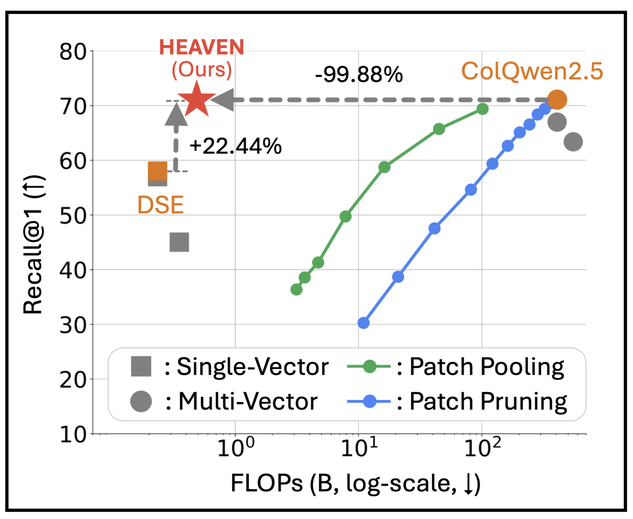
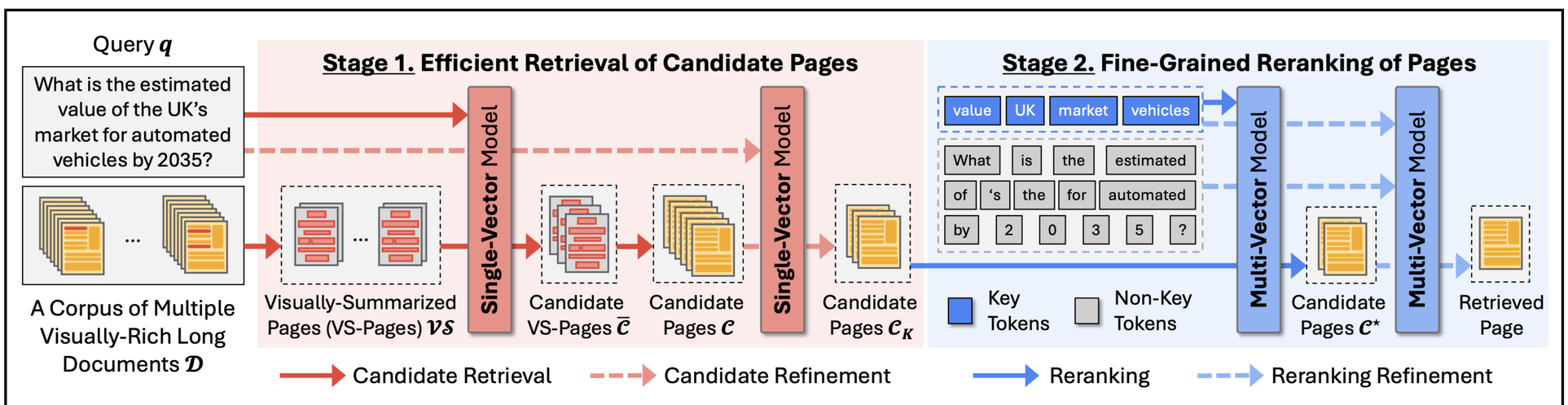
Hybrid-Vector Retrieval for Visually Rich Documents: Combining Single-Vector Efficiency and Multi-Vector Accuracy

Juyeon Kim — KAIST — juyeonkim@kaist.ac.kr
Geon Lee — KAIST — geonlee0325@kaist.ac.kr
Dongwon Choi — KAIST — cookie000215@kaist.ac.kr
Taeuk Kim — Hanyang University — kimtaeuk@hanyang.ac.kr
Kijung Shin — KAIST — kijungs@kaist.ac.kr

