Sassioui, Rachid Benouini, Yasser El Ouargui, Mohamed El-Kamili, Meriyem Chergui, and Mohammed Ouzzif. Visually-rich document understanding: concepts, taxonomy and challenges. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1–7. IEEE, 2023. doi: 10.1109/WINCOM59760.2023.10322990. URL https://doi.org/10.1109/WINCOM59760.2023.10322990.

Rahul Seetharaman, Kaustubh D Dhole, and Aman Bansal. Insertrank: Llms can reason over bm25 scores to improve listwise reranking. *arXiv preprint arXiv:2506.14086*, 2025.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*, 2025.

Wenxuan Shen, Mingjia Wang, Yaochen Wang, Dongping Chen, Junjie Yang, Yao Wan, and Weiwei Lin. Are we on the right way for assessing document retrieval-augmented generation? *arXiv preprint arXiv:2508.03644*, 2025.

Connor Shorten, Augustas Skaburskas, Daniel M Jones, Charles Pierse, Roberto Esposito, John Trengrove, Etienne Dilocker, and Bob van Luijt. Irpapers: A visual document benchmark for scientific retrieval and question answering. *arXiv preprint arXiv:2602.17687*, 2026.

Abhinav Shukla, Sai Vemprala, Aditya Kusupati, and Ashish Kapoor. Matmamba: A matryoshka state space model. *arXiv preprint arXiv:2410.06718*, 2024.

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V. Vasilakos. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025.

Dingjie Song, Sicheng Lai, Mingxuan Wang, Shunian Chen, Lichao Sun, and Benyou Wang. Both text and images leaked! a systematic analysis of data contamination in multimodal LLM. *arXiv preprint arXiv:2411.03823*, 2024.

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin Zhang, and Meng Wang. How to bridge the gap between modalities: Survey on multimodal large language model. *IEEE Transactions on Knowledge and Data Engineering*, 2025. doi: 10.1109/TKDE.2025.3527978. URL https://doi.org/10.1109/TKDE.2025.3527978.

Zhivar Sourati, Zheng Wang, Marianne Menglin Liu, Yazhe Hu, Mengqing Guo, Sujeeth Bharadwaj, Kyu Han, Tao Sheng, Sujith Ravi, Morteza Dehghani, et al. Lad-rag: Layout-aware dynamic rag for visually-rich document understanding. *arXiv preprint arXiv:2510.07233*, 2025.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *arXiv preprint arXiv:2505.13965*, 2025a.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025b.

Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*, 2020.

Duolin Sun, Meixiu Long, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, Jiahai Wang, and Jinjie Gu. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*, 2025a.

Hao Sun, Yingyan Hou, Jiayan Guo, Bo Wang, Chunyu Yang, Jinsong Ni, and Yan Zhang. Unveil: Unified visual-textual integration and distillation for multi-modal document retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23935–23945. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.1166. URL https://aclanthology.org/2025.acl-long.1166/.

Yiqun Sun, Pengfei Wei, and Lawrence B. Hsieh. Very efficient listwise multimodal reranking for long documents. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.11864. arXiv:2605.11864.

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Sen Mei, Chi Chen, and Maosong Sun. VisRAG2.0: Mitigating visual hallucinations via evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025c.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6088–6109, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.310. URL https://aclanthology.org/2025.naacl-long.310/.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24827–24837, 2025. doi: 10.1109/CVPR52734.2025.02312. URL https://doi.org/10.1109/CVPR52734.2025.02312.

Yixuan Tang and Yi Yang. Do we need domain-specific embedding models? an empirical investigation. *arXiv preprint arXiv:2409.18511*, 2024.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19254–19264, 2023. doi: 10.1109/CVPR52729.2023.01845. URL https://doi.org/10.1109/CVPR52729.2023.01845.

Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview. *arXiv preprint arXiv:2412.12591*, 2024.

