



Retrieval in Vision Space

25-2 Lab Seminar

심규호

ColPali: EFFICIENT DOCUMENT RETRIEVAL WITH VISION LANGUAGE MODELS

Manuel Faysse^{*1,3} **Hugues Sibille**^{*1,4} **Tony Wu**^{*1} **Bilel Omrani**¹

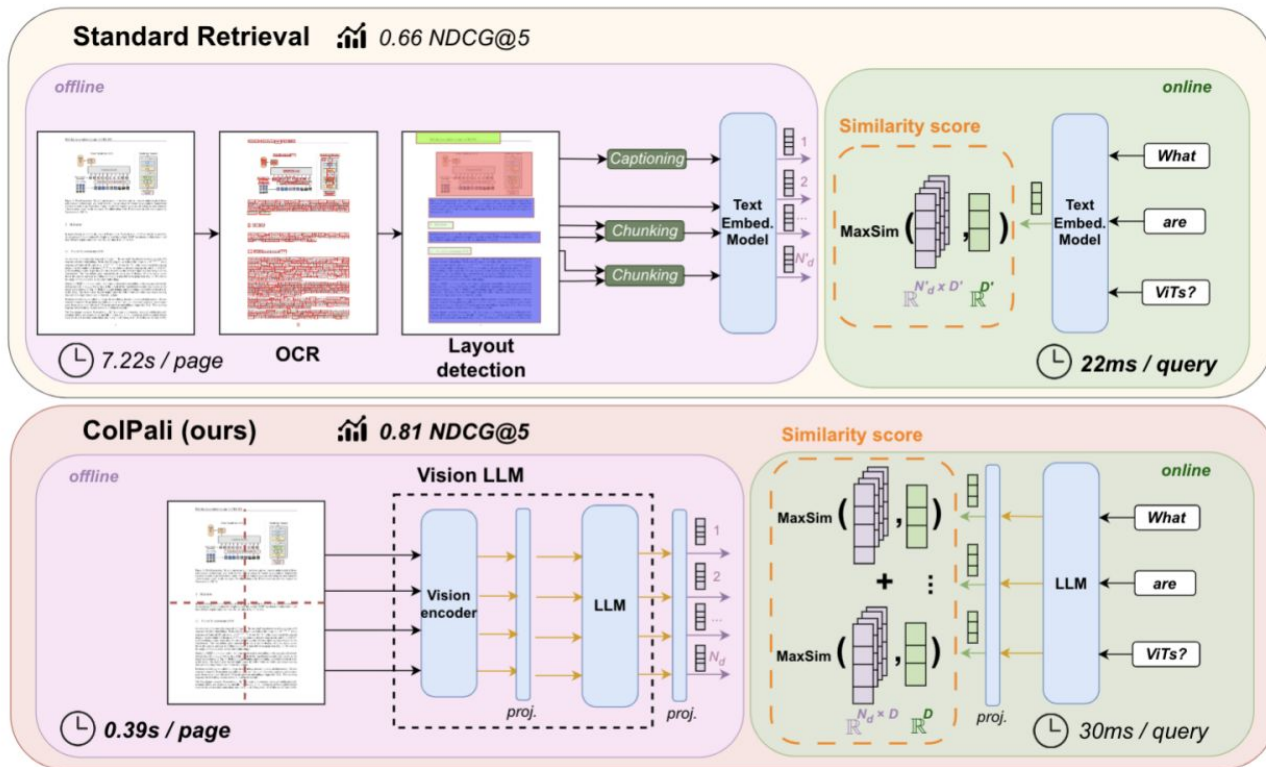
Gautier Viaud¹ **Céline Hudelot**³ **Pierre Colombo**^{2,3}