ACL 2026 SAN DIEGO JULY 2-7



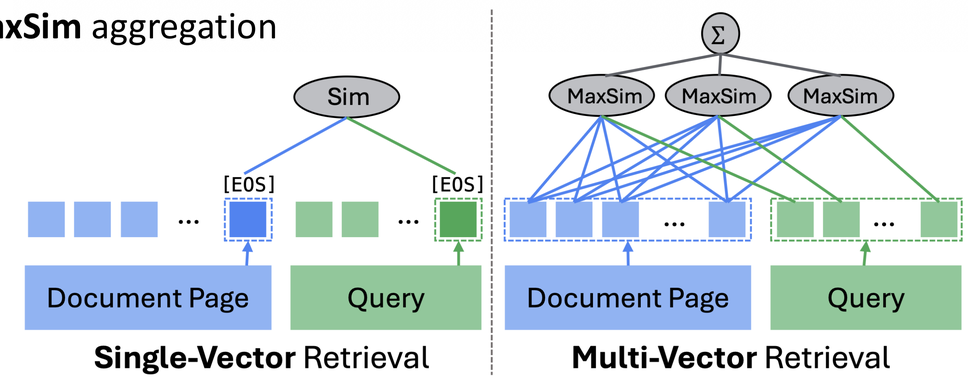


Summary

- **Visual Document Retrieval (VDR)**: Retrieving relevant pages from a corpus of **visually rich documents** (pages contain charts, tables, and figures)
- **Key Observations**
- **O1. Single-Vector** is **significantly more efficient** than Multi-Vector
- **O2. Single-Vector** is **acceptable** for **coarse-grained retrieval**, but **insufficient** for **fine-grained retrieval**.
- **Proposed Method: HEAVEN**
- **Plug-and-play Two-stage Hybrid-Vector** Retrieval framework
- **Stage 1. Single-Vector** candidate retrieval using **VS-pages**
- **Stage 2. Multi-Vector** reranking using **Key tokens**
- **Proposed Benchmark: ViMDoc**
- Benchmark for **Visually rich + Multi**-document + **Long**-document Retrieval
- Enables evaluation over a **Large Search Space**
- **Experimental Result**
- **99.87%** of ColQwen2.5 R@1 at only **0.18%** of its FLOPs (−99.82% FLOPs)

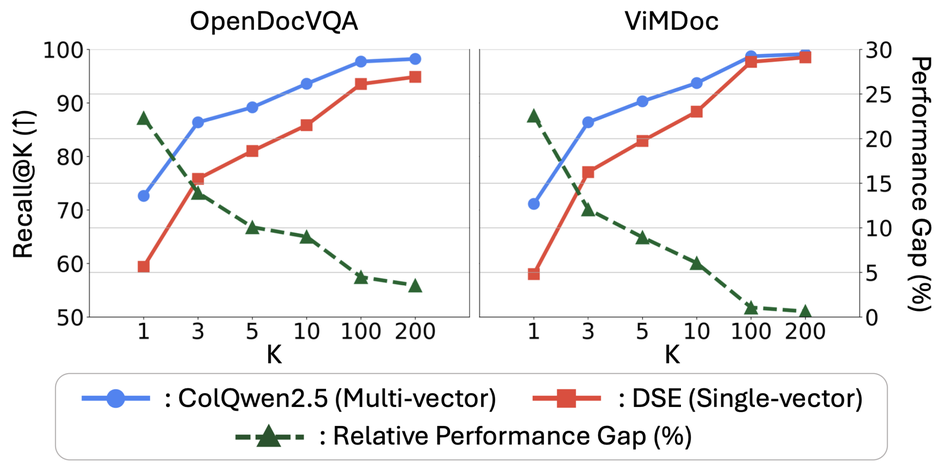
Preliminary

- **Visual Document Retrieval (VDR):** Retrieve relevant pages from a document corpus with **visually rich content** (charts, tables, figures)
- **Traditional**: OCR or complex layout parsing → process into **text**
- **Recent LVLMs**: Encode each page **directly as an image** → no OCR needed
- **Retrieval Frameworks:**
- **Single-Vector Retrieval: Single** embedding per Query/Page
- **Multi-Vector Retrieval**: **Multiple** Token/Patch embeddings per Query/Page
→ **MaxSim** aggregation



