ColMate : Contrastive Late Interaction and Masked Text for Multimodal Document Retrieval

Ahmed Masry[♣♡], Megh Thakkar[♣], Patrice Bechard [♣], Sathwik Tejaswi Madhusudhan[♣], Rabiul Awal[♣♠◇], Shambhavi Mishra[♣△], Akshay Kalkunte Suresh[♣], Srivatsava Daruru[♣], Enamul Hoque[♡], Spandana Gella[♣], Torsten Scholak[♣], Sai Rajeswar[♣*]

[♣]ServiceNow, [♡]York University, [♠]MILA - Quebec AI Institute,
[◇]Université de Montréal, [△]École de technologie supérieure

Abstract

Retrieval-augmented generation has proven practical when models require specialized knowledge or access to the latest data. However, existing methods for multimodal document retrieval often replicate techniques developed for text-only retrieval, whether in how they encode documents, define training objectives, or compute similarity scores. To address these limitations, we present ColMate, a document retrieval model that bridges the gap between multimodal representation learning and document retrieval. ColMate utilizes a novel OCR-based pretraining objective, a self-supervised masked contrastive learning objective, and a late interaction scoring mechanism more relevant to multimodal document structures and visual characteristics. ColMate obtains 3.61% improvements over existing retrieval models on the ViDoRe V2 benchmark, demonstrating stronger generalization to out-of-domain benchmarks.

1 Introduction

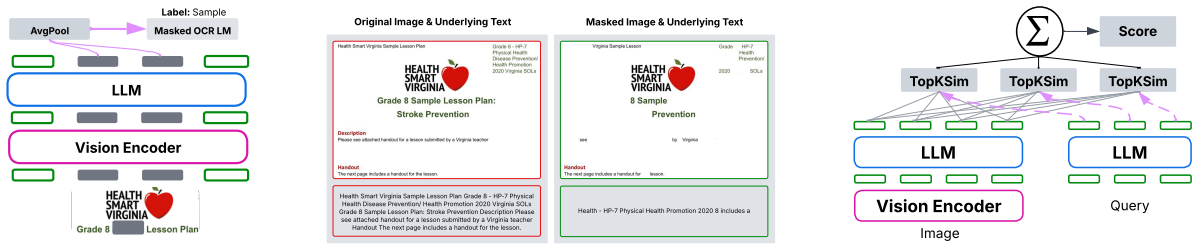
The information assimilated by LLMs (Large Language Models) during pretraining and stored in their parametric memory can get outdated with time (Zhang et al., 2024). LLMs also face challenges when tackling tasks requiring highly-specialized knowledge or scenarios requiring information not present in their training data, such as confidential information (Gao et al., 2024). RAG (Retrieval Augmented Generation) offers a solution to this issue by providing a way to provide external knowledge as context to the models (Lewis et al., 2020). Given its practicality, multimodal RAG has also been explored. However, compared to text-only retrieval, it involves retrieving documents comprising of rich information in the form of figures, charts, and tables along with text, increasing complexity and requiring understanding visual and spatial representations (Mei et al., 2025).

Despite these challenges, multimodal RAG has made notable progress with the development of models like ColPali (Faysse et al., 2025). Although representing meaningful progress, these methods exhibit several limitations: *(i)* They use pretrained VLMs such as PaliGemma (Beyer et al., 2024) to obtain visual representations; however, these models do not explicitly optimize output visual tokens during pretraining, as the autoregressive loss is applied only to subsequent text tokens, limiting their effectiveness for fine-grained visual retrieval. *(ii)* They rely heavily on supervised fine-tuning with annotated query–document pairs to achieve cross-modal alignment, which restricts their applicability in domains lacking such labeled data. *(iii)* Existing methods adopt late-interaction mechanisms such as MaxSim (Khattab and Zaharia, 2020), originally designed for text retrieval, which assume a one-to-one correspondence between query and document tokens. In visual contexts, this assumption breaks down, as patch-based image tokenization can fragment words or merge multiple words into a single patch, introducing noise in similarity computation during training and ultimately degrading retrieval performance. These shortcomings in the design of existing methods may limit their capabilities for multimodal document retrieval.