Granite Vision Team, Leonid Karlinsky, Assaf Arbelle, Abraham Daniels, Ahmed Nassar, Amit Alfassi, Bo Wu, Eli Schwartz, Dhiraj Joshi, Jovana Kondic, et al. Granite vision: a lightweight, open-source multimodal model for enterprise intelligence. *arXiv preprint arXiv:2502.09927*, 2025a.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025b.

Paul Teiletche, Quentin Macé, Max Conti, Antonio Loison, Gautier Viaud, Pierre Colombo, and Manuel Faysse. Modernvbert: Towards smaller visual document retrievers. *arXiv preprint arXiv:2510.01149*, 2025.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, and Bhuwan Dhingra. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *arXiv preprint arXiv:2505.11293*, 2025.

Pascal Tilli and Mohsen Mesgar. Beyond bag-of-patches: Learning global layout via textual supervision for late-interaction visual document retrieval. *arXiv preprint arXiv:2605.08421*, 2026. URL https://arxiv.org/abs/2605.08421.

Anyang Tong, Xiang Niu, ZhiPing Liu, Chang Tian, Yanyan Wei, Zenglin Shi, and Meng Wang. Hkrag: Holistic knowledge retrieval-augmented generation over visually-rich documents. *arXiv preprint arXiv:2511.20227*, 2025.

Yu-Che Tsai, Kuan-Yu Chen, Yuan-Chi Li, Yuan-Hao Chen, Ching-Yu Tsai, and Shou-De Lin. Let llms speak embedding languages: Generative text embeddings via iterative contrastive refinement. *arXiv preprint arXiv:2509.24291*, 2025.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

Chetan Verma, Aditya Srinivas Timmaraju, Cho-Jui Hsieh, Suyash Damle, Ngot Bui, Yang Zhang, Wen Chen, Xin Liu, Prateek Jain, and Inderjit S. Dhillon. Matryoshka model learning for improved elastic student models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4935–4944, 2025. doi: 10.1145/3711896.3737245. URL https://doi.org/10.1145/3711896.3737245.

Haitian Wang, Ruoxi Sun, Quantong Qiu, Juntao Li, Junhui Li, Hua Chen, Jinxiong Chang, and Min Zhang. MMLongEmbed: Benchmarking multimodal embedding models in long-context scenarios. *arXiv preprint arXiv:2606.14747*, 2026a. URL https://arxiv.org/abs/2606.14747.

Haocheng Wang, Tongkun Guan, Wei Shen, and Xiaokang Yang. VaRS-Doc: Interpretation-aware variant representations via latent self-probing for visual document retrieval. *arXiv preprint arXiv:2608.01211*, 2026b. URL https://arxiv.org/abs/2608.01211.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.166. URL https://aclanthology.org/2025.naacl-long.166/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*, 2025b.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*, 2025c.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*, 112(11):1716–1754, 2024b. doi: 10.1109/JPROC.2024.3525147. URL https://doi.org/10.1109/JPROC.2024.3525147.

Wei-Yao Wang, Kazuya Tateishi, Qiyu Wu, Shusuke Takahashi, and Yuki Mitsufuji. Virtue: Visual-interactive text-image universal embedder. *arXiv preprint arXiv:2510.00523*, 2025d.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024c. doi: 10.52202/079017-0649. URL https://doi.org/10.52202/079017-0649.

Yuchi Wang, Dingkang Yang, Haiyang Yu, Weikang Bian, Jiefeng Long, Xiao Liang, Chao Feng, and Hongsheng Li. Mmemb-r1: Reasoning-enhanced multimodal embedding with pair-aware selection and adaptive control. *arXiv preprint arXiv:2604.06156*, 2026c. URL https://arxiv.org/abs/2604.06156.

Yueqi Wang, Zhenrui Yue, Huimin Zeng, Dong Wang, and Julian McAuley. Train once, deploy anywhere: Matryoshka representation learning for multimodal recommendation. *arXiv preprint arXiv:2409.16627*, 2024d.

Zhicheng Wang, Chen Ju, Xu Chen, Shuai Xiao, Jinsong Lan, Xiaoyong Zhu, Ying Chen, and Zhiguo Cao. Explore more, learn better: Parallel mllm embeddings under mutual information minimization. *arXiv preprint arXiv:2511.01588*, 2025e.

