

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



Juyeon Kim



Geon Lee



Dongwon Choi



Taeuk Kim



Kijung Shin



Contents

- **Preliminary**
- **Motivation**
- **Proposed Method: HEAVEN**
- **Proposed Benchmark: ViMDoc**
- **Experiments**



Preliminary: Visual Document Retrieval

Visual Document Retrieval (VDR): Retrieve relevant pages from a corpus of visually rich documents

- Real-world documents are **rich with visual elements** (e.g., charts, tables, figures).
- Cannot be captured by **text alone**.

Table of Contents

Item 7: Management's Discussion and Analysis of Financial Condition and Results of Operations

Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A) is designed to provide a reader of TM's financial statements with a narrative from the perspective of management. TM's MD&A is presented in eight sections:

- Overview
- Results of Operations
- Performance by Business Segment
- Performance by Geographic Area
- Critical Accounting Estimates
- New Accounting Pronouncements
- Financial Condition and Liquidity
- Financial Instruments

Forward-looking statements in Item 7 may involve risks and uncertainties that could cause results to differ materially from those projected (refer to the section entitled "Cautionary Note Concerning Factors That May Affect Future Results" in Item 1 and the risk factors provided in Item 1A for discussion of these risks and uncertainties).

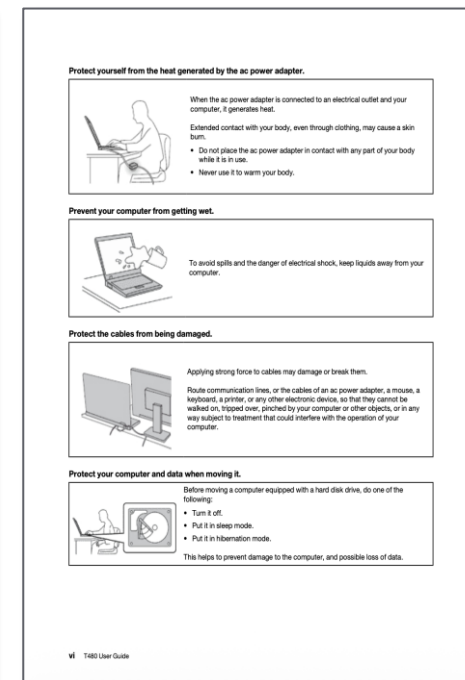
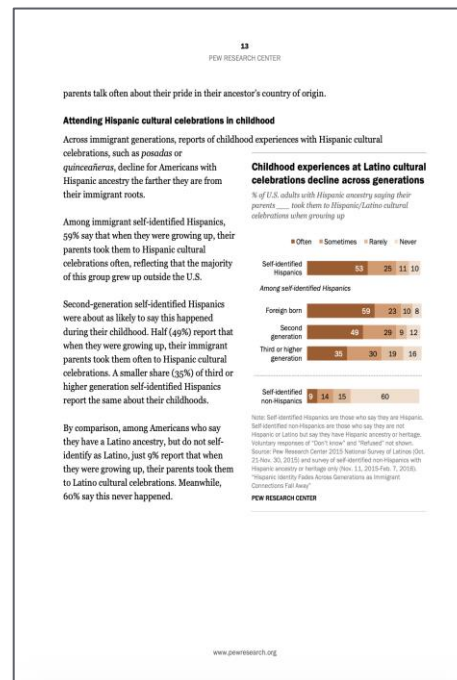
OVERVIEW

TM is a diversified global manufacturer, technology innovator and marketer of a wide variety of products and services. As described in Item 1B, effective in the first quarter of 2018, TM improved the alignment of its business around markets and customers. Segment information presented herein reflects the impact of these changes for all periods presented. TM manages its operations in five operating business segments: Industrial, Safety and Graphics, Health Care, Electronics and Energy, and Consumer. From a geographic perspective, any reference to EMEA refers to Europe, Middle East and Africa on a combined basis.

Earnings per share (EPS) attributable to TM common shareholders - diluted

The following table provides the increase (decrease) in diluted earnings per share for the fourth quarter and year 2018 compared to the same period last year, in addition to 2017 compared to 2016.

