Retrieval in Vision Space

25-2 Lab Seminar

심규호

Natural Language Processing & Artificial Intelligence

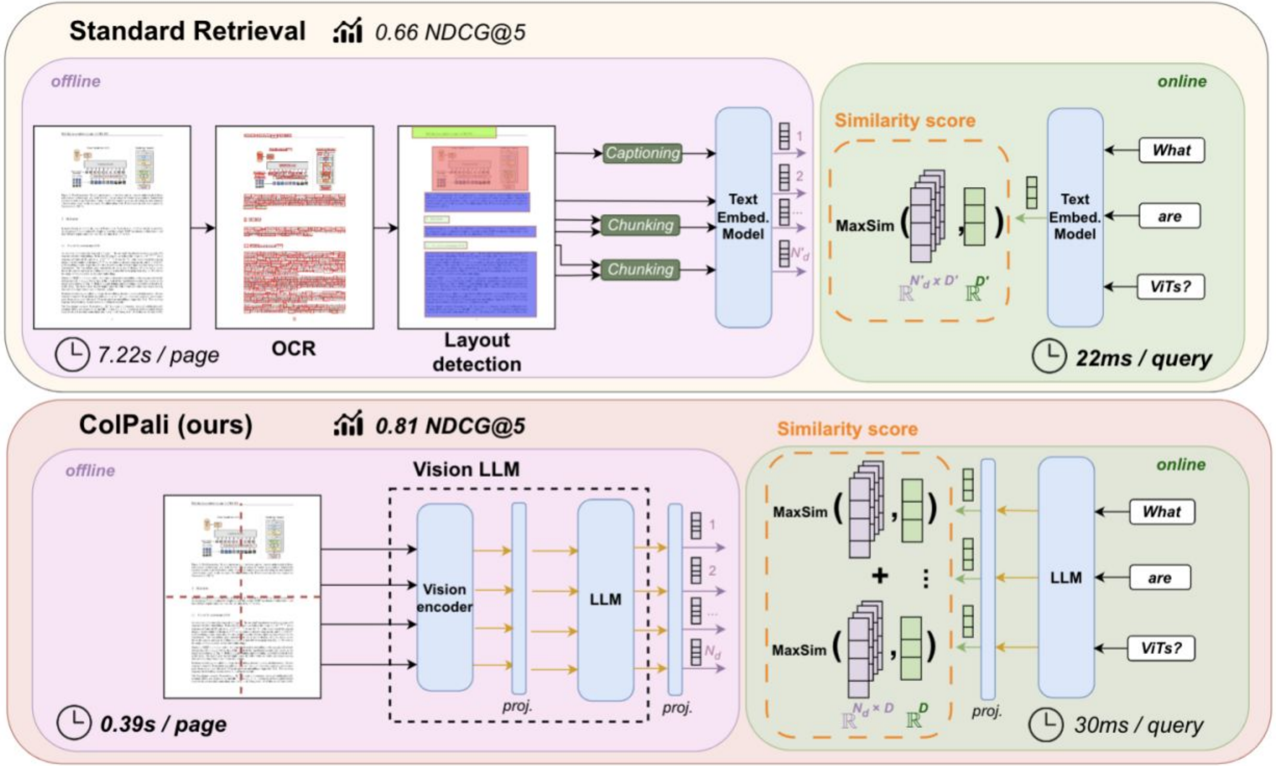
고려대학교 KOREA UNIVERSITY

ColPali: EFFICIENT DOCUMENT RETRIEVAL WITH VISION LANGUAGE MODELS

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

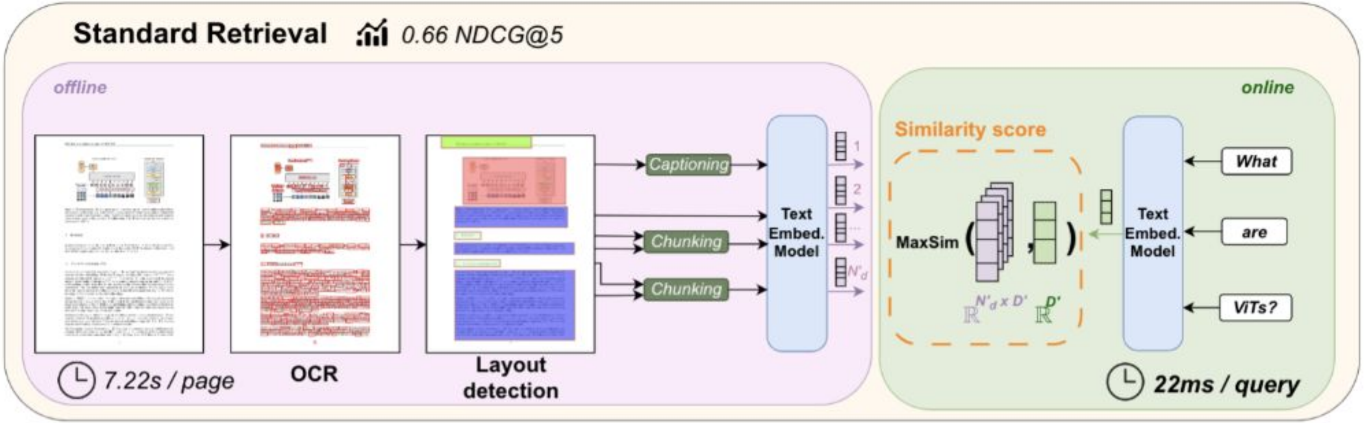
2025 ICLR

Colpali simplifies Document Retrieval



Colpali simplifies Document Retrieval

Standard Retrieval - prior data ingestion pipeline

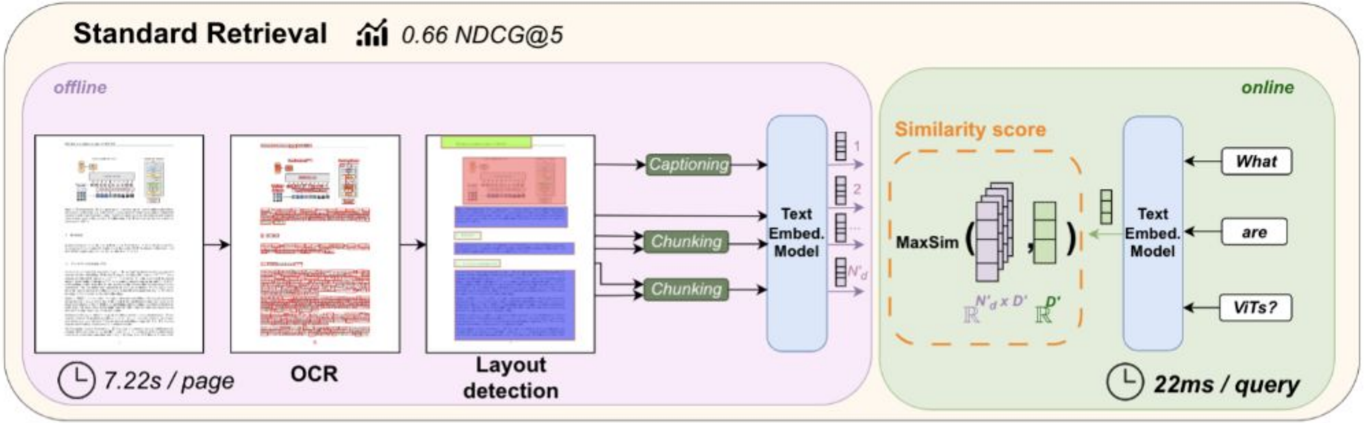