Zora Zhiruo Wang, Akari Asai, Xinyan Velocity Yu, Frank F. Xu, Yiqing Xie, Graham Neubig, and Daniel Fried. Coderag-bench: Can retrieval augment code generation?, 2024e. URL https://arxiv.org/abs/2406.14497.

Navve Wasserman, Oliver Heinimann, Yuval Golbari, Tal Zimbalist, Eli Schwartz, and Michal Irani. Docrerank: Single-page hard negative query generation for training multi-modal rag rerankers. *arXiv preprint arXiv:2505.22584*, 2025a.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*, 2025b.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer, 2024a. doi: 10.1007/978-3-031-73021-4_23. URL https://doi.org/10.1007/978-3-031-73021-4_23.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*, 2024b.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions, 2024.

Orion Weller, Benjamin Chang, Eugene Yang, Mahsa Yarmohammadi, Samuel Barham, Sean MacAvaney, Arman Cohan, Luca Soldaini, Benjamin Van Durme, and Dawn J. Lawrie. mfollowir: A multilingual benchmark for instruction following in retrieval. In *European Conference on Information Retrieval*, pp. 295–310. Springer, 2025a. doi: 10.1007/978-3-031-88711-6_19. URL https://doi.org/10.1007/978-3-031-88711-6_19.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*, 2025b.

Tiansheng Wen, Yifei Wang, Zequn Zeng, Zhong Peng, Yudi Su, Xinyang Liu, Bo Chen, Hongwei Liu, Stefanie Jegelka, and Chenyu You. Beyond matryoshka: Revisiting sparse coding for adaptive representation. *arXiv preprint arXiv:2503.01776*, 2025.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, et al. Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*, 2024a.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, Trevor Darrell, and David M Chan. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. *arXiv preprint arXiv:2407.13766*, 2024b.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14024–14045, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.708. URL https://aclanthology.org/2025.emnlp-main.708/.

Yebo Wu, Feng Liu, Ziwei Xie, Zhiyuan Liu, Changwang Zhang, Jun Wang, and Li Li. Tsembed: Unlocking task scaling in universal multimodal embeddings. *arXiv preprint arXiv:2603.04772*, 2026.

Haowen Xiang, Yibo Yan, Jiahao Huo, Yu Huang, Yi Cao, Mingdong Ou, and Xuming Hu. MM-Matryoshka: Towards budget-elastic visual document retrieval via a 2d multimodal matryoshka training framework. *arXiv preprint arXiv:2606.07654*, 2026. URL https://arxiv.org/abs/2606.07654.

Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *Nature Machine Intelligence*, 7(11):1823–1833, 2025a. doi: 10.1038/s42256-025-01137-0. URL https://doi.org/10.1038/s42256-025-01137-0.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*, 2024.

Chenghao Xiao, Hou Pong Chan, Hao Zhang, Weiwen Xu, Mahani Aljunied, and Yu Rong. Scaling language-centric omnimodal representation learning. *arXiv preprint arXiv:2510.11693*, 2025b.

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth C. Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22187–22198, 2025c. doi: 10.1109/ICCV51701.2025.02060. URL https://doi.org/10.1109/ICCV51701.2025.02060.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. *arXiv preprint arXiv:2509.18095*, 2025d.

Peijin Xie, Shun Qian, Bingquan Liu, Dexin Wang, Lin Sun, and Xiangzheng Zhang. Textlessrag: End-to-end visual document rag by speech without text. *arXiv preprint arXiv:2509.07538*, 2025.

Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Haike Xu and Tong Chen. Beyond sequential reranking: Reranker-guided search improves reasoning intensive retrieval. *arXiv preprint arXiv:2509.07163*, 2025.

Mengyao Xu, Gabriel Moreira, Ronay Ak, Radek Osmulski, Yauhen Babakhin, Zhiding Yu, Benedikt Schifferer, and Even Oldridge. Llama nemoretriever colembed: Top-performing text-image retrieval model. *arXiv preprint arXiv:2507.05513*, 2025a.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. Mm-r5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*, 2025b.