	Three months ended December 31, 2018	2018	2017
Diluted earnings per share			
Same period last year	\$ 0.82	\$ 0.82	\$ 0.82
2017 Earnings of TCCA Impact	1.28	1.24	1.24
Same period last year, excluding 2017 Tax Cuts and Jobs Act (TCJA)	\$ 1.28	\$ 1.24	\$ 1.24
Increase (decrease) in earnings per share - diluted, due to:			
2017 Earnings of TCCA Impact	—	(0.42)	0.34
Operating productivity and other	0.18	0.92	0.47
Acquisition of the former TCCA	(0.05)	(0.15)	—
Foreign exchange impacts	(0.03)	(0.03)	(0.13)
Legal-related charges	—	(0.44)	—
Legal - regulator seek actualized reserve	—	(0.16)	(0.07)
Other adjustment	0.06	0.18	(0.22)
Increase in tax, excluding Tax Cuts and Jobs Act (TCJA) measurement period adjustment	0.05	0.43	0.34
TCJA measurement period adjustment	0.08	0.18	0.08
Share of common stock outstanding	—	—	—
2018 Earnings of Communications Markets Division, net of related restructuring actions	0.02	0.02	—
Current period, excluding MNC Natural Resource Damage (NRD)	\$ 1.20	\$ 1.46	\$ 1.17
TCJA measurement period adjustment	0.07	(0.29)	(1.24)
MNC NRD resolution	(0.11)	(1.28)	—
Current period	\$ 1.16	\$ 1.07	\$ 1.03



Preliminary: Visual Document Retrieval (Cont.)

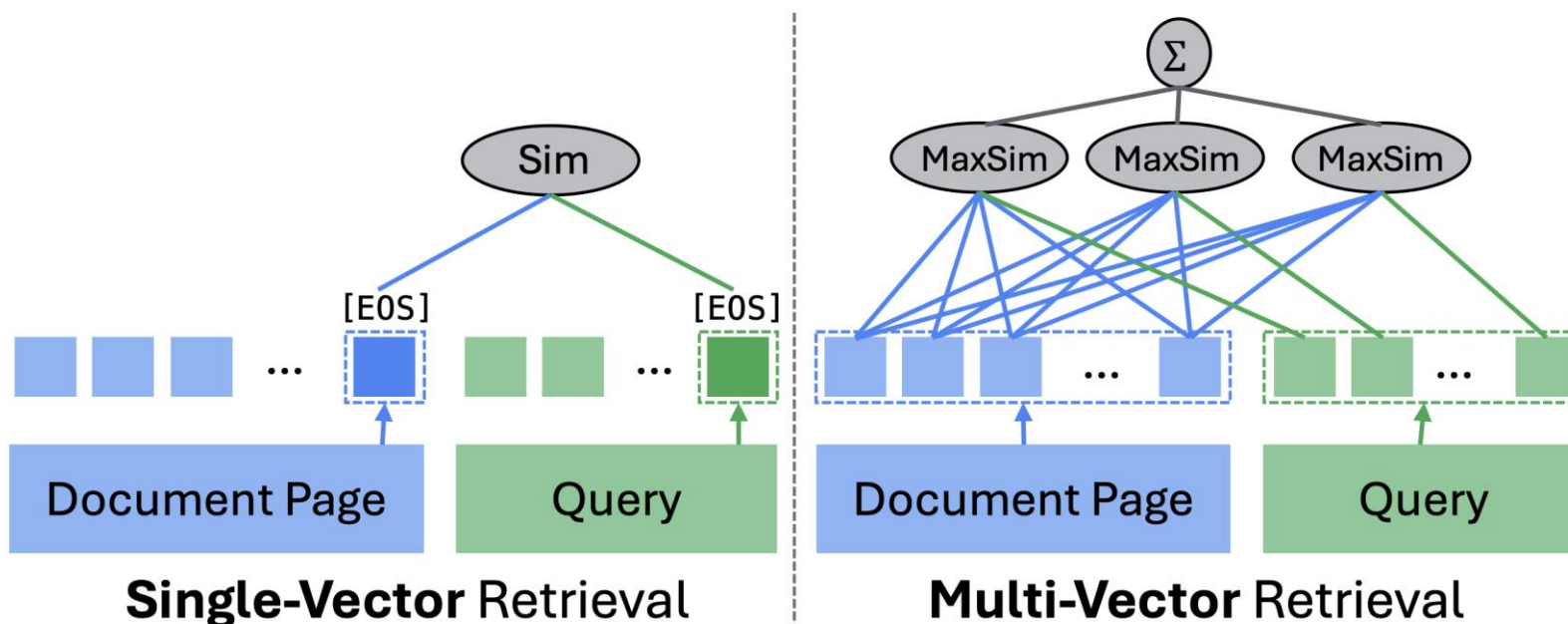
Visual Document Retrieval (VDR)

- Traditional methods: **OCR** or **complex layout parsing** → process **visual content** into **text**
- **Recent LVLMs**: encode each page **directly as an image** → **visual embeddings** for retrieval
→ No **OCR** or **complex layout parsing** needed!