Prior Data Ingestion Pipeline

1. PDF parsers/OCR로 page에서 텍스트 추출 진행
2. Document Layout Detection 모델로 page 의 구성대로 segment → e.g., paragraph, title, table, figure, etc.
3. (Optional) Captioning 진행
4. Chunking 진행

Colpali simplifies Document Retrieval

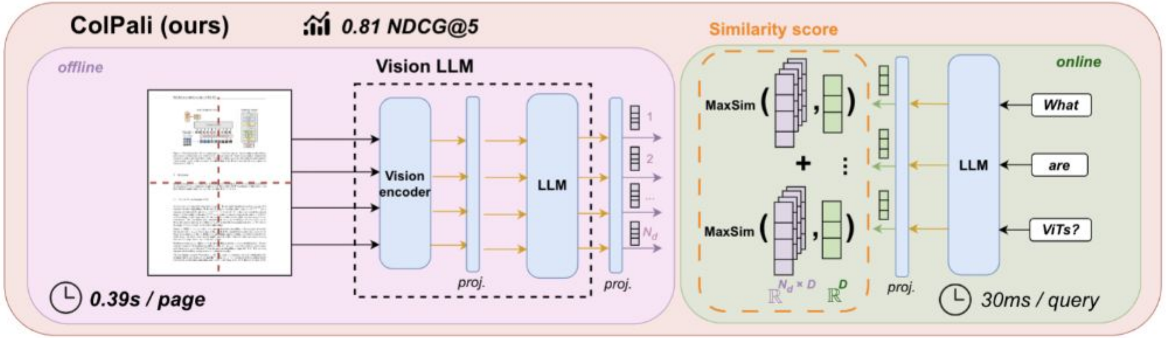
Standard Retrieval - Pain Points



1. Visual Cue들에 대한 활용 부족
   a. Modern retrieval system들은 텍스트 정보에 대게 의존
2. 길고 취약한 데이터 수집 및 색인 파이프라인 (e.g., PDF parsing, layout detection, chunking, captioning)