We address these limitations by introducing ColMate, a multimodal document retrieval model designed to capture the rich information embedded in visually dense documents. ColMate introduces three complementary components across pretraining, fine-tuning, and late-interaction: *(i)* a masked OCR language modeling objective that explicitly optimizes visual token representations during pretraining, *(ii)* a self-supervised contrastive learning objective that enables cross-modal alignment without relying on annotated query–document pairs, and *(iii)* a refined

*Correspondance to ahmed.masry@servicenow.com and sai.rajeswar@servicenow.com.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2071–2080
November 4-9, 2025



(a) Masked OCR LM (b) Masked Contrastive Learning (c) Training with TopKSim

Figure 1: Overview of ColMate's key components. *(a) Masked OCR Language Modeling (MOLM)* explicitly optimizes visual token representations by predicting masked OCR tokens during pretraining. *(b) Masked Contrastive Learning (MaskedCL)* enables self-supervised cross-modal alignment between masked text and document features when query–document pairs are unavailable. *(c) TopKSim* refines the late-interaction mechanism by averaging top-K similarity scores during training to reduce noise from patch-based tokenization in visual documents.

late-interaction mechanism, TopKSim, that alleviates training noise arising from patch-based tokenization in visual documents. Together, these components bridge the gap between existing retrieval approaches and recent advances in multimodal representation learning, yielding more robust and generalizable multimodal document retrieval.

ColMate improves performance over existing methods on both in-domain and out-of-domain ViDoRe benchmarks (Faysse et al., 2025; Macé et al., 2025), with particularly notable gains in out-of-domain generalization. The contributions of this work are summarized as follows: (i) we present ColMate, a multimodal document retrieval model that integrates three complementary components across pretraining, self-supervised fine-tuning, and late interaction; (ii) we demonstrate consistent performance improvements compared to existing methods across diverse domains; and (iii) we provide detailed ablations and analyses to quantify the contribution of each component. To support future research, we release the model weights publicly at https://huggingface.co/ahmed-masry/ColMate-3B.

2 Related Work

2.1 Multimodal Retrieval & Late Interaction

Multimodal retrieval models have significantly advanced through contrastive learning approaches that align visual and textual representations in shared embedding spaces (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). These models utilize contrastive learning techniques (Hadsell et al., 2006; Chen et al., 2020) and are trained on large-scale image-text datasets (Lin et al., 2014; Schuhmann et al., 2022) to enable cross-modal retrieval capabilities, facilitating downstream applications such as visual question answering (VQA) (Liu et al., 2023; Beyer et al., 2024) and multimodal RAG (Chen et al., 2022; Yasunaga et al., 2022). For efficient dense retrieval, late interaction mechanisms like ColBERT (Khattab and Zaharia, 2020) have proven effective by computing fine-grained token-level similarities between queries and documents, and recent works have extended these approaches to multimodal document retrieval, with ColPali (Faysse et al., 2025) directly applying ColBERT's MaxSim operation (Khattab and Zaharia, 2020) to vision-language models like PaliGemma (Beyer et al., 2024). However, this direct adaptation introduces noise during training because MaxSim assumes a one-to-one correspondence between query and document tokens. This assumption breaks down in visual contexts where patch-based tokenization can fragment words across multiple patches.

2.2 Visual Document Understanding & Pretraining

Visual document understanding focuses on enabling models to comprehend and process documents that often contain complex layouts, diverse fonts, and multimodal content such as images, tables, and charts. Models like DocFormer (Appalaraju et al., 2021), LayoutLMv3 (Huang et al., 2022), and DiT (Li et al., 2022) integrate textual content with layout and visual information, while approaches like StructTextV2 (Yu et al., 2023) and UniDoc (Gu et al., 2021) refine pretraining objectives specifically for document understanding through masked image and language modeling over structured layouts. Recent large-scale efforts like BigDocs (Rodriguez et al., 2025) have

demonstrated the critical importance of document-specific training data and pretraining for improving document comprehension capabilities. Despite these advances, most multimodal document retrieval methods build on general-purpose VLMs like PaliGemma (Beyer et al., 2024), whose pretraining objectives do not explicitly optimize visual token representations in the final layer. As a result, visual embeddings may not capture the fine-grained document structure and semantics, limiting retrieval effectiveness.

