Multimodal RAG in Long-Context DocVQA

120425 Weekly Seminar

심규호

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

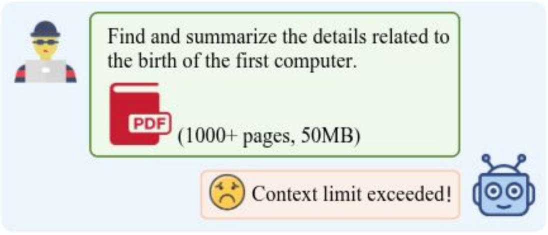
Multimodal RAG in Long-Context Document Understanding

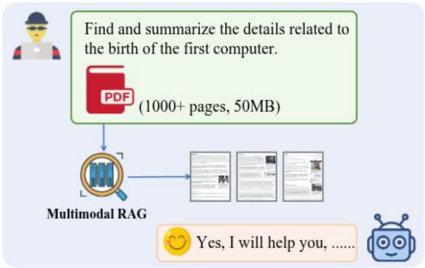
Document Understanding

1. ***LLMs & Text-based RAG methods***
   a. *Convert the document (e.g., via OCR) into text for processing*
   b. *Strip away critical multimodal information (e.g., figures)*
2. ***LVLMs (Large Vision-Language Models)***
   a. Enhanced understanding of multi-modal information
   b. Constrained input size →*Suffer from multi-page document comprehension*

⇒ *Multimodal-RAG methods*

- *Image representation*
- *cross-modal representation (text + image)*





Multimodal RAG in Long-Context Document Understanding

Overview

Image-based Multimodal RAG

1. Collections of **PDFs/Documents**
2. Conversion to **IMGs** (e.g., PDF2Img)
3. **Retrieval**
   a. Page-Query Relevance (ColPali)
4. **Generation** (LVLM)

(a) Image based Multimodal RAG

$$D = \{d_i\}_{i=1}^{N}$$

$$z_i^{\text{img}} = \text{Enc}_{\text{img}}(d_i) \quad e_q^{\text{text}} = \text{Enc}_{\text{text}}(q)$$

$$s_{\text{img}}(e_q, z_i) = \langle e_q^{\text{text}}, z_i^{\text{img}} \rangle$$

$$X_{\text{img}} = \{\, d_i \in D \mid s_{\text{img}}(e_q, z_i) \geq \tau_{\text{img}} \,\}$$

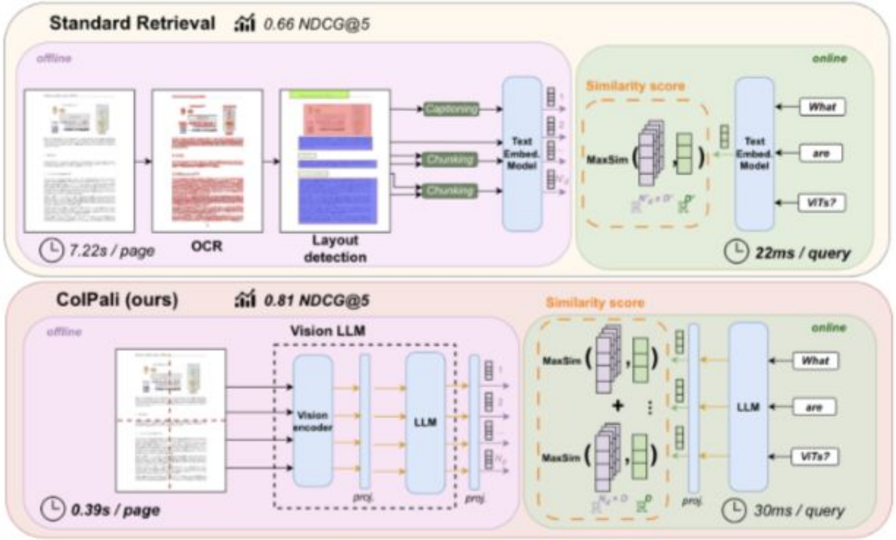
Multimodal RAG in Long-Context Document Understanding

ColPali-based Multimodal Retrieval

ColPali - *Retrieval in Vision Space*

1. ***Encode Query***
2. ***Late Interaction Mechanism***

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



M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding

Jaemin Cho[1*] Debanjan Mahata[2] Ozan İrsoy[2] Yujie He[2] Mohit Bansal[1]
[1]UNC Chapel Hill [2]Bloomberg
{jmincho,mbansal}@cs.unc.edu {dmahata,oirsoy,yhe247}@bloomberg.net