¹Illuin Technology ²Equall.ai ³CentraleSupélec, Paris-Saclay ⁴ETH Zürich

manuel.faysse@centralesupelec.fr

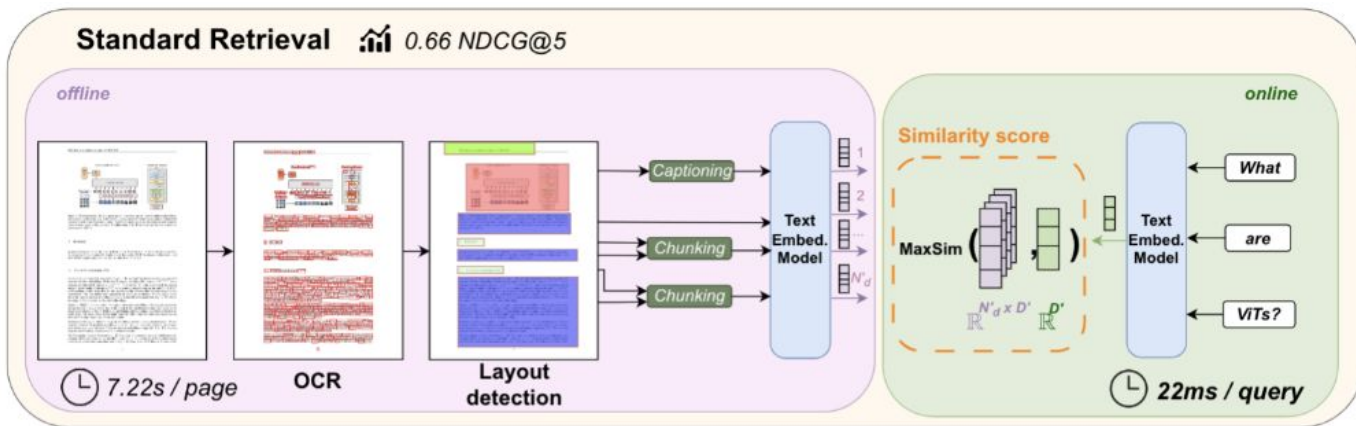
2025 ICLR

Colpali simplifies Document Retrieval



Colpali simplifies Document Retrieval

Standard Retrieval - prior data ingestion pipeline

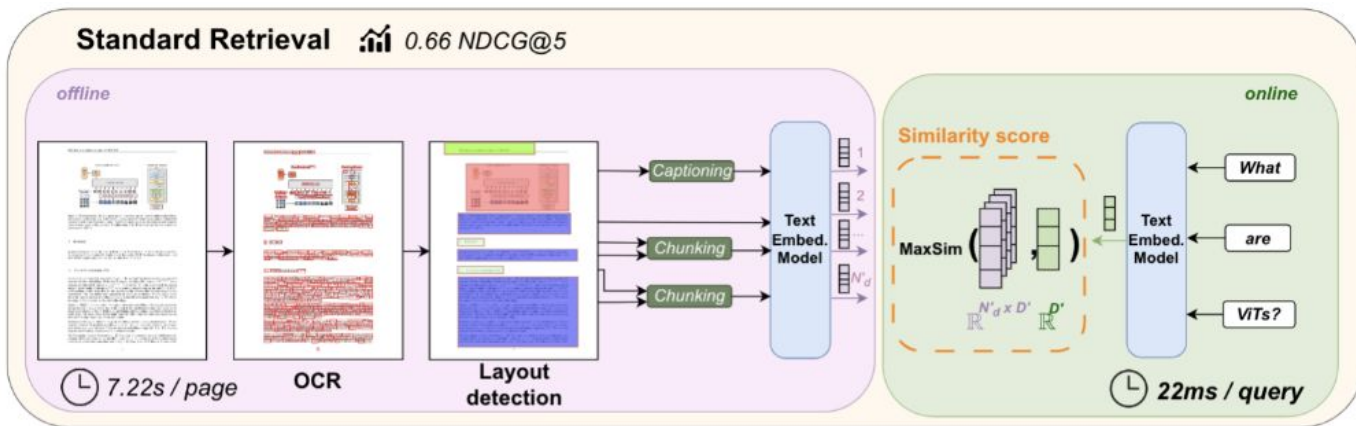


Prior Data Ingestion Pipeline

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

Colpali simplifies Document Retrieval

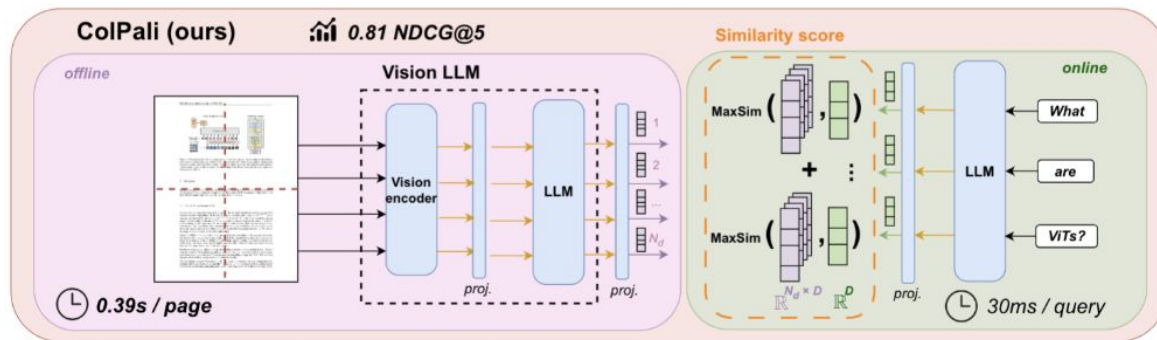
Standard Retrieval - *Pain Points*