Colpali simplifies Document Retrieval

Ours - Late Interaction Based Vision Retrieval



ColPali = ColBERT + PaliGemma

ColBERT → Late Interaction Mechanism

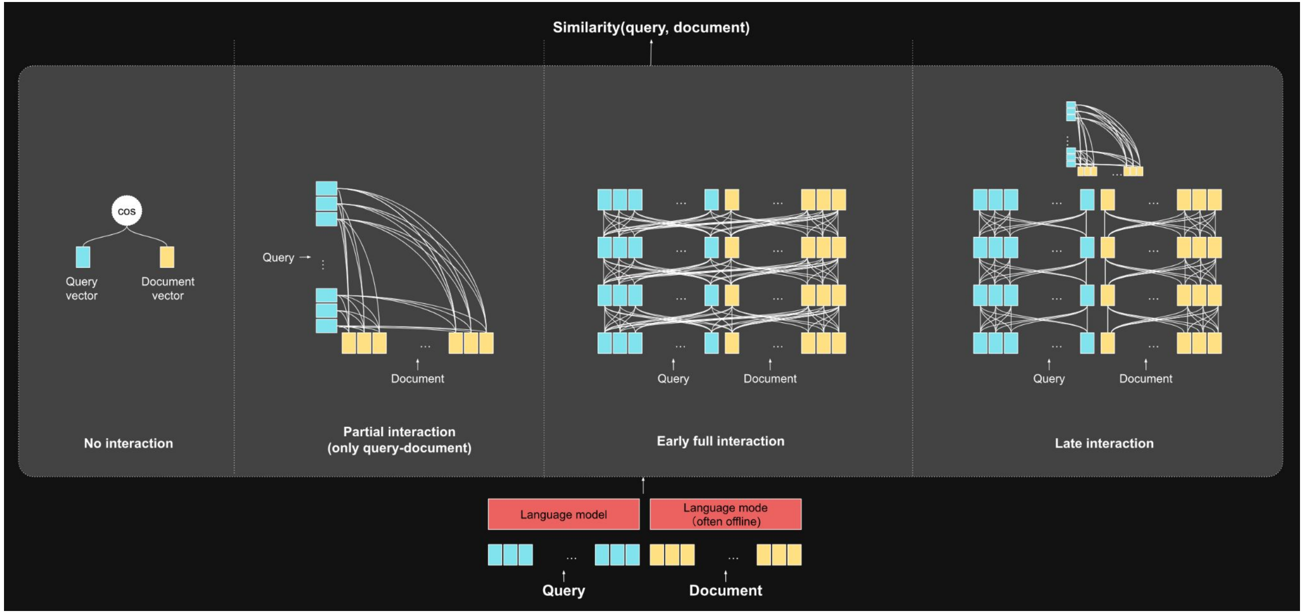
PaliGemma → Vision Language Model

Architecture

1. Vision-Language Model: Multi-modal 학습에서 얻게 되는 text & image token들의 alignment를 활용하자!
2. Late Interaction: 개별 query와 document 용어 간의 풍부한 상호작용
3. Contrastive Loss

Colpali simplifies Document Retrieval

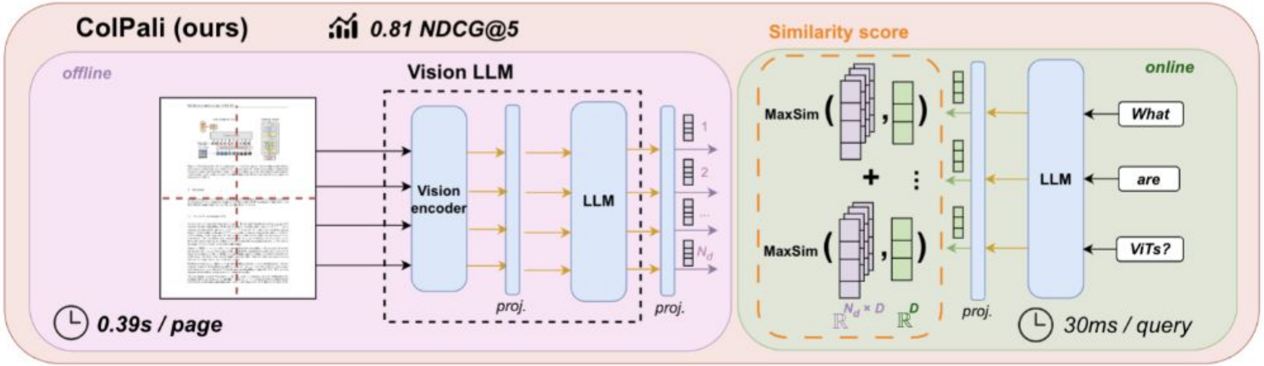
Ours - ColBERT's Late Interaction Mechanism



$$\mathrm{LI}(q,d) = \sum_{i \in [|1, N_q|]} \max_{j \in [|1, N_d|]} \langle \mathbf{E_q}^{(i)} | \mathbf{E_d}^{(j)} \rangle$$