Preliminary: Retrieval Framework

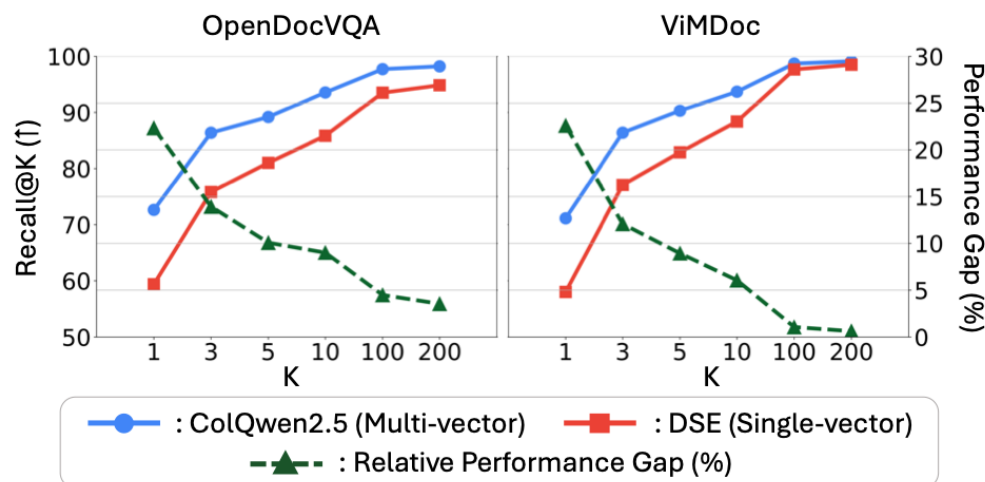
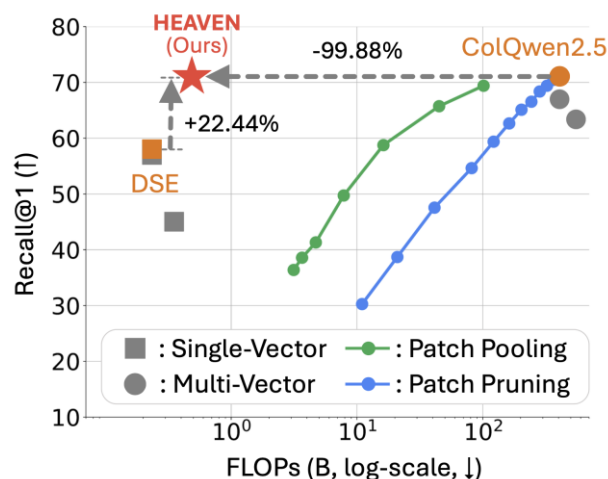
- **Single-Vector Retrieval:** **Single** embedding per Query/Page
- **Multi-Vector Retrieval:** **Multiple** Token/Patch embeddings per Query/Page → **MaxSim** aggregation



Motivation

Two Key Observations

- [Obs 1] **Single-Vector Retrieval** is **More Efficient** than **Multi-Vector Retrieval**
 - Up to **99.94%** fewer FLOPs per query
 - [Obs 2] **Single-Vector Retrieval** can be acceptable for **Coarse-Grained Retrieval**.
- The **performance gap** narrows from **22.5%** at K=1 to only **0.63%** at K=200