Youze Xue, Dian Li, and Gang Liu. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. *arXiv preprint arXiv:2506.02020*, 2025.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*, 2024.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11798–11827, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.614. URL https://aclanthology.org/2025.findings-acl.614/.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025b.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*, 2025c.

Yibo Yan, Guangwei Xu, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. *arXiv preprint arXiv:2509.23883*, 2025d.

Yibo Yan, Mingdong Ou, Yi Cao, Jiahao Huo, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Visual late chunking: An empirical study of contextual chunking for efficient visual document retrieval. *arXiv preprint arXiv:2604.10167*, 2026a.

Yibo Yan, Mingdong Ou, Yi Cao, Xin Zou, Jiahao Huo, Shuliang Liu, James Kwok, and Xuming Hu. Sculpting the vector space: Towards efficient multi-vector visual document retrieval via prune-then-merge framework. *arXiv preprint arXiv:2602.19549*, 2026b.

Yibo Yan, Mingdong Ou, Yi Cao, Xin Zou, Shuliang Liu, Jiahao Huo, Yu Huang, James Kwok, and Xuming Hu. Beyond the grid: Layout-informed multi-vector retrieval with parsed visual document representations. *arXiv preprint arXiv:2603.01666*, 2026c.

Binhan Yang, Yuxiang Zeng, Hengxin Zhang, Zhuanglin Zheng, Yunzhen Chi, Yongxin Tong, and Ke Xu. Unified and efficient approach for multi-vector similarity search. *arXiv preprint arXiv:2604.02815*, 2026. URL https://arxiv.org/abs/2604.02815.

Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*, 2025a.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, et al. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *arXiv preprint arXiv:2506.02454*, 2025b.

Jinsung Yoon, Rajarishi Sinha, Sercan O Arik, and Tomas Pfister. Matryoshka-adaptor: Unsupervised and supervised tuning for smaller embedding dimensions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10318–10336, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.576. URL https://aclanthology.org/2024.emnlp-main.576/.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. *arXiv preprint arXiv:2503.19900*, 2025a.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024.

Xinlei Yu, Chengming Xu, Zhangquan Chen, Yudong Zhang, Shilin Lu, Cheng Yang, Jiangning Zhang, Shuicheng Yan, and Xiaobin Hu. Visual document understanding and reasoning: A multi-agent collaboration framework with agent-wise adaptive test-time scaling. *arXiv preprint arXiv:2508.03404*, 2025b.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024a.

Biao Zhang, Lixin Chen, Tong Liu, and Bo Zheng. SMEC:rethinking matryoshka representation learning for retrieval embedding compression. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26209–26222, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1332. URL https://aclanthology.org/2025.emnlp-main.1332/.

Charles Zhang, Benji Peng, Xintian Sun, Qian Niu, Junyu Liu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Ming Liu, et al. From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models. *arXiv preprint arXiv:2411.05036*, 2024b.

Jianrui Zhang, Anirudh Sundara Rajan, Brandon Han, Soochahn Lee, Sukanta Ganguly, and Yong Jae Lee. Reasoning-augmented representations for multimodal retrieval. *arXiv preprint arXiv:2602.07125*, 2026a.

Junyuan Zhang, Qintong Zhang, Bin Wang, Linke Ouyang, Zichen Wen, Ying Li, Ka-Ho Chow, Conghui He, and Wentao Zhang. Ocr hinders rag: Evaluating the cascading impact of ocr on retrieval-augmented generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17443–17453, 2025b. doi: 10.1109/ICCV51701.2025.01620. URL https://doi.org/10.1109/ICCV51701.2025.01620.

Kun Zhang, Jingyu Li, Zhe Li, Jingjing Zhang, Fan Li, Yandong Liu, Rui Yan, Zihang Jiang, Nan Chen, Lei Zhang, et al. Composed multi-modal retrieval: A survey of approaches and applications. *arXiv preprint arXiv:2503.01334*, 2025c.