Colpali simplifies Document Retrieval

Ours - Late Interaction Based Vision Retrieval

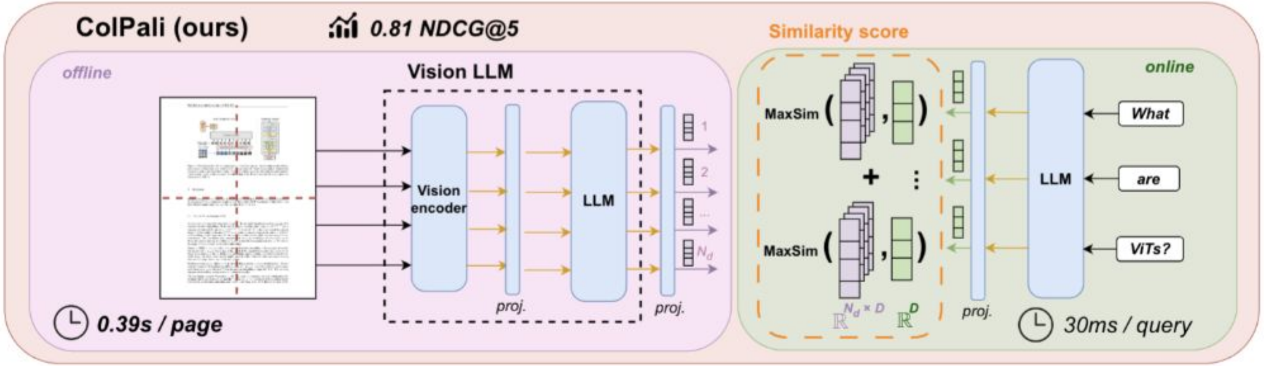


Architecture - VLM; leverage the alignment b/w output embeddings of text & image tokens

1. PaliGemma.
   a. SigLIP의 patch embedding들로 Gemma-2B의 text vector space에 projection
   b. Pali3
2. *Bag of Embedding*.
   a. 축소된 차원 (D=128)의 vector space에 language model의 token embedding을 매핑하기 위한 projection layer 추가

Colpali simplifies Document Retrieval

Ours - Late Interaction Based Vision Retrieval



Architecture - Contrastive Loss

$$\mathcal{L} = -\frac{1}{b}\sum_{k=1}^{b}\log\left[\frac{\exp\left(s_k^+\right)}{\exp\left(s_k^+\right)+\exp\left(s_k^-\right)}\right] = \frac{1}{b}\sum_{k=1}^{b}\log\left(1+\exp\left(s_k^- - s_k^+\right)\right)$$

1. Relevant query-doc pairs에 높은 similarity score 지정
2. 가장 큰 negative score에 집중함으로써,
   모델은 배치내에서 가장 혼란스럽거나 어려운 negative example으로부터 positive pair를 구별하도록 유도

$$s_k^+ = \mathrm{LI}\left(q_k, d_k\right)$$

$$s_k^- = \max_{l, l \neq k} \quad \mathrm{LI}\left(q_k, d_l\right)$$

Colpali simplifies Document Retrieval

ViDoRe - Visual Document Retrieval Benchmark

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

Colpali simplifies Document Retrieval

Three main properties for an Efficient Document Retrieval System

(R1) Strong Retrieval Performance

(R2) Fast *Online* Querying

(R3) High Throughput Corpus Indexation (*Offline* Indexing)

(R1) Strong Retrieval Performance

성능에 가장 중요한 세가지 요소
(1) a carefully crafted task-specific dataset
(2) 이미지 안의 텍스트 의미를 더 잘 활용하기 위해 사전학습된 **LLM**과 **Vision Model**의 결합
(3) 문서 내 방대한 시각 정보를 더 잘 포착하기 위해 **single-vector representation** 대신 **multi-vector embedding** 사용

BiSigLIP → (1); Document retrieval-oriented dataset에 파인튜닝된 **SigLIP**
BiPali →(1)(2); PaliGemma bi-encoder model. SigLIP이 생성한 **patch embedding**들을 **LLM**에 입력하여 문맥화된 패치 임베딩 생성
ColPali →(1)(2)(3); ColBERT strategy