2.3 Visually Rich Document Retrieval.

Retrieving visually rich documents is challenging, as traditional OCR-based text retrieval struggles to capture layout information and visual semantics. Recent models such as DSE (Ma et al., 2024) and VisRAG (Yu et al., 2024) leverage vision–language models (VLMs) directly for document retrieval, reducing the need for complex OCR preprocessing pipelines. ColPali (Faysse et al., 2025) further extends late-interaction architectures like ColBERT (Khattab and Zaharia, 2020) to efficiently match query and document embeddings. Self-supervised objectives such as masked language modeling (Devlin et al., 2019) and contrastive learning (Hadsell et al., 2006; Chen et al., 2020) have proven effective for improving representation learning, with extensions like SpanBERT (Joshi et al., 2020) and hard-negative contrastive training (Robinson et al., 2021) enhancing robustness. However, existing multimodal document retrieval models rely predominantly on supervised fine-tuning with annotated query–document pairs, limiting scalability and generalization to domains where such data is scarce.

3 Methodology

As presented in Fig. 1, ColMate incorporates three novel components: *(i)* MOLM, an OCR-based masked language modeling training objective for improved feature representation, *(ii)* TopKSim, a late-interaction mechanism optimized for vision domain retrieval, and *(iii)* MaskedCL, a self-supervised contrastive learning objective for scenarios without query-image pairs.

3.1 Masked OCR Language Modeling (MOLM)

ColPali initializes from PaliGemma's pretrained weights (Beyer et al., 2024). However, the pretraining of PaliGemma and similar VLMs often does not directly optimize the vision tokens representations in the last layer, as the autoregressive loss is typically applied only to subsequent text tokens. Consequently, vision features may be suboptimal, hindering retrieval tasks that heavily rely on them.

To address this, we introduce a novel pretraining objective, *Masked OCR Language Modeling (MOLM)*, formulated as follows:

(1) Given an input document image, we obtain contextualized visual token embeddings $V = \{v_1, v_2, \ldots, v_N\} \subseteq \mathbb{R}^d$ from the LLM output. Suppose we have an OCR word token $\{w_j\}$ with its corresponding bounding box. Let $B_j \subseteq \{1, \ldots, N\}$ denote the indices of visual tokens that spatially overlap with the bounding box of word w_j.

(2) We randomly select 30% of the OCR tokens to mask, forming the set $\mathcal{M} \subset \{w_j\}$. For each masked word $w_j \in \mathcal{M}$, we compute a pooled visual representation over the tokens in B_j:

$$\bar{v}_j = \frac{1}{|B_j|} \sum_{i \in B_j} v_i$$

(3) Using these pooled embeddings, the model is trained with a masked language modeling objective to predict the masked OCR tokens:

$$\mathcal{L}_{\text{MOLM}} = - \sum_{w_j \in \mathcal{M}} \log p(w_j \mid \bar{v}_j, \theta)$$

where $p(w_j \mid \bar{v}_j, \theta)$ denotes the probability of predicting the token w_j from its visual context by the model with parameters θ.

By directly optimizing visual token embeddings through this OCR-based masking objective, MOLM significantly enriches the visual representations, thereby enhancing downstream multimodal retrieval performance.

3.2 TopKSim

Current late-interaction methods like MaxSim, originally designed for text-based retrieval models such as ColBERT (Khattab and Zaharia, 2020), are suboptimal when directly applied to document images (Faysse et al., 2025). This limitation stems from fundamental differences in how text and images are tokenized. In text, tokens usually align well with semantic units such as words. In contrast, visual tokens (derived from image patches)

may span multiple words or capture only parts of a word, depending on factors like patch size and font rendering. To address this, we propose *TopKSim*, a novel method that averages the top-K similarity scores during training instead of relying on the single maximum. The hyperparameter K controls the number of top-scoring document tokens considered for each query token. This approach reduces training noise and mitigates the over-reliance on single image patches, acting as a regularizer. This results in more robust retrieval and matching of query-image pairs.

Formally, given an encoded query $q = \{q_1, q_2, \ldots, q_n\} \subset \mathbb{R}^d$ and an encoded document $d = \{d_1, d_2, \ldots, d_m\} \subset \mathbb{R}^d$, we define the similarity score as:

$$\text{Score}(q, d) = \sum_{i=1}^{n} \frac{1}{K} \sum_{j \in \mathcal{I}_i} \langle q_i \mid d_j \rangle,$$

where $\langle q_i \mid d_j \rangle$ denotes the dot product between the i-th query token and the j-th document token, and $\mathcal{I}_i \subseteq \{1, \ldots, m\}$ is the set of indices corresponding to the K largest dot products $\langle q_i \mid d_j \rangle$ for fixed i, defined as:

$$\mathcal{I}_i = \arg \text{topK}_{j \in \{1,\ldots,m\}} \langle q_i \mid d_j \rangle$$

We employ TopKSim only during training; at inference, we revert to MaxSim, which has shown superior empirical performance.