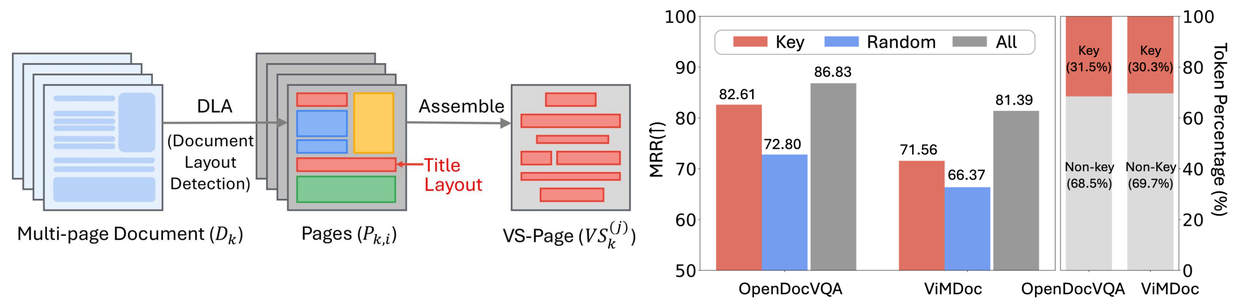
Key Observations

- **[Observation 1] Efficiency-Accuracy Tradeoff**
- **Single-Vector** retrieval is efficient (up to **99.94% fewer FLOPs**) but **coarse**
- **Multi-Vector** retrieval is accurate but **computationally expensive**
- **[Observation 2] Single-Vector is Acceptable for Coarse Retrieval**
- The **performance gap narrows** from 22.5% at K=1 to only 0.63% at K=200



Proposed Method: HEAVEN

- **[Stage 1] Single-Vector Retrieval of Candidate Pages**
- **Visually-Summarized pages (VS-pages):** Compress **multiple pages** into **single VS-pages** by assembling representative **title layouts**
- [1] **Candidate VS-page Retrieval** → $\mathcal{C}$
- [2] **Candidate Refinement**: Combine VS-page + page scores → $\mathcal{C}_K$
- **[Stage 2] Multi-Vector Reranking of Pages**
- **Key tokens: Filtered query tokens** by **linguistic importance** (≈30%)
- [1] **Reranking with Key Tokens** → $\mathcal{C}^*$
- [2] **Reranking Refinement**: Combine Stage 1 + Stage 2 scores



Proposed Benchmark: ViMDoc

- Benchmark for **visually rich**, **multi-document**, and **long-document** retrieval
- **Query Processing: Two-stage Filtering** retains only **self-contained** queries

Benchmarks (Test Split)	Document Page #Total	Document Page Avg./Doc	Query #Total	Query #Cross	Multi-Doc	Long-Doc	Cross-Query
MP-DocVQA	6.2k	6.5	5.0k	-			
VisR-Bench	24.2k	18.5	35.6k	-			
SlideVQA	8.0k	20.0	2.2k	0.6k			✓
MMDocIR	20.4k	65.1	1.7k	0.3k		✓	✓
MMLongBench-Doc	6.5k	47.5	1.1k	0.4k		✓	✓
LongDocURL	33.9k	85.6	2.3k	1.2k		✓	✓
ViDoRE	8.3k	1.0	3.8k	-	✓		
ViDoSeek	5.4k	18.4	1.1k	-	✓		
REAL-MM-RAG	8.6k	52.8	4.6k	-	✓	✓	
M3DocVQA	41.1k	12.2	2.4k	†	✓		✓
VisDoM	21.0k	16.4	2.3k	†	✓		✓
OpenDocVQA	106.7k	3.1	2.5k	0.3k	✓		✓
ViMDoc (Ours)	76.3k	55.4	10.9k	0.7k	✓	✓	✓