(R1) Strong Retrieval Performance

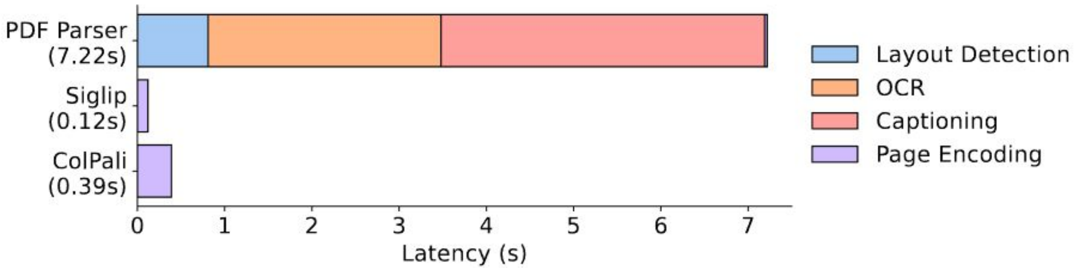
	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
`Unstructured` text-only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	$28.4_{\downarrow 5.7}$	-	-	$36.1_{\downarrow 7.9}$	$68.5_{\uparrow 8.9}$	$88.4_{\downarrow 2.0}$	$76.8_{\downarrow 1.5}$	$77.7_{\downarrow 1.1}$	$84.6_{\uparrow 2.0}$	-
`Unstructured` + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	$31.4_{\downarrow 0.2}$	$25.7_{\downarrow 11.1}$	$60.1_{\downarrow 2.8}$	$70.8_{\uparrow 24.3}$	$50.5_{\downarrow 12.2}$	$\mathbf{73.2}_{\uparrow 8.9}$	$90.2_{\downarrow 2.6}$	$83.6_{\downarrow 2.3}$	$84.9_{\uparrow 1.0}$	$91.1_{\uparrow 3.9}$	$66.1_{\uparrow 0.6}$
`Unstructured` + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	$35.7_{\downarrow 4.4}$	$32.9_{\downarrow 5.4}$	$71.9_{\uparrow 1.9}$	$69.1_{\uparrow 33.7}$	$43.8_{\downarrow 17.7}$	$73.1_{\uparrow 12.2}$	$88.8_{\uparrow 0.8}$	$83.3_{\downarrow 1.4}$	$80.4_{\downarrow 2.3}$	$91.3_{\uparrow 2.1}$	$67.0_{\uparrow 1.9}$
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	$58.5_{\uparrow 15.3}$	$32.9_{\uparrow 2.6}$	$70.5_{\uparrow 6.4}$	$62.7_{\uparrow 4.6}$	$30.5_{\uparrow 4.3}$	$26.5_{\uparrow 7.8}$	$74.3_{\uparrow 11.8}$	$73.7_{\uparrow 8.0}$	$74.2_{\uparrow 8.1}$	$82.3_{\uparrow 3.2}$	$58.6_{\uparrow 7.2}$
BiPali (+LLM)	$56.5_{\downarrow -2.0}$	$30.0_{\downarrow -2.9}$	$67.4_{\downarrow -3.1}$	$76.9_{\uparrow 14.2}$	$33.4_{\uparrow 2.9}$	$43.7_{\uparrow 17.2}$	$71.2_{\downarrow -3.1}$	$61.9_{\downarrow -11.7}$	$73.8_{\downarrow -0.4}$	$73.6_{\downarrow -8.8}$	$58.8_{\uparrow 0.2}$
ColPali (+Late Inter.)	$\mathbf{79.1}_{\uparrow 22.6}$	$\mathbf{54.4}_{\uparrow 24.5}$	$\mathbf{81.8}_{\uparrow 14.4}$	$\mathbf{83.9}_{\uparrow 7.0}$	$\mathbf{65.8}_{\uparrow 32.4}$	$\mathbf{73.2}_{\uparrow 29.5}$	$\mathbf{96.2}_{\uparrow 25.0}$	$\mathbf{91.0}_{\uparrow 29.1}$	$\mathbf{92.7}_{\uparrow 18.9}$	$\mathbf{94.4}_{\uparrow 20.8}$	$\mathbf{81.3}_{\uparrow 22.5}$

(R2) Fast *Online* Querying

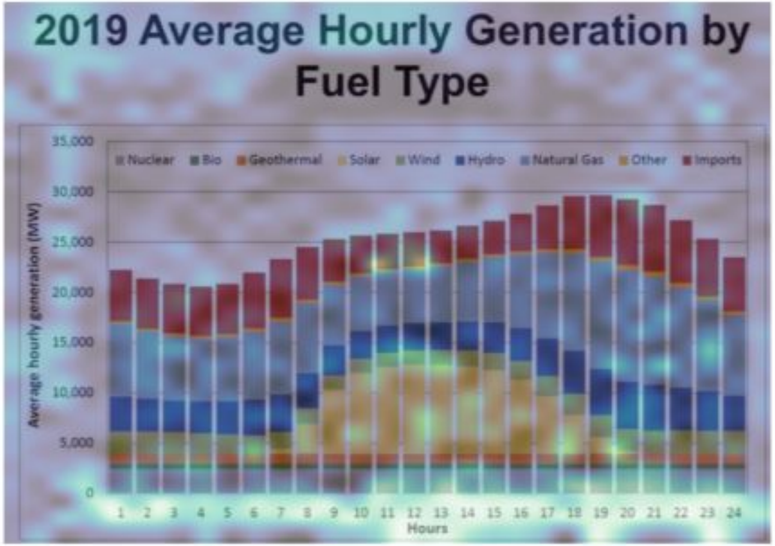
- **Querying Latency:** ColPali (30ms/15 tokens) vs. BGE-M3 (22ms/15 tokens)
- **Similarity Computation: Late Interaction >> Bi-encoder**

(R3) High Throughput Corpus Indexation (*Offline* Indexing)

- **PDF processing**을 통해 서로 다른 **chunk**를 얻는 것이 가장 시간 소모적인 부분
- 더불어, **Multimodal** 데이터를 처리하기 위해 **Captioning**을 사용하는 것은 이미 긴 이 과정을 더욱 악화
- 반면, **ColPali**는 문서 **page**의 **image representation**을 직접 **encoding**함



Interpretability



Query: "Which hour of the day had the highest overall eletricity generation in 2019?"