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



$$\text{ColPali} = \text{ColBERT} + \text{PaliGemma}$$

Late Interaction
Mechanism

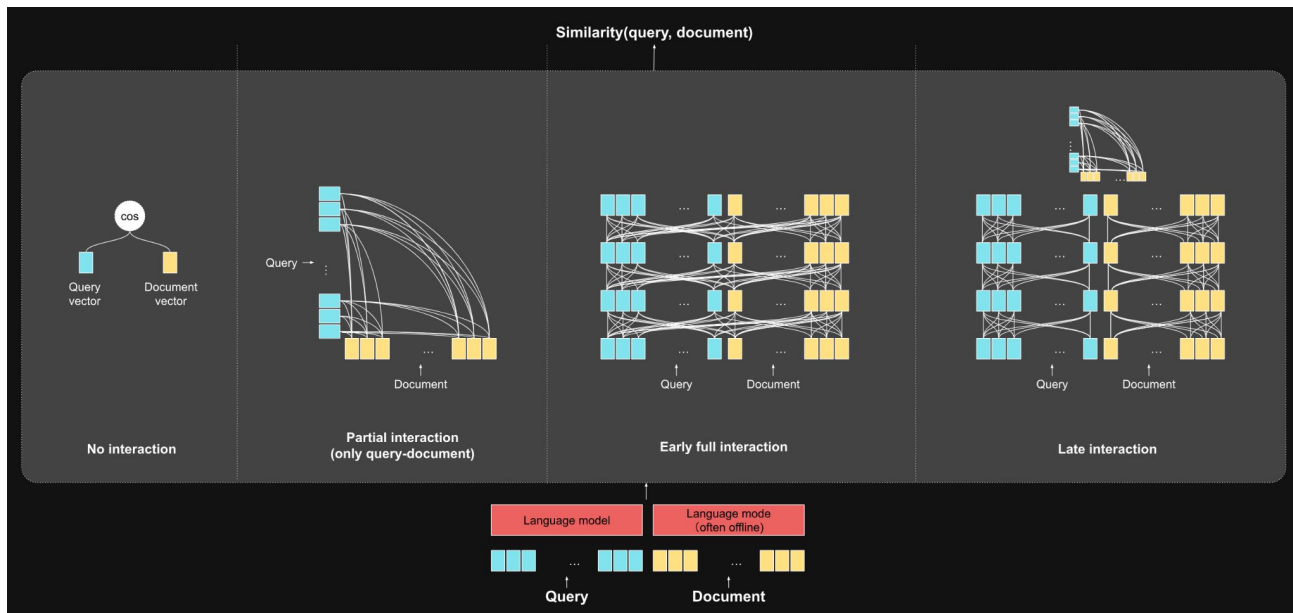
Vision Language Model

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

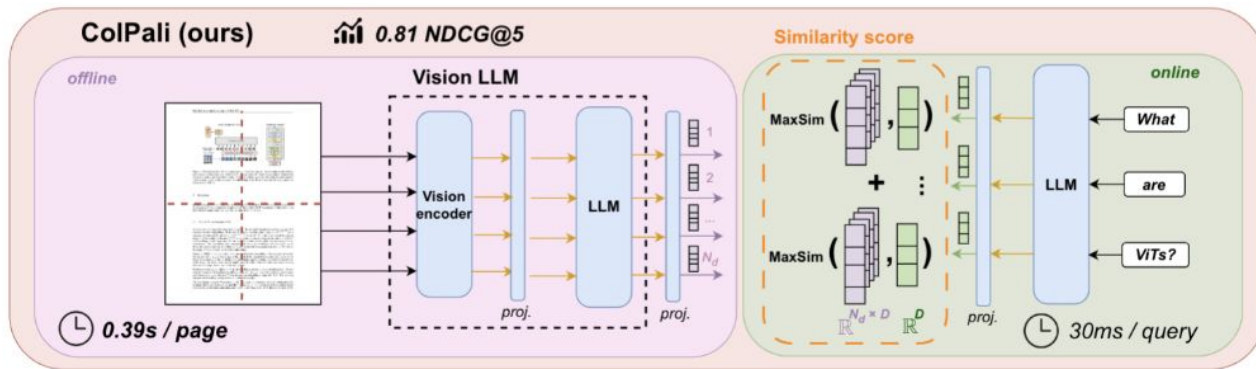
Ours - ColBERT's Late Interaction Mechanism



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

Colpali simplifies Document Retrieval

Ours - Late Interaction Based Vision Retrieval

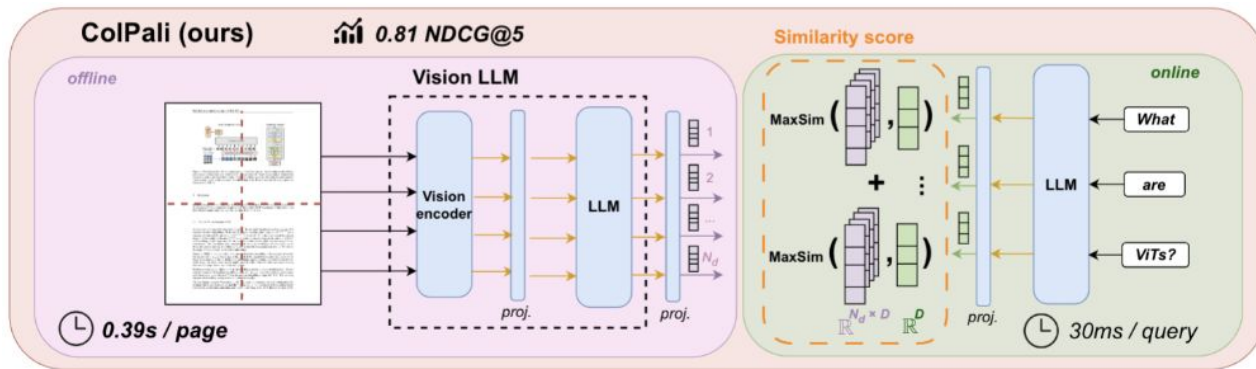


Architecture - VLM; leverage the alignment b/w output embeddings of text & image tokens

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

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

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

$$s_k^+ = \text{LI}(q_k, d_k)$$

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

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 LLM과 Vision Model의 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 embedding들을 LLM에 입력하여 문맥화된 패치 임베딩 생성