3.3 Self-supervised Masked Contrastive Learning (MaskedCL).

In practical scenarios, there are often abundant PDF documents available, but corresponding annotated query-document pairs may be lacking. To address this scenario, we propose MaskedCL, a self-supervised contrastive learning approach designed to mimic supervised contrastive fine-tuning without relying on labeled queries. MaskedCL integrates contrastive learning principles with a masking-based pretext task specifically tailored for multimodal document retrieval: *(i)* We construct pseudo-queries by randomly masking spans within the text content extracted from PDFs. *(ii)* Correspondingly, we generate masked versions of document images by overlaying white masks on random patches. *(iii)* Finally, we perform contrastive alignment between the masked textual representations and masked visual representations using our TopKSim late interaction mechanism. This strategy encourages robust cross-modal representation learning, even when large portions of textual or visual information are masked.

Formally, following prior methods (Faysse et al., 2025; Khattab and Zaharia, 2020), we define the MaskedCL loss using in-batch softmax cross-entropy over positive and hardest negative scores. Given a pseudo-query q_k and its corresponding masked document image d_k, we compute:

$$s_k^+ = \text{Score}(q_k, d_k),$$
$$s_k^- = \max_{l \neq k} \text{Score}(q_k, d_l),$$

where $\text{Score}(q, d)$ denotes the TopKSim score. The MaskedCL loss is then defined as:

$$\mathcal{L}_{\text{MaskedCL}} = -\frac{1}{b} \sum_{k=1}^{b} \log \left(\frac{\exp(s_k^+)}{\exp(s_k^+) + \exp(s_k^-)} \right),$$

with b representing the batch size.

4 Experimental Setup

Datasets & Benchmarks For masked OCR language modeling (MOLM), we use 4M digital PDF documents from the pdfa-eng-wds dataset[1], rendered as images at 96 dpi. From these, 1M documents with their underlying metadata (words and bounding boxes) are used for self-supervised contrastive masked learning. For supervised fine-tuning, we adopt the ViDoRe training split (Faysse et al., 2025), which includes 118,695 query-page pairs from synthetic and public datasets.

For evaluation, we use two benchmarks: *(i)* ViDoRe V1 (in-domain) (Faysse et al., 2025), which covers 10 academic and real-world datasets (e.g., DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2021a), arXivVQA (Li et al., 2024), and real-world domains like energy, government, healthcare, and AI), and *(ii)* ViDoRe V2 (out-of-domain), comprising documents from 9 diverse real-world domains such as biomedical, economics, and ESG. Both benchmarks are multilingual, while training data is exclusively English.

Modeling ColMate builds upon the ColPali model (Faysse et al., 2025), chosen for its faster runtime. To compare the models efficiency, we measured forward pass times for image encoding on a single H100 GPU using a batch size of 4 and 500 images from the DocVQA benchmark. The average batch times were: ColPali-3B (76 ms)

[1] https://huggingface.co/datasets/pixparse/pdfa-eng-wds

	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health	Avg.
Self-supervised Models											
ColMate-Pali-3B (CL)	58.78	41.19	76.97	69.06	46.97	67.31	91.65	88.97	88.02	87.32	71.62
ColMate-Pali-3B (MaskedCL) (**Ours**)	67.38	44.03	77.81	71.37	49.84	66.96	92.25	91.88	92.75	90.95	**74.52**
Supervised Models											
ColPali-3B	83.03	58.45	85.71	87.44	70.36	77.38	97.43	95.40	96.21	96.91	84.93
ColPali-3B (Reproduced)	84.55	57.65	86.43	86.65	71.31	76.40	96.62	94.64	94.84	97.76	84.68
ColMate-Pali-3B (**Ours**)	83.68	57.52	84.15	87.65	74.06	79.84	98.36	94.15	95.34	96.65	**85.14**

Table 1: nDCG@5 scores of ColMate and baselines on the ViDoRe V1 (in-domain) benchmark 10 academic and real-world datasets. ColMate achieves the highest average performance. In self-supervised settings, MaskedCL outperforms standard contrastive learning (CL).