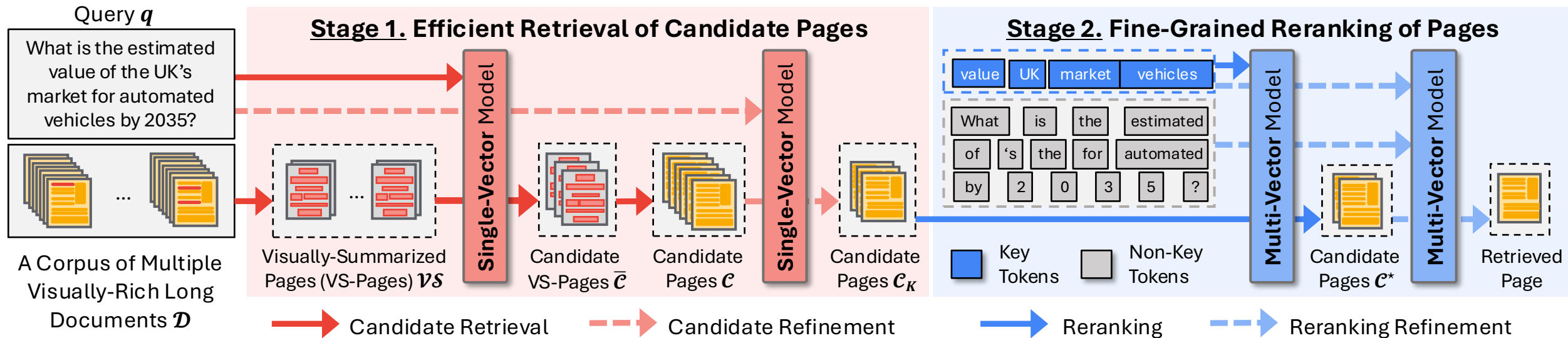


→ Combining **Single-Vector Efficiency** and **Multi-Vector Accuracy**



Proposed Method: HEAVEN

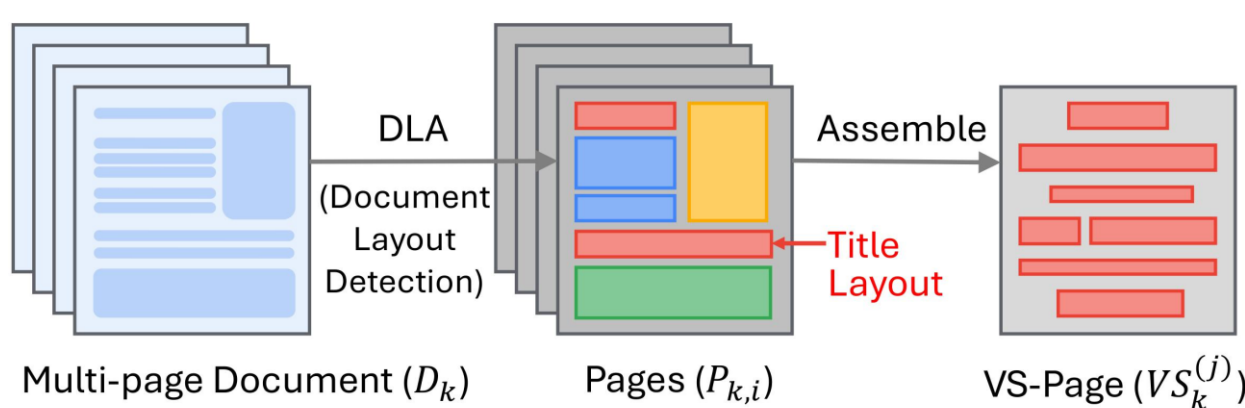
- **HEAVEN: Plug-and-Play Hybrid-Vector Two-stage** Retrieval framework
 - Stage 1. **Single-Vector** Retrieval of Candidate Pages
 - Stage 2. **Multi-Vector** Reranking of Pages



Proposed Method: HEAVEN (cont.)

- 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**
- [Motivation]** Most pages are **irrelevant** or contain **repetitive/uninformative** content – only a **few elements** are **actually informative**



NBC Sunday Night Football

NBC Sunday Night Football

Reporters:

Studio show [edit]

Related

Production

Original release

Main article: [Football Night in America](#)

Opening game [edit]

Contract [edit]

Scheduling [edit]

Connections to other sports [edit]

World Series conflicts [edit]

Flexible scheduling [edit]

Automation with CLI scripting

3.1 Secure restricted shell

- Non-interactive shell
- Interactive shell
- Implement error checking

3.2 General tips

- Automating IBM storage FlashSystem
- Prefixes svcinfo and svctask
- Automating IBM storage FlashSystem
- Remove unwanted information
- Selectable delimiter for outputs
- Extended data tables
- Automating IBM storage FlashSystem

3.3 Internal scripting commands

- Control command rate
- Loops: for and while

3.4 Script examples

- Data parsing
- Automating IBM storage FlashSystem

Automation with REST API

API limits

4.1.1 REST API Explorer

4.1 REST API on IBM Storage Virtualize

Proposed Method: HEAVEN (cont.)