2025 ICCV Workshop

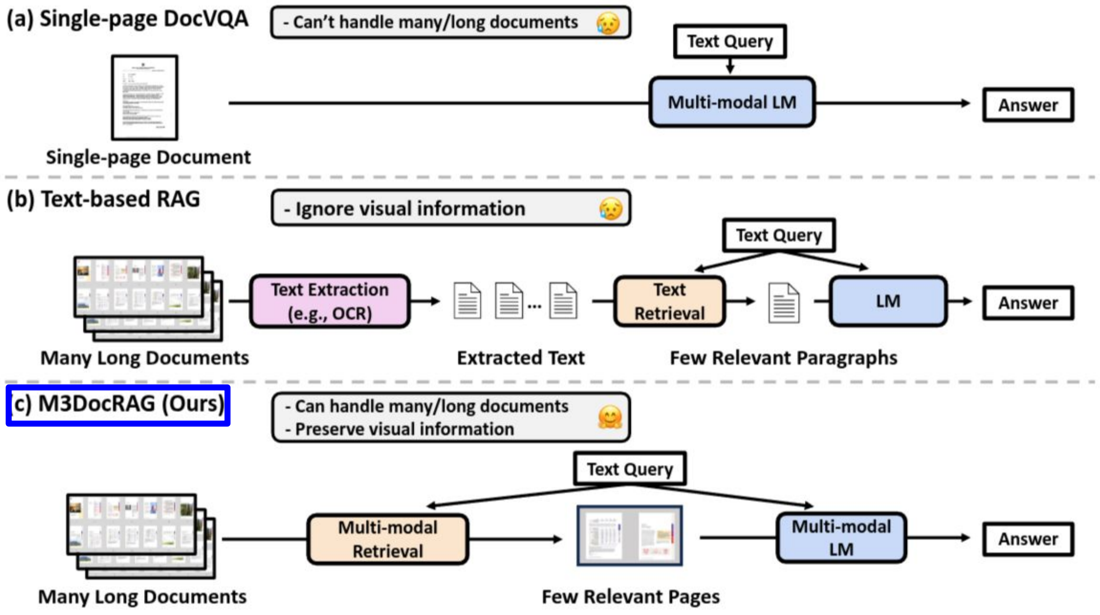
M3DocRAG

Real-world Document Understanding Scenarios

Framework

1. Information across different pages or documents
   a. existing VQA methods cannot handle many long documents
2. Complex Visual Formats
   a. tables, charts, mixed layouts

Accurately & Efficiently answering questions across numerous, lengthy documents w/intricate layouts



⇒ **M3DocRAG**

Multi-modal **M**ulti-page **M**ulti-**Doc**ument **R**etrieval-**A**ugmented **G**eneration

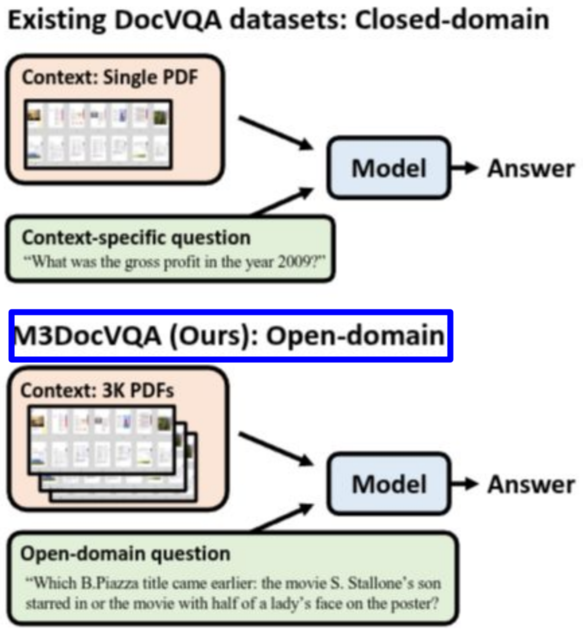
M3DocRAG

Real-world Document Understanding Scenarios

Dataset

1. Existing DocVQA datasets are *not adequate for* **open-domain setting**
   a. **Closed domain:** grounding to a single source document
   b. **Open domain:** searching a large corpus

*Large ‘**haystack**’ of multi-modal documents & retrieve relevant information to generate the final answer*
→ 2,441 multi-hop questions, 3,368 PDF docs, 41,005 pages