(Faysse et al., 2025), ColQwen2.5-3B (188 ms) (Wang et al., 2024), and ColSmolVLM-256M (100 ms) (Marafioti et al., 2025). ColPali-3B was the fastest. Notably, ColSmolVLM-256M, despite being 12× smaller than ColPali-3B, was slower, primarily due to differences in image preprocessing. ColPali-3B processes images at 448×448 resolution, while ColSmolVLM-256M uses a 2048×2048 resolution split into 17 crops of 512×512. Given its superior speed, we selected ColPali-3B for our ColMate framework. We initialize our training from the *paligemma-448-base* checkpoint.

Hyperparameters For MOLM, we train on the 4M pages from PDFA for 1 epoch, with a learning rate of 3e-5 and batch size 64 using Adam (Kingma and Ba, 2017). For MaskedCL, we train on 1M documents for one epoch with a learning rate of 2e-5 and batch size 256 using paged AdamW (8-bit). We use LoRA (Hu et al., 2021) with $\alpha = 32$ and rank $r = 32$ for the LLM layers and the projection layer. For effective masking, we adopt the SpanBERT masking mechanism (Joshi et al., 2020), masking contiguous spans sampled from a geometric distribution with a maximum length of 10 and probability $p = 0.2$. To simulate real-world queries referencing small document sections, we apply an aggressive 80% masking probability to text. To avoid trivial text-to-image matching, we also randomly mask 50% of OCR words in the images with white masks. Finally, in the full supervised finetuning setup, we pre-train ColMate using MaskedCL before the full supervised fine-tuning.

For supervised fine-tuning, we follow ColPali v1.3 settings[2], using a learning rate of 5e-5, batch size 256, 1000 warmup steps, and 3 training epochs of the ViDoRe training set (Faysse et al., 2025). Finally, we set $K = 5$ for TopKSim, as it provided the best performance in preliminary experiments comparing $K \in \{3, 5, 10\}$. Experiments were conducted on machines equipped with 8×H100 and 4×H100 GPUs.

[2]https://huggingface.co/vidore/colpali-v1.3

5 Evaluation

5.1 Main Results

We present results on ViDoRe V1 (in-domain) and ViDoRe V2 (out-of-domain) in Tables 1 and 2. ColMate outperforms the ColPali baseline on both benchmarks. On ViDoRe V1, ColMate-Pali-3B achieves an average nDCG@5 of 85.14, surpassing both the original ColPali-3B (84.93) and our reproduction (84.68). On ViDoRe V2, ColMate reaches 57.61, compared to 54.60 for ColPali-3B and 54.00 for our reproduction, demonstrating strong generalization to unseen domains.

To simulate scenarios without image-query pairs, we evaluate self-supervised Masked Contrastive Learning (MaskedCL) against standard Contrastive Learning (CL) without masking. MaskedCL, even without supervised finetuning, achieves competitive performance—74.52 on ViDoRe V1 and 41.50 on ViDoRe V2—highlighting its effectiveness in low-resource settings. The performance gap with supervised finetuning is especially narrow on subsets such as AI, Energy, Gov., and Health, which represent real-world documents. MaskedCL also outperforms standard CL by 2.90 and 3.39 on V1 and V2, respectively, making it a strong choice when training data lacks image-query pairs.

5.2 Ablation Studies

To better understand the individual contributions of each proposed component of ColMate, we perform a detailed ablation study using the PaliGemma-3B model (Beyer et al., 2024). The effectiveness of each component is measured using the ViDoRe V1 and ViDoRe V2 benchmarks, and results are reported using the nDCG@5 metric in Figure 2 and more detailed numbers on Table 3.

	MIT Bio.	Econ. Mac.	ESG	AXA	ESG-M	AXA-M	MIT Bio.-M	Econ. Mac-M	ESG Human	Avg.
Self-supervised Models										
ColMate-Pali-3B (CL)	42.18	47.16	20.89	46.64	31.90	34.76	35.68	38.86	44.95	38.11
ColMate-Pali-3B (MaskedCL) (**Ours**)	50.18	47.21	24.53	40.24	35.36	36.61	44.61	40.22	54.51	**41.50**
Supervised Models										
ColPali-3B	59.70	51.60	57.00	59.80	55.70	50.10	56.50	49.90	51.10	54.60
ColPali-3B (Reproduced)	60.20	54.01	52.14	54.61	53.13	45.34	58.43	49.01	59.17	54.00
ColMate-Pali-3B (**Ours**)	60.99	55.99	54.15	67.01	53.44	50.61	59.31	54.14	62.82	**57.61**

Table 2: nDCG@5 scores on ViDoRe V2 (out-of-domain) across 9 real-world domains. ColMate achieves the best average and demonstrates stronger generalization.