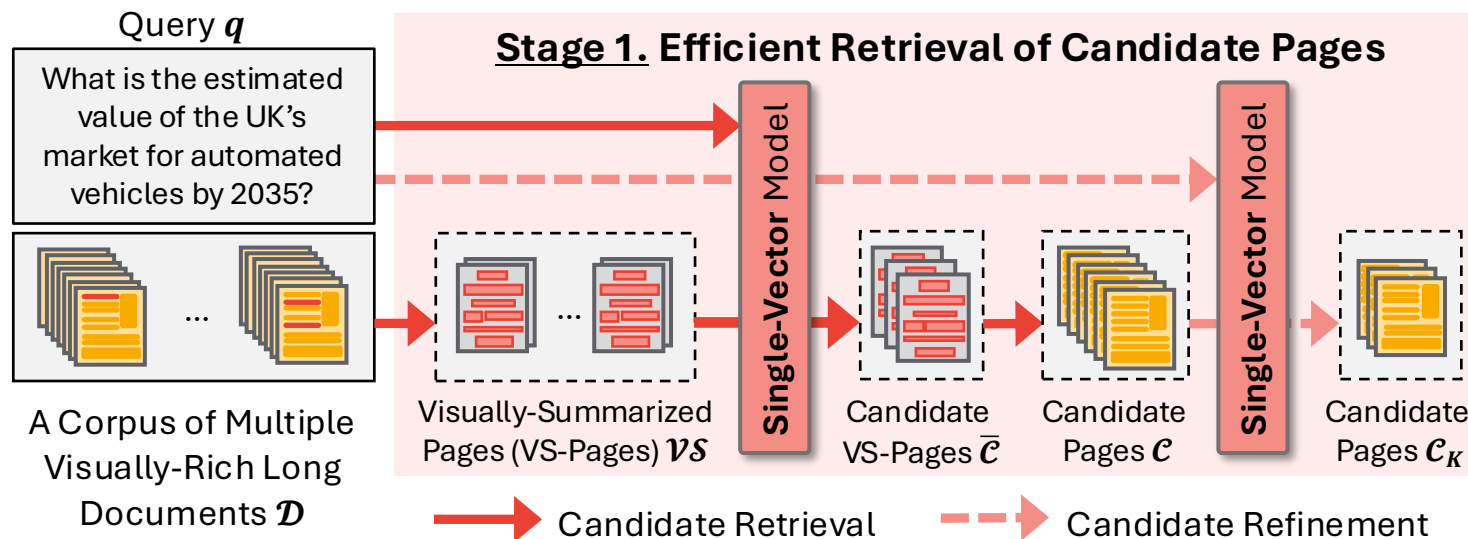
- Stage 1. Single-Vector Retrieval of Candidate Pages

- [1] Candidate VS-page Retrieval

→ Candidates $\mathcal{C}$

- [2] Candidate Page Refinement: combine VS-page + page scores

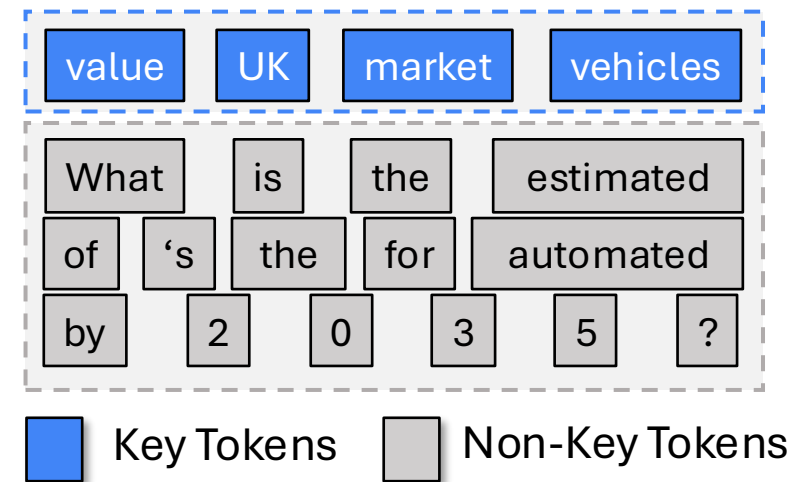
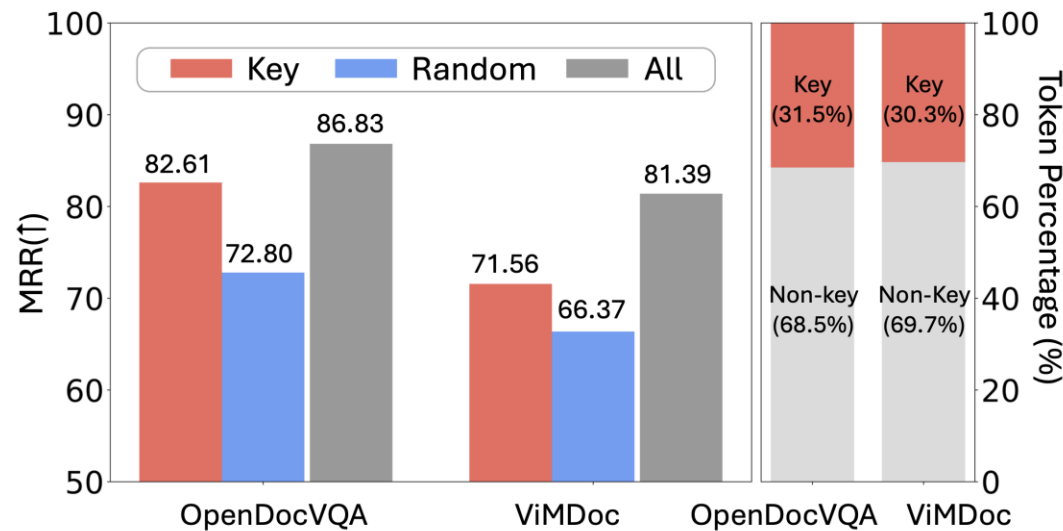
→ Candidates $\mathcal{C}_K$



Proposed Method: HEAVEN (cont.)