ColPali →(1)(2)(3); ColBERT strategy

(R1) Strong Retrieval Performance

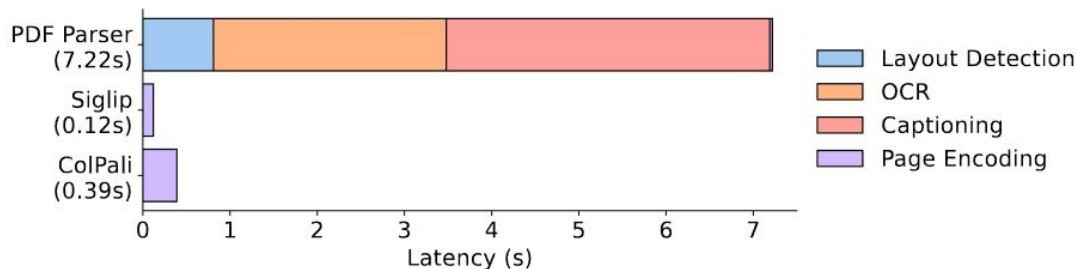
	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured text-only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4 _{↓5.7}	-	-	36.1 _{↓7.9}	68.5 _{↑8.9}	88.4 _{↓2.0}	76.8 _{↓1.5}	77.7 _{↓1.1}	84.6 _{↑2.0}	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4 _{↓0.2}	25.7 _{↓11.1}	60.1 _{↓2.8}	70.8 _{↑24.3}	50.5 _{↓12.2}	73.2 _{↑8.9}	90.2 _{↓2.6}	83.6 _{↓2.3}	84.9 _{↑1.0}	91.1 _{↑3.9}	66.1 _{↑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 _{↓4.4}	32.9 _{↓5.4}	71.9 _{↑1.9}	69.1 _{↑33.7}	43.8 _{↓17.7}	73.1 _{↑12.2}	88.8 _{↑0.8}	83.3 _{↓1.4}	80.4 _{↓2.3}	91.3 _{↑2.1}	67.0 _{↑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 _{↑15.3}	32.9 _{↑2.6}	70.5 _{↑6.4}	62.7 _{↑4.6}	30.5 _{↑4.3}	26.5 _{↑7.8}	74.3 _{↑11.8}	73.7 _{↑8.0}	74.2 _{↑8.1}	82.3 _{↑3.2}	58.6 _{↑7.2}
BiPali (+LLM)	56.5 _{↓-2.0}	30.0 _{↓-2.9}	67.4 _{↓-3.1}	76.9 _{↑14.2}	33.4 _{↑2.9}	43.7 _{↑17.2}	71.2 _{↓-3.1}	61.9 _{↓-11.7}	73.8 _{↓-0.4}	73.6 _{↓-8.8}	58.8 _{↑0.2}
ColPali (+Late Inter.)	79.1 _{↑22.6}	54.4 _{↑24.5}	81.8 _{↑14.4}	83.9 _{↑7.0}	65.8 _{↑32.4}	73.2 _{↑29.5}	96.2 _{↑25.0}	91.0 _{↑29.1}	92.7 _{↑18.9}	94.4 _{↑20.8}	81.3 _{↑22.5}

(R2) Fast *Online* Querying

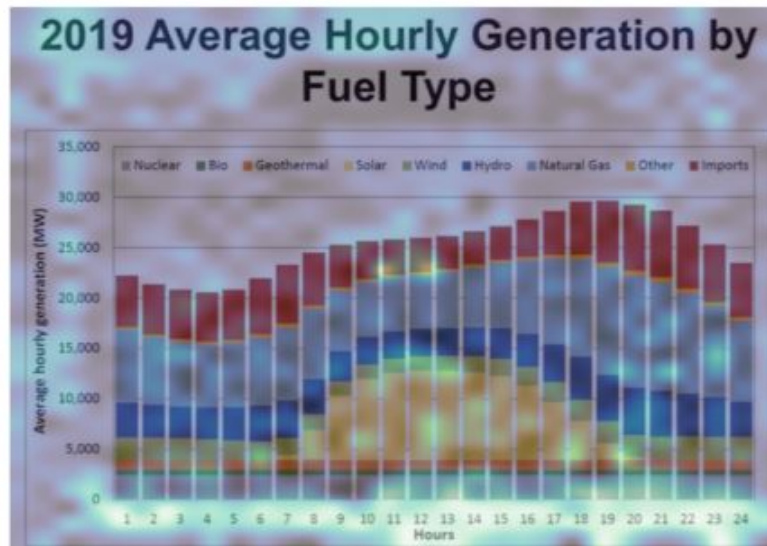
- Querying Latency: ColPali (30ms/15 tokens) vs. BGE-M3 (22ms/15 tokens)
- Similarity Computation: Late Interaction >> Bi-encoder

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

- PDF processing을 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, ColPali는 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-Domain에 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