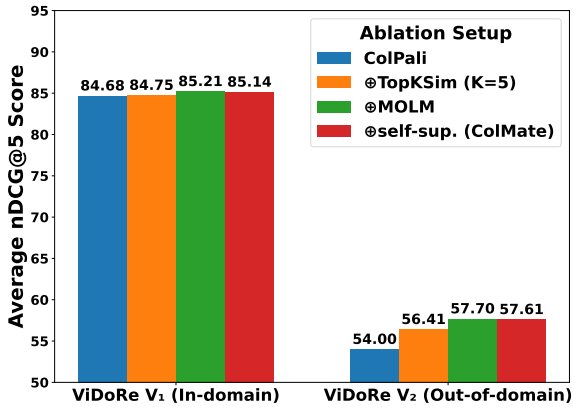


Figure 2: **Ablation Studies:** Impact of ColMate Components on ViDoRe Benchmarks (Average nDCG@5)

TopKSim vs MaxSim We compare our TopKSim mechanism with the MaxSim baseline (Khattab and Zaharia, 2020; Faysse et al., 2025) for multimodal document retrieval. Two ColPali models are fine-tuned from *paligemma-448-base* using identical hyperparameters from Section 4, one with TopKSim (K=5), the other with MaxSim.

As shown in Figure 2, TopKSim achieves better generalization on the out-of-domain ViDoRe V2 benchmark (56.41 vs. 54.00) and performs slightly better on the in-domain ViDoRe V1. The improvement is especially notable on the TabFQuAD subset (see Table 3), which lacks similar examples in the training set, further highlighting the improved robustness and generalization offered by TopKSim.

Adding Masked OCR Modeling We assess the impact of applying Masked OCR Language Modeling (MOLM) before supervised finetuning and observe consistent performance gains across both benchmarks. On the out-of-domain ViDoRe V2, MOLM increases average score from 56.41 to 57.70, supporting our hypothesis that modeling OCR-masked tokens improves vision-language representations for visual documents.

Adding Self-Supervised Finetuning We evaluate the impact of applying our self-supervised Masked Contrastive Learning (MaskedCL) framework before supervised finetuning. As shown in Figure 2, MaskedCL offers no performance gains when supervised data is abundant, suggesting its limited utility in data-rich scenarios. However, Tables 1 and 2 illustrate that MaskedCL alone achieves performance comparable to supervised finetuning across several ViDoRe subsets, highlighting its effectiveness in low-resource contexts.

5.3 Experiments with More Powerful Backbone Models

To examine the scalability of our framework, we applied ColMate to a more capable backbone, Qwen2.5-VL-3B (Wang et al., 2024), following the same pretraining and fine-tuning procedure used for ColPali. Table 4 reports performance on ViDoRe V1 (in-domain) and ViDoRe V2 (out-of-domain) benchmarks. ColMate improves the out-of-domain performance of Qwen2.5-VL-3B by 0.86 nDCG@5, indicating that the inductive bias introduced by ColMate has a smaller effect as backbone models become stronger and benefit from richer pretraining data. Nevertheless, the framework delivers larger gains with ColPali, which is substantially faster (86 ms vs. 188 ms for ColQwen2.5-VL), translating into tangible benefits in resource-constrained or latency-sensitive settings.

6 Conclusion

We introduce ColMate, a multimodal document retrieval model that addresses key limitations of existing methods by combining three simple but effective ideas: masked OCR language modeling for richer vision-text representations, masked contrastive learning for effective self-supervised alignment, and TopKSim for robust similarity aggregation during training. ColMate consistently improves retrieval performance over existing methods on both in-domain and out-of-domain ViDoRe benchmarks, with particularly strong generaliza-

tion to unseen domains. Our extensive ablation studies show the individual contribution of each component, validating our design choices. Our results highlight ColMate's utility for practical retrieval scenarios, particularly where annotated image-query pairs are scarce.

As future work, we plan to extend ColMate to additional model architectures and evaluate it on broader multimodal document retrieval benchmarks.