- Stage 2. Multi-Vector Reranking of Pages

- Key Token:** Filtered query tokens by **linguistic importance** (nouns/named entities)
- [Motivation]** Many query tokens (e.g., stopwords) are **less informative** for the task
 - computing similarities for all tokens causes **redundant computation**



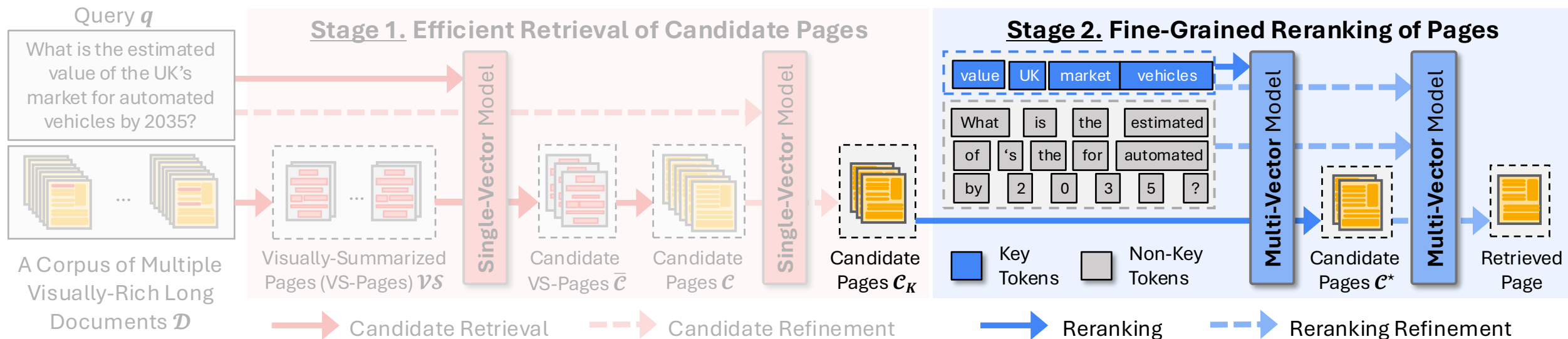
→ Reasonably **effective**, despite their **small proportion (~30%)** within queries



Proposed Method: HEAVEN (cont.)

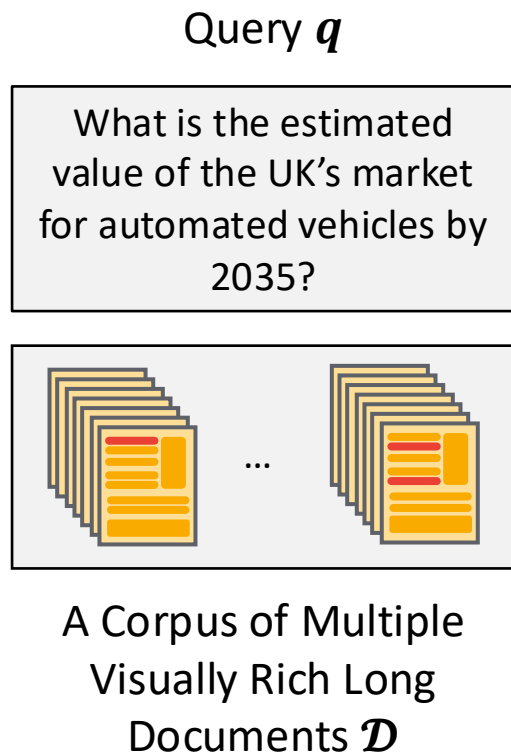
- **Stage 2. Multi-Vector Reranking of Pages**

- [1] **Reranking** with Filtered Query Tokens $\rightarrow$ Candidates $\mathcal{C}^*$
- [2] **Reranking Refinement**: combine Stage 1 + Stage 2 scores $\rightarrow$ Final Page Ranking



Proposed Benchmark: ViMDoc

- ViMDoc: Benchmark for **Visually Rich**, **Multi-Document**, and **Long-Document** retrieval



Benchmarks (Test Split)	Document Page		Query		Multi- Doc	Long- Doc	Cross- Query
	#Total	Avg./Doc	#Total	#Cross			
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	✓	✓	✓



Proposed Benchmark: ViMDoc (cont.)