User Query의 각 용어에 대해, ColPali는 가장 관련성 높은 문서 이미지 patch를 식별하고 query-page 매칭 점수를 계산함

Conclusions

1. **ViDoRe (Visual Document Retrieval Benchmark)**
   a. Visually Complex document들이 포함된 realistic setting에서 문서 검색 시스템 평가
2. **ColPali**
   a. 고품질의 multi-vector embedding을 생성하기 위해 Vision-Language Model 활용하는 새로운 검색 방법
3. **향후**
   a. **Out-of-Domain**에 대한 generalization
   b. **End-to-End RAG system** 구축 (+ visually grounded Query Answering)

Lost in OCR Translation? Vision-Based Approaches to Robust Document Retrieval

Alexander Most
AMost@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

Joseph Winjum
Joseph.Winjum@gmail.com
Gianforte School of Computing
Montana State University
Bozeman, MT, USA

Ayan Biswas
Ayan@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

Shawn Jones
SMJones@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

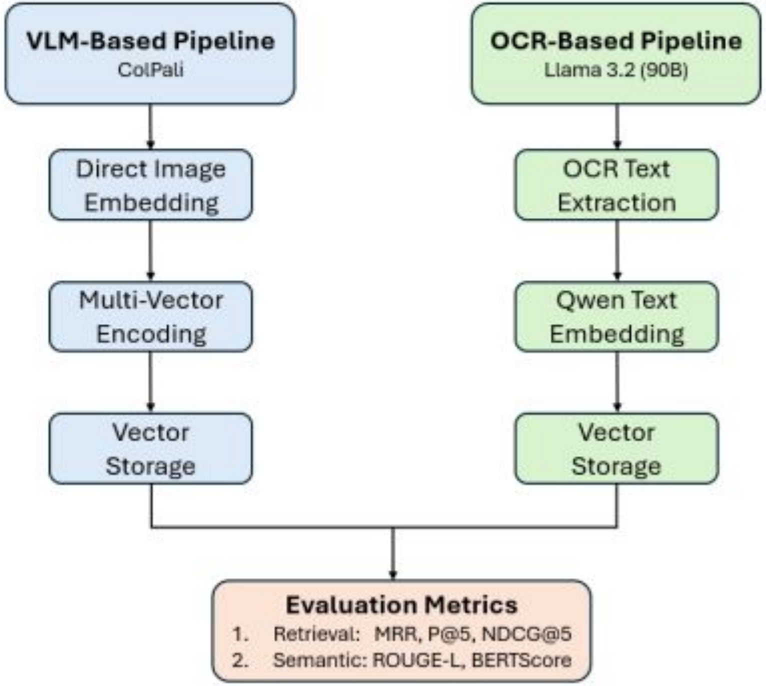
Nishath Rajiv Ranasinghe
Ranasinghe@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

Dan O'Malley
OMalleD@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

Manish Bhattarai
Ceodspspectrum@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