Shawn Jones
SMJones@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

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

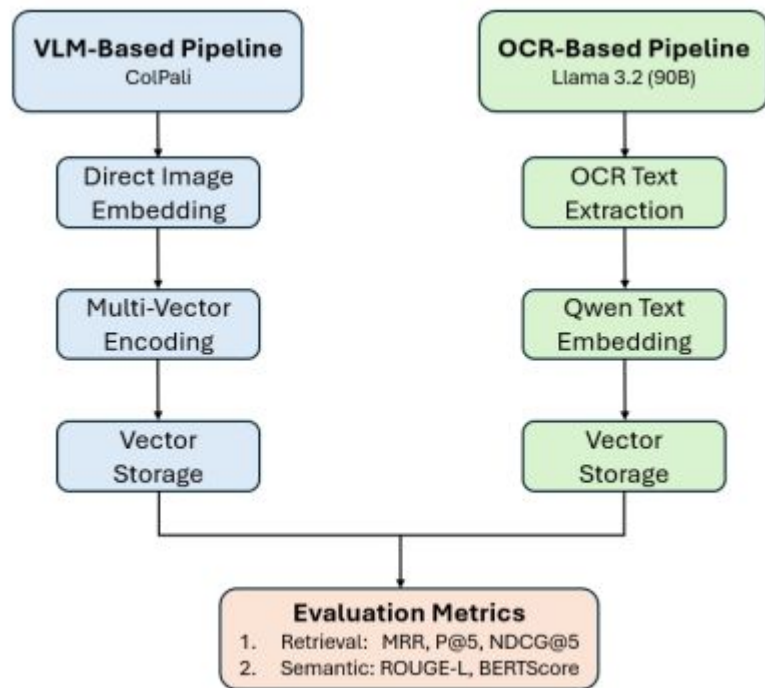
Manish Bhattarai
Ceodsppectrum@lanl.gov
Los Alamos National Laboratory
Los Alamos, NM, USA

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

Dan O'Malley
OMalleD@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

ry, MST-7 Engineered Materials, Materials Science and Techn

imensional correlation spectroscopy has improved on traditional ng samples to a variable amount they have been probed spectro- etc. These perturbed signals are nt wavenumber axis and alter amic spectrum is created that hronous spectra. The dynamic hanges in the material are corre- sm of material degradation can that Noda proposed.² This tech- ic studies because of improved ation of complex or overlapping i in traditional analytical meth- y utilized since its inception, and lies, as well as even wine chem- bilities of band separation. Two- ation analysis has also been de- tra with different spectroscopic chnique has increasing avenues used in many other fields in the

urier Transform Infrared- h was created by Isao Noda, shows an increase in resolu- dotting spectra against a sec- sion.¹ The desire for this tech- nical spectroscopy became R field. As 2D spectroscopy ic spectroscopy began to be 2D-FTIR arose from relaxa-

Generalized Two-Dimensic (2D-COS). This technique is spectra and overlapped band: in traditional IR spectroscopy the correlation in bands betw methods (IR, Raman, etc). A nique is that it can be univ copy or analytical techniqu exposed to any of the previo external perturbations befo ment (Figure 1). By increas turbation, distinct changes served by overlaying multj changes in a material. Specif generate a 2D correlation co dynamic spectrum.²

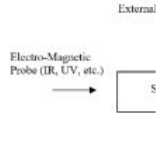


Figure 1. General process to obtain a sample already affected by the dynamic spectrum, $\tilde{\chi}(v)$

ng experience in the context of existing behavior ise of ATHEANA is that *error-forcing contexts* uipment/material conditions and performance shaping is under which *unsafe actions* (UAs) can occur. (See Manual [Taylor et. al., 1997] and Implementation Gu formation.)

IA relies heavily on the analysis of operational events ; generating creative thinking about possible EFCs, a called the Human Events Reference for ATHEANA (H odology. This report documents the initial develop

id
pment effort is a follow-on activity to two earlier db n Action Classification Scheme (HACS) (Barriere, 4) pre-dates ATHEANA. It was patterned on existing tigation Process (HPIP) (Paradies, Unger, Haas, and ized Library for Assessing Reactor Reliability (NUCL), so represented the commonly accepted aspects of h phases, and error types. This strong link between vventions of HRA/PRA, however, rendered the db itRA framework of ATHEANA. Thus, a second evc system Event Classification Scheme or HSECS, was f ATHEANA. (See Cooper, Luckas, and Wreathall,

at a definition for the highlighted term is available in the glossa

3 ps was accurately tracked. It is the ultimate limit of a low-probe. A smaller test cell has set creates a uniform voltage set risetime for adjusting probe

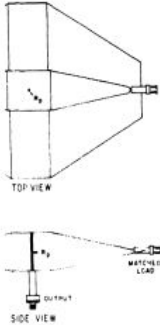


Figure 2. Schematic of hollow, coaxial capacitor (CC) with wall thickness less than the diameter of the probe.

and a 20-mV subnanosecond pulse a testing the response of a vol- ating capacitor is adjusted for minimum overshoot. The output i the response through the inter- hat cable losses are subtracted, a compensation capacitor found to

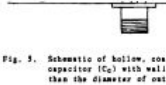


Figure 3. Schematic of hollow, coaxial capacitor (CC) with wall thickness less than the diameter of the probe.

response times. Physically, it is design that provides a centering transmission line is long while it small physical as well as electri experiment.