- **ViMDoc**: Benchmark for **Visually Rich**, **Multi-Document**, and **Long-Document** retrieval
 - **Challenge**: Most sourced queries are **Context-dependent** (e.g., “What is the title?”)
 - **Unsuitable** for Multi-document Retrieval
 - **Two-stage Query Filtering** (Heuristic + LLM) → Retain only **Self-contained** queries
 - Enables evaluation over a **large, multi-document** search space

Benchmarks	Avg. Search Space		Corpus
	# Page	# Doc	
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



Experimental Result

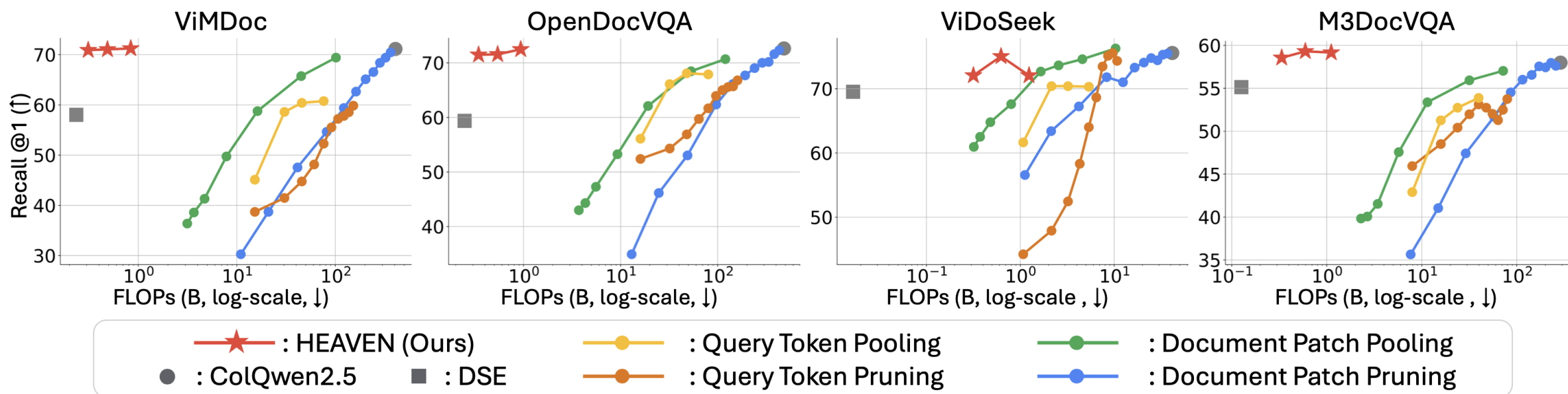
- **[1] Main Result:** HEAVEN preserves **99.87%** of the state-of-the-art retrieval performance of ColQwen2.5 (in terms of Recall@1) while **reducing per-query FLOPs by 99.82%**

		ViMDoc (Proposed)			OpenDocVQA			ViDoSeek			M3DocVQA			AVERAGE		
Model		R@1	R@3	FLOPs	R@1	R@3	FLOPs	R@1	R@3	FLOPs	R@1	R@3	FLOPs	R@1	R@3	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%



Experimental Result (cont.)

- [2] Comparison with Efficiency-oriented Variants of Multi-Vector model:
 - Patch Pooling, Patch Pruning, Token Pooling, Token Pruning
- In the **accuracy-efficiency trade-off**, HEAVEN achieves **higher accuracy** with **significantly fewer FLOPs**.



Experimental Result (cont.)

- **[3] Ablation Study:** All components are effective
 - removing any component **degrades either efficiency or accuracy.**

	ViMDOC (Proposed)			OpenDocVQA		
Stage 1	R@100	R@200	FLOPs	R@100	R@200	FLOPs
HEAVEN	97.20	97.96	0.134	93.02	94.59	0.147
w/o VS-pages	97.68	98.51	0.235	93.54	94.86	0.247
w/o refinement	59.47	68.02	0.017	59.47	68.02	0.024
Stage 2	R@1	R@3	FLOPs	R@1	R@3	FLOPs
HEAVEN	71.05	86.41	0.486	71.56	84.53	0.541
w/o query filtering	71.08	86.38	0.871	71.76	85.10	0.957
w/o refinement	58.05	77.08	0.358	59.53	75.82	0.402



Experimental Result (cont.)