ACM DocEng'25

VLM-based vs. OCR-based RAG



VLM-Based

1) **ColQwen2 (7B)**

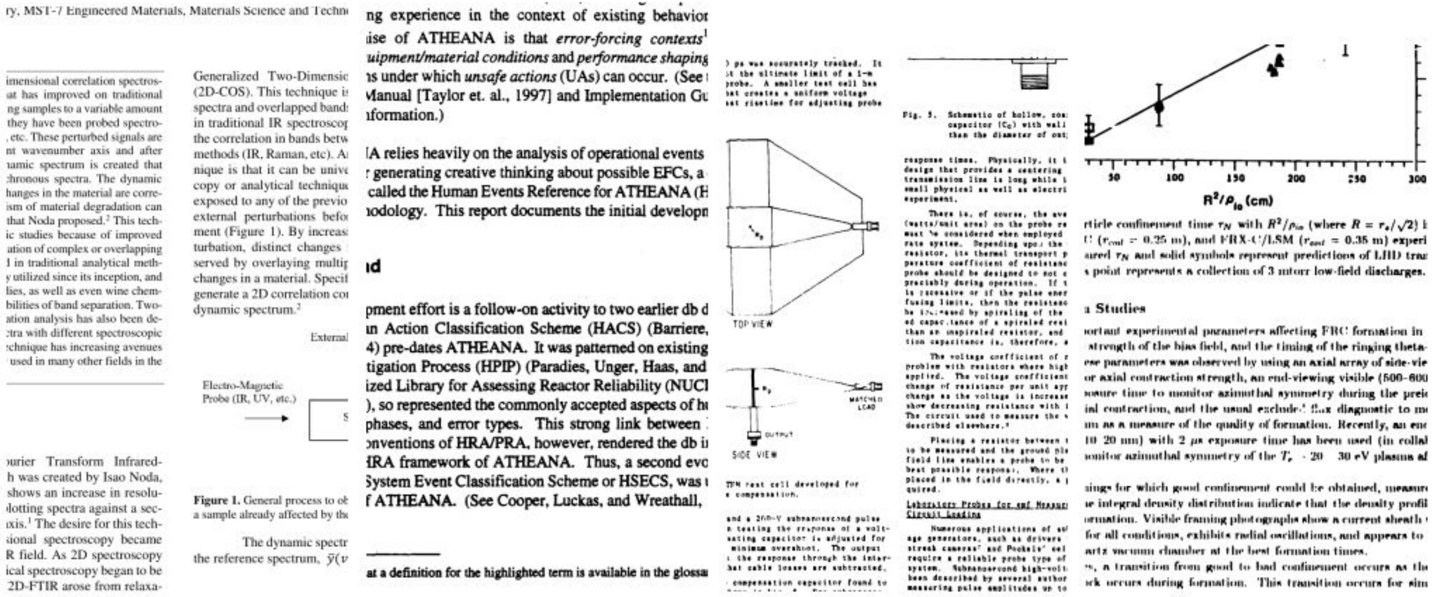
OCR-Based

1) **OCR pipeline w/Nougat OCR**
2) **OCR pipeline w/Llama 3.2 OCR**

1. Retrieval Performance
2. Semantic Answer Accuracy
3. Computational Efficiency
4. In-Domain Performance

Documents with Different Degradation Levels

DocDeg Dataset



(a) level-0 → Digital Copy. No Visible Degradation

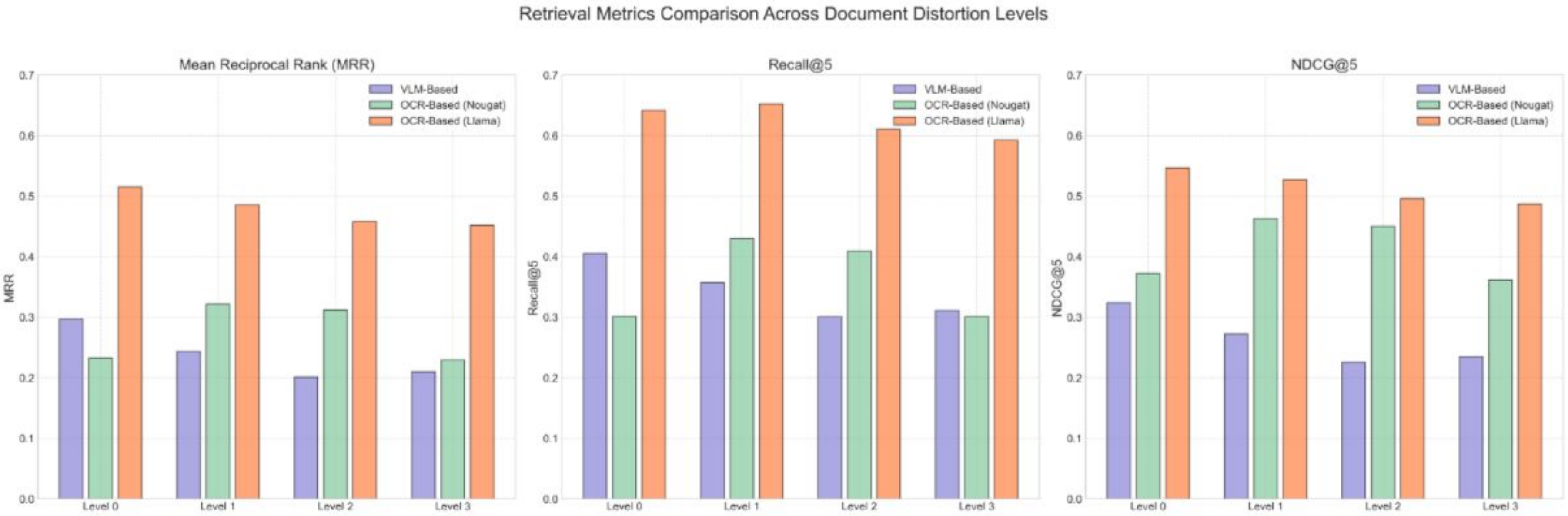
(b) level-1 → Perfect Scan. Minimal Degradation

(c) level-2 → Imperfect Scan. Noticeable Degradation

(d) level-3 → Severely Degraded.

Retrieval Performance

Across Realistic, visually degraded Conditions



1. Llama OCR 기반 파이프라인이 VLM 기반 파이프라인 지속적으로 능가
2. Nougat OCR 기반 파이프라인이 levels 1~3에서 VLM 기반 RAG를 능가
3. Nougat은 최고 품질 문서(level 0)에서는 가장 낮은 성능

Semantic Answer Quality

End-to-End performance

Distortion Level	Questions	Metric	VLM-Based	OCR-Based
Level 0	11190	Exact Match	0.0046	**0.0080**
		Average BLEU	0.0589	**0.0753**
		Average ROUGE-1	0.2088	**0.2547**
		Average ROUGE-L	0.1912	**0.2370**
Level 1	19154	Exact Match	0.0038	**0.0066**
		Average BLEU	0.0628	**0.0804**
		Average ROUGE-1	0.2167	**0.2643**
		Average ROUGE-L	0.1960	**0.2430**
Level 2	6864	Exact Match	0.0033	**0.0057**
		Average BLEU	0.0587	**0.0751**
		Average ROUGE-1	0.1975	**0.2410**
		Average ROUGE-L	0.1785	**0.2213**
Level 3	4748	Exact Match	0.0038	**0.0066**
		Average BLEU	0.0595	**0.0762**
		Average ROUGE-1	0.2044	**0.2493**
		Average ROUGE-L	0.1826	**0.2263**
Total	**41956**	Exact Match	0.0039	**0.0068**
		Average BLEU	0.0607	**0.0777**
		Average ROUGE-1	0.2100	**0.2556**
		Average ROUGE-L	0.1903	**0.2355**