Ruichen Zhang, Shunpu Tang, Yinqiu Liu, Dusit Niyato, Zehui Xiong, Sumei Sun, Shiwen Mao, and Zhu Han. Toward agentic ai: generative information retrieval inspired intelligent communications and networking. *arXiv preprint arXiv:2502.16866*, 2025d.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. In *International Conference on Learning Representations (ICLR)*, 2026b. URL https://arxiv.org/abs/2510.09510.

Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025e.

Wenzheng Zhang, Xi Victoria Lin, Karl Stratos, Wen-tau Yih, and Mingda Chen. Imprag: Retrieval-augmented generation with implicit queries. *arXiv preprint arXiv:2506.02279*, 2025f.

Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558*, 2025g.

Xiantao Zhang. Roles of MLLMs in visually rich document retrieval for RAG: A survey. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 19–36, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.2. URL https://aclanthology.org/2025.ijcnlp-long.2/.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024c.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://doi.org/10.1162/tacl_a_00595.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025h.

Yimeng Zhang, Jiri Gesi, Ran Xue, Tian Wang, Ziyi Wang, Yuxuan Lu, Sinong Zhan, Huimin Zeng, Qingjun Cui, Yufan Guo, et al. See, think, act: Online shopper behavior simulation with vlm agents. *arXiv preprint arXiv:2510.19245*, 2025i.

Zhibo Zhang, Yang Xu, Kai Ming Ting, and Cam-Tu Nguyen. Llms meet isolation kernel: Lightweight, learning-free binary embeddings for fast retrieval. *arXiv preprint arXiv:2601.09159*, 2026c.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*, 2023.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, et al. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *arXiv preprint arXiv:2506.20923*, 2025.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*, 2025a.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025b.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. *arXiv preprint arXiv:2406.04292*, 2024.

Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, et al. Mr^2-bench: Going beyond matching to reasoning in multimodal retrieval. *arXiv preprint arXiv:2509.26378*, 2025a.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.935. URL https://aclanthology.org/2025.acl-long.935/.

Wengang Zhou, Houqiang Li, and Qi Tian. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*, 2017.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. Longembed: Extending embedding models for long context retrieval. *arXiv preprint arXiv:2404.12096*, 2024a.

Dawei Zhu, Rui Meng, Jiefeng Chen, Sujian Li, Tomas Pfister, and Jinsung Yoon. Doclens: A tool-augmented multi-agent framework for long visual document understanding. *arXiv preprint arXiv:2511.11552*, 2025a.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4857–4866, 2022. doi: 10.1145/3503161.3548422. URL https://doi.org/10.1145/3503161.3548422.

Fengbin Zhu, Zijing Cai, Yuzhe Wang, Pengyang Shao, Wenjie Wang, Fuli Feng, Richang Hong, and Tat-Seng Chua. Mure: Hierarchical multi-resolution encoding via vision-language models for visual document retrieval. *arXiv preprint arXiv:2603.13349*, 2026.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Haiyang Wu, and Shiqi Wang. Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. *Advances in Neural Information Processing Systems*, 2025b. doi: 10.52202/085713-5719. URL https://doi.org/10.52202/085713-5719.

Yichen Zhu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Retrieval-augmented embodied agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17985–17995, 2024b. doi: 10.1109/CVPR52733.2024.01703. URL https://doi.org/10.1109/CVPR52733.2024.01703.

Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers. *arXiv preprint arXiv:2509.24621*, 2025c.

Shengyao Zhuang, Shuai Wang, Fabio Zheng, Bevan Koopman, and Guido Zuccon. Starbucks-v2: Improved training for 2d matryoshka embeddings. *arXiv preprint arXiv:2410.13230*, 2024.

Haoyu Zuo, Yibo Yan, Xin Zou, Shuliang Liu, Yi Cao, Mingdong Ou, and Xuming Hu. AnchorFold: A focus-then-fold framework via recursive attention propagation for efficient multi-vector visual document retrieval. *arXiv preprint arXiv:2608.08732*, 2026. URL https://arxiv.org/abs/2608.08732.