There is, of course, the ave (watts/unit area) on the probe re must be considered when employed rate system. Depending upon the resistor, its thermal transport p perature coefficient of resistance probe should be designed to not a preciously during operation. If it is too reactive or if the pulse and fusing limits, then the resistance be in- used by spiraling of the ed capacitor of a spiraled reel than an unsprayed resistor, and tion capacitance is, therefore, a

The voltage coefficient of a problem with resistors where high applied. The voltage coefficient change of resistance per unit app change as the voltage is increased show decreasing resistance with l The circuit used to measure the a described elsewhere.²

Placing a resistor between i to be measured and the ground il field line enables a probe to be heat possible response. Where it placed in the field directly, it is quired.

Labels: Probe for use with: Numerous applications of all age generators, such as delivery several cases and Probe's can require a reliable probe type of system. Subnanosecond high-volt been described by several author measuring pulse amplitudes up to

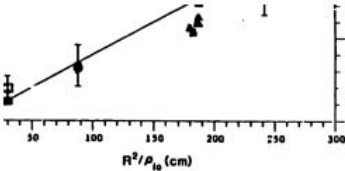


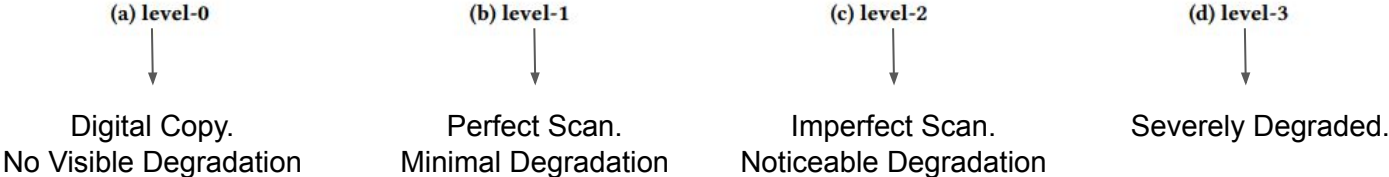
Figure 4. Relationship between R^2/ρ_{10} (cm) and response times. Solid symbols represent predictions of LHD; open symbols represent a collection of 3 mtorr low-field discharges.

Studies

ortant experimental parameters affecting ERC formation in strength of the bias field, and the timing of the ringing these parameters was observed by using an axial array of side-view or axial contraction strength, an end-viewing visible (500-600 ns) time to monitor azimuthal symmetry during the pre- al contraction, and the usual exclude. diagnostic to m m as a measure of the quality of formation. Recently, an en 10-20 nm) with 2 ps exposure time has been used (in colla nitor azimuthal symmetry of the $T_e \sim 20-30$ eV plasma ad

ings for which good confinement could be obtained, measure se integral density distribution indicate that the density profil formation. Visible framing photographs show a current sheath for all conditions, exhibits radial oscillations, and appears to artiz vacuum chamber at the best formation times.