Benchmarks	Avg. Search Space # Page	Avg. Search Space # Doc	Corpus
VisR-Bench	18.5	1.0	Single-Doc
MMDocIR	65.1	1.0	Single-Doc
MMLongBench-Doc	47.5	1.0	Single-Doc
LongDocURL	85.6	1.0	Single-Doc
REAL-MM-RAG	2151.0	40.8	Multi-Doc
ViMDoc (Ours)	76347.0	1379.0	Multi-Doc

→ Enables evaluation over a **large, multi-document search space**

Experimental Results

- **Main Result**
- **99.87% of ColQwen2.5 R@1** while reducing **FLOPs by 99.82%**

	Model	ViMDoc (Proposed) R@1	ViMDoc (Proposed) R@3	ViMDoc (Proposed) FLOPs	OpenDocVQA R@1	OpenDocVQA R@3	OpenDocVQA FLOPs	ViDoSeek R@1	ViDoSeek R@3	ViDoSeek FLOPs	M3DocVQA R@1	M3DocVQA R@3	M3DocVQA FLOPs	AVERAGE R@1	AVERAGE R@3	AVERAGE FLOPs
Single-Vector	BGE-M3 (dense)	44.89	61.62	0.156	38.67	52.45	0.165	54.82	76.01	0.011	57.56	77.11	0.084	48.99	66.80	0.104
	NV-Embed-V2	50.06	69.08	0.625	47.37	64.59	0.658	66.46	83.89	0.044	62.65	85.77	0.336	56.64	75.83	0.416
	VisRAG	45.03	64.49	0.352	51.41	67.39	0.370	63.31	84.33	0.025	50.39	69.59	0.189	52.53	71.45	0.234
	GME	57.01	76.62	0.235	54.19	71.93	0.247	71.19	89.40	0.017	59.85	77.83	0.126	60.56	78.95	0.156
	DSE	58.03	77.08	0.235	59.38	75.82	0.247	69.53	87.13	0.017	55.14	71.30	0.126	60.52	77.83	0.156
	HEAVEN (only Stage 1)	57.64	76.58	0.134	59.89	76.05	0.147	69.35	87.39	0.010	59.40	73.96	0.026	61.57	78.50	0.079
	(vs. DSE)	99.34%	99.35%	-42.76%	100.87%	100.32%	-40.47%	99.75%	100.30%	-42.40%	107.73%	103.73%	-79.67%	101.74%	100.86%	-49.31%
Multi-Vector	BGE-M3 (multi)	53.60	70.66	5863.387	46.60	62.80	888.097	56.74	76.01	30.777	56.88	76.96	1408.110	53.46	71.61	2047.590
	ColPali	63.38	80.58	669.670	67.50	82.13	665.152	66.29	84.41	57.086	58.01	78.65	398.295	63.79	81.44	447.551
	ColQwen2	66.99	82.51	407.320	72.27	86.14	482.049	75.13	90.63	41.514	59.32	80.69	288.507	68.43	84.99	304.847
	ColQwen2.5	71.13	86.39	407.320	72.63	86.38	482.049	75.57	91.94	41.514	57.99	78.73	288.507	69.33	85.86	304.847
	HEAVEN	71.05	86.41	0.486	71.56	84.53	0.541	75.04	91.33	0.623	59.31	78.66	0.545	69.24	85.23	0.549
	(vs. ColQwen2.5)	99.88%	100.02%	-99.88%	98.52%	97.86%	-99.89%	99.30%	99.33%	-98.50%	102.27%	99.90%	-99.81%	99.87%	99.27%	-99.82%

- **Comparison with Efficiency-oriented Variants of Multi-Vector Models**
- HEAVEN achieves the **best efficiency-accuracy trade-off**