Full RAG pipeline에서 Colpali를 사용하려면 **추가적인 단계**가 필요

1) Colpali로 검색된 문서 이미지에 OCR을 적용하고 추출된 텍스트를 LLM에 전달 (*ours*)
2) 이미지에 대한 QA가 가능한 VLM을 사용

시각적 뉘앙스를 전달하는 VLM의 잠재적 benefit은 개선된 의미적 성능으로는 이어지지 못함

Computational Efficiency Analysis
OCR Time, Embedding Time, Retrieval Latency, Memory Usage

Metric	VLM-Based	OCR-Based
OCR Time (per document)	N/A	6s
Embedding Time (per document)	.4739s	.5503s
Retrieval Latency (per query)	0.04252s	.0311s
Memory Usage (per 1k documents)	1.38 GB	9.5 GB

1. OCR로 텍스트 추출하는데 소모되는 시간: 0s vs. 6s
2. Embedding 시간 VLM < OCR
3. 1,000개 문서당 메모리 소비량은 VLM 기반 파이프라인이 현저히 낮음

⇒ VLM-기반 접근법은 계산 효율성과 자원 활용에서 상당한 이점 제공

In-Domain Performance

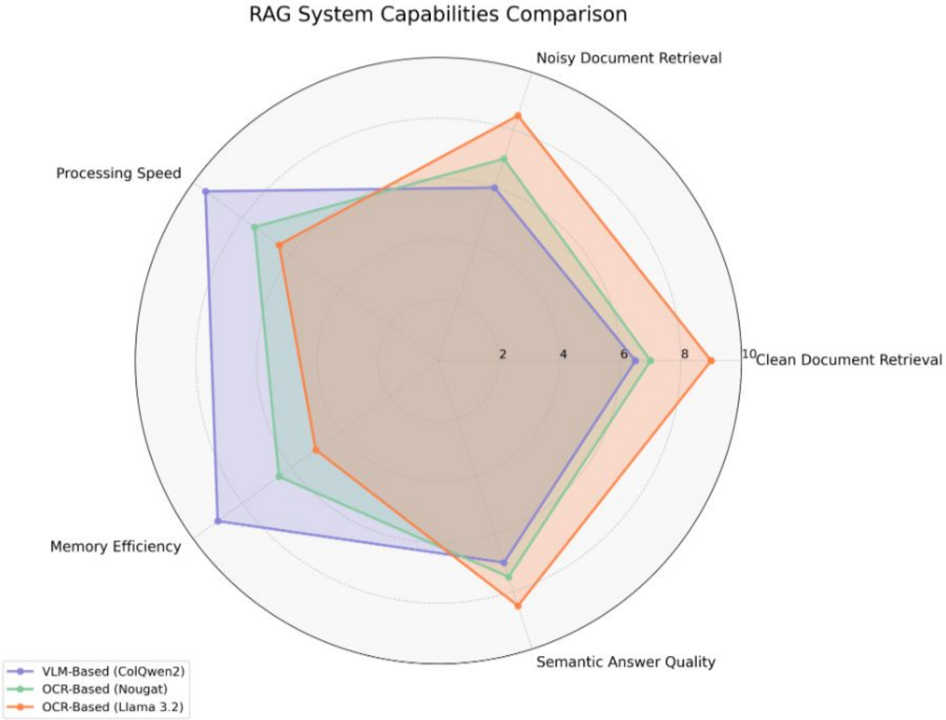
cleaner document conditions →ViDoRe-DocVQA

Metric (Top-5)	ColQwen2	OCR+Qwen2 (Nougat)	OCR+Qwen2 (Llama)
NDCG@5	0.6027	0.3373	0.3373
MAP@5	0.5813	0.3147	0.3147
MRR@5	0.5851	0.3164	0.3164
Recall@5	0.6674	0.4058	0.4058

1. DocVQA로 fine-tuning된 ColQwen2가 가장 강력한 검색 성능을 달성
2. ViDoRe는 노이즈가 없지만, 그 쿼리들은 페이지 텍스트와 느슨하게만 연결되어 있음
   → *subtle semantic link*
3. ColQwen2가 높은 성능을 달성한 이유는 ViDoRE 자체의 query-page 쌍들로 fine-tuning되어, 그 embedding들이 정확히 그런 쌍들의 idiosyncratic link들을 알아차리도록 학습되었기 때문임
4. 두 OCR 기반 파이프라인은 동일한 점수 기록 → 이러한 깨끗한 pdf에서는 OCR 품질이 더 이상 병목이 아니라, Query-Page embedding간의 semantic alignment이 병목임.

Conclusion

OCR-based vs. VLM-based RAG pipelines



1. Task-specific fine-tuning 없이도 OCR 기반 접근법은 향상된 검색 및 생성 성능 제공
2. Standard Retrieval이 인덱싱 시간은 더 오래 걸리지만 query time에서는 더 효율적인 성능
3. ColPali는 End-to-End QA를 직접적으로 지원하지 않음

감사합니다