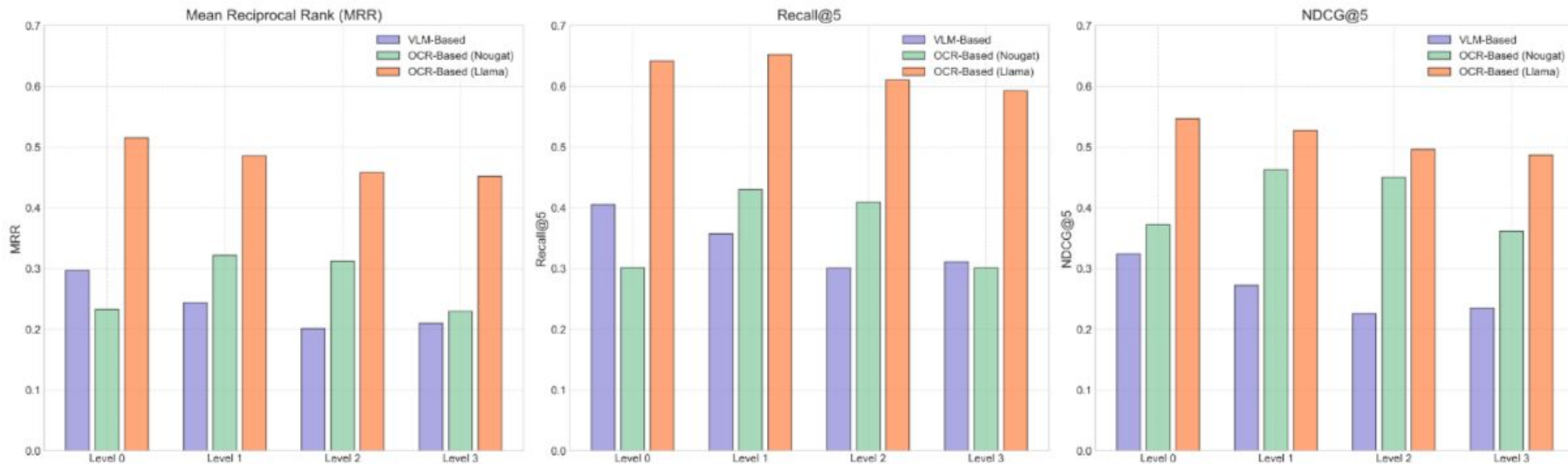
ly, a transition from good to bad confinement occurs as the ck occurs during formation. This transition occurs for sim



Retrieval Performance

Across Realistic, visually degraded Conditions

Retrieval Metrics Comparison Across Document Distortion Levels



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

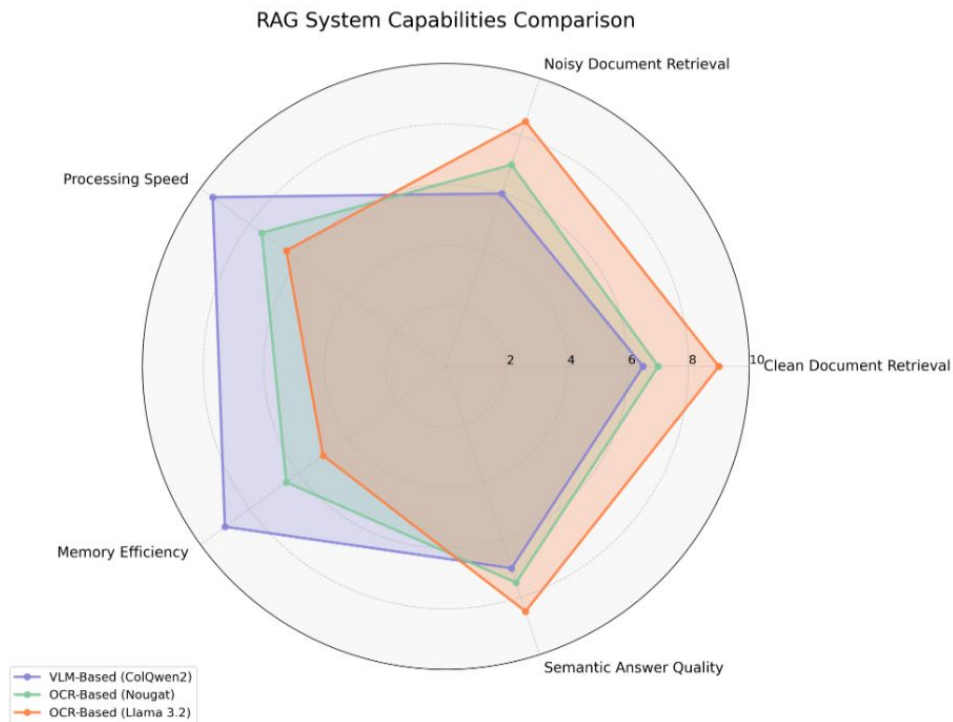
cleaner document conditions → *ViDoRe-DocVQA*

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