- **[4] Plug-and-play Analysis:** Consistent performance across **different encoder combinations**, confirming HEAVEN's **modular design**.

Base Encoder		VIMDOC (Proposed)			OpenDocVQA		
Stage1	Stage2	R@1	R@3	FLOPs	R@1	R@3	FLOPs
VisRAG	ColPali	60.95	79.19	1.298	65.27	80.02	1.287
VisRAG	ColQwen2	62.45	80.07	0.553	67.79	81.28	0.615
VisRAG	ColQwen2.5	64.06	82.07	0.553	67.54	81.40	0.615
DSE	ColPali	63.46	80.68	1.231	67.50	81.81	1.214
DSE	ColQwen2	66.63	82.63	0.486	69.91	83.31	0.541
DSE	ColQwen2.5	71.05	86.41	0.486	71.56	84.53	0.541



Experimental Result (cont.)

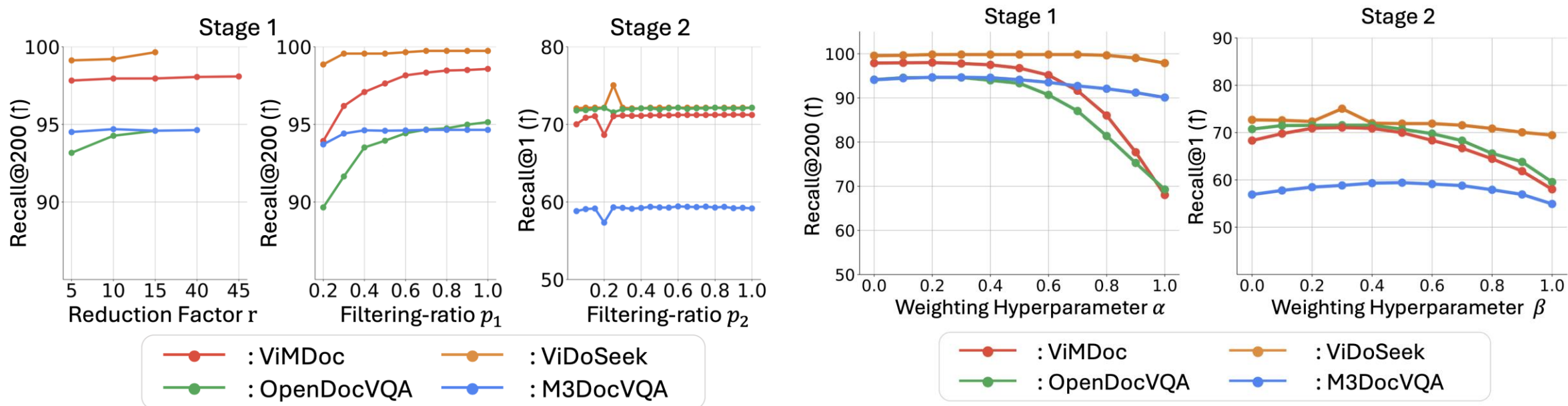
- **[5] Efficiency Analysis:** HEAVEN achieves **99.88% faster online retrieval latency** than ColQwen2.5.

		ViMDoC (Proposed)				
		Offline Indexing (min)			Online Retrieval (sec, B)	
		OCR	VS-page	Encode	Latency	FLOPs
Single	NV-Embed-V2 	473.0	NA	99.5	0.133	0.625
	DSE 	NA	NA	146.7	0.115	0.235
Multi	BGE-M3 (multi) 	473.0	NA	32.3	20953.468	5863.387
	ColQwen2.5 	NA	NA	129.1	2006.361	407.320
HEAVEN		NA	58.9	286.4	2.412	0.486
Stage 1		NA	58.9	157.3	0.079	0.134
Stage 2		NA	NA	129.1	2.333	0.352



Experimental Result (cont.)

- [6] **Hyperparameter Analysis:** HEAVEN is **robust to hyperparameter choices** across reduction factor r , filtering ratios p_1/p_2 , and weighting parameters α/β .



Takeaway

- **Single-Vector** is **Efficient** but **Coarse**; **Multi-Vector** is **Accurate** but **Computationally Expensive**
- **HEAVEN** bridges the **Efficiency-Accuracy Tradeoff**
 - **99.87%** of ColQwen2.5 **Recall@1** at **99.82% fewer FLOPs**
- **VimDoc** benchmark for **Visually Rich**, **Multi-Document**, and **Long-Document** retrieval
 - Enables Evaluation over a **Large, Multi-Document Search Space**



Thank You !

Paper



GitHub



HF Dataset