Limitations

This work applies the ColMate framework primarily to the PaliGemma model (Beyer et al., 2024) for computational efficiency. Future work will extend ColMate to other models such as SmolVLM (Marafioti et al., 2025). While ColMate provides only modest gains on in-domain benchmarks (Table 1), it delivers substantial improvements on out-of-domain tasks (Table 2). TopKSim also introduces an additional hyperparameter (K) that may require tuning. Lastly, MaskedCL offers limited benefits when annotated query-image pairs are abundant, but is highly effective as a self-supervised method when such data is unavailable.

Ethical Considerations

We complied with the terms of use and licenses for the ViDoRe benchmarks and the PaliGemma model, which were used solely for research puroses in our work. Our models are not generative; they encode documents and queries for retrieval tasks. Therefore, we do not anticipate risks typically associated with large language models, such as hallucinations or harmful content generation. Still, proper evaluation is necessary before deploying these models in real-world scenarios. AI writing tools were used only to enhance the paper's writing.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3530–3539.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *Preprint*, arXiv:2403.00231.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. *Preprint*, arXiv:2406.11251.

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

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *Preprint*, arXiv:2504.05299.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. *Preprint*, arXiv:2010.04592.

Juan Rodriguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte, François Savard, Ahmed Masry, Shravan Nayak, Rabiul Awal, Mahsa Massoud, Amirhossein Abaskohi, Zichao Li, Suyuchen Wang, Pierre-André Noël, Mats Leon Richter, Saverio Vadacchino, Shubham Agarwal, and 24 others. 2025. Bigdocs: An open dataset for training multimodal models on document and code tasks. *Preprint*, arXiv:2412.04626.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun

Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. Structextv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, and Ruiming Tang. 2024. Evaluating the external and parametric knowledge fusion of large language models. *Preprint*, arXiv:2405.19010.

A Appendices

A.1 Extended and Detailed Results

We present detailed ablation results in Table 3.

ViDoRe V_1 (nDCG@5) (In-domain)											
Method	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health	Avg.
ColPali	84.55	57.65	86.43	86.65	71.31	76.40	96.62	94.64	94.84	97.76	84.68
⊕ TopKSim (K=5)	83.08	57.40	85.11	90.43	71.02	76.41	97.63	94.66	94.84	96.94	84.75
⊕ MOLM	83.80	55.98	84.68	88.60	74.79	80.51	98.01	94.95	94.93	95.89	85.21
⊕ self-sup. (ColMate)	83.68	57.52	84.15	87.65	74.06	79.84	98.36	94.15	95.34	96.65	85.14

ViDoRe V_2 (nDCG@5) (Out-of-domain)										
Method	MIT Bio.	Econ. Mac.	ESG	AXA	ESG–M	AXA–M	Bio.–M	Econ.–M	ESG Human	Avg.
ColPali (Reproduced)	60.20	54.01	52.14	54.61	53.13	45.34	58.43	49.01	59.17	54.00
⊕ TopKSim (K=5)	60.53	54.46	55.87	61.13	55.91	47.75	57.99	52.34	61.71	56.41
⊕ MOLM	60.55	57.64	58.50	54.08	57.84	50.16	58.28	57.00	65.22	57.70
⊕ self-sup. (ColMate)	60.99	55.99	54.15	67.01	53.44	50.61	59.31	54.14	62.82	57.61

Table 3: Ablation studies showing the impact of the different ColMate components on the ViDoRe V_1 and V_2 benchmarks (nDCG@5).

ViDoRe V_1 (nDCG@5) (In-domain)											
Method	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health	Avg.
ColQwen2.5-VL-3B	89.22	63.22	92.37	91.12	81.10	87.30	99.63	95.89	96.41	97.89	89.41
ColMate-Qwen2.5-VL-3B	90.24	61.10	93.68	91.51	81.88	90.24	99.26	96.39	96.54	98.13	**89.89**

ViDoRe V_2 (nDCG@5) (Out-of-domain)										
Method	MIT Bio.	Econ. Mac.	ESG	AXA	ESG–M	AXA–M	Bio.–M	Econ.–M	ESG Human	Avg.
ColQwen2.5-VL-3B	63.64	59.78	57.39	60.29	57.38	53.19	61.12	56.52	68.39	59.74
ColMate-Qwen2.5-VL-3B	62.06	59.55	60.06	65.82	60.21	56.39	60.33	52.05	68.94	**60.60**

Table 4: nDCG@5 results for applying ColMate to Qwen2.5-VL-3B.