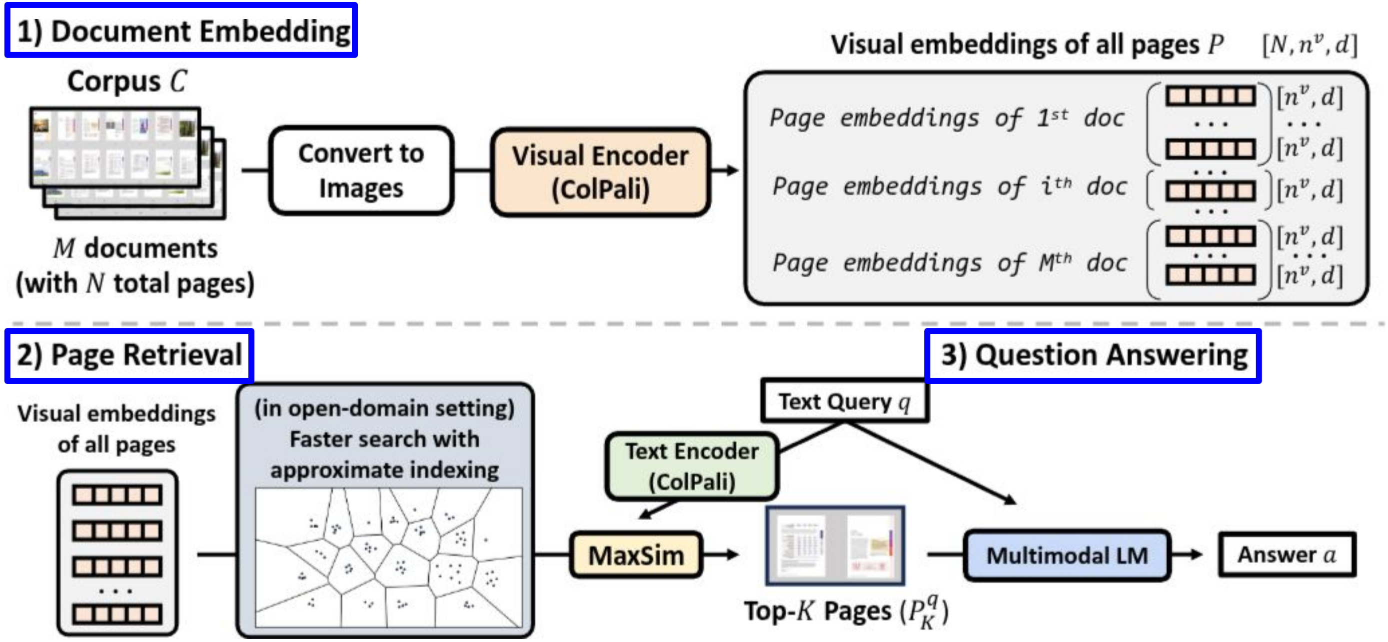


⇒ **M3DocVQA**

Multi-modal **M**ulti-page **M**ulti-**Doc**ument **V**isual **Q**uestion **A**nswering

M3DocRAG

M3DocRAG framework



M3DocRAG

Experiments - M3DocvQA (Open-Domain)

Method	# Pages	Evidence Modalities			Question Hops		Overall	
		Image	Table	Text	Single-hop	Multi-hop	EM	F1
Text RAG (w/ ColBERT v2)								
Llama 3.1 8B	1	8.3	15.7	29.6	25.3	12.3	15.4	20.0
Llama 3.1 8B	2	7.7	16.8	31.7	27.4	12.1	15.8	21.2
Llama 3.1 8B	4	7.8	21.0	34.1	29.4	15.2	17.8	23.7
M3DocRAG *(w/ ColPali)*								
Qwen2-VL 7B (Ours)	1	25.1	27.8	39.6	37.2	25.0	27.9	32.3
Qwen2-VL 7B (Ours)	2	**26.8**	**30.4**	**42.1**	41.0	25.2	29.9	34.6
Qwen2-VL 7B (Ours)	4	24.7	**30.4**	41.2	**43.2**	**26.6**	**31.4**	**36.5**

M3DocRAG

Experiments - MMLongBench-Doc (Closed-domain)

Method	# Pages	Evidence Modalities					Evidence Locations			Overall	
		TXT	LAY	CHA	TAB	IMG	SIN	MUL	UNA	ACC	F1
Text Pipeline											
LMs											
ChatGLM-128k [5]	up to 120	23.4	12.7	9.7	10.2	12.2	18.8	11.5	18.1	16.3	14.9
Mistral-Instruct-v0.2 [25]	up to 120	19.9	13.4	10.2	10.1	11.0	16.9	11.3	24.1	16.4	13.8
Text RAG											
ColBERT v2 + Llama 3.1	1	20.1	14.8	12.7	17.4	7.4	21.8	7.8	**41.3**	21.0	16.1
ColBERT v2 + Llama 3.1	4	23.7	17.7	14.9	**24.0**	11.9	25.7	12.2	38.1	**23.5**	19.7
Multi-modal Pipeline											
Multi-modal LMs											
DeepSeek-VL-Chat [38]	up to 120	7.2	6.5	1.6	5.2	7.6	5.2	7.0	**12.8**	7.4	5.4
Idefics2 [33]	up to 120	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5 [61, 64]	up to 120	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD [15]	up to 120	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5 [22]	up to 120	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat [4]	up to 120	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat [36]	up to 120	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
M3DocRAG											
ColPali + Idefics2 (Ours)	1	10.9	11.1	6.0	7.7	15.7	15.4	7.2	8.1	11.2	11.0
ColPali + Qwen2-VL 7B (Ours)	1	25.7	21.0	18.5	16.4	19.7	30.4	10.6	5.8	18.8	20.1
ColPali + Qwen2-VL 7B (Ours)	4	**30.0**	**23.5**	**18.9**	20.1	**20.8**	**32.4**	**14.8**	5.8	21.0	**22.6**

MMLongBench-Doc

1. **Closed-domain**
2. **Models must handle a long PDF document (up to 120 pages)**
    a. Concatenation strategy that combines all screenshot pages into either 1 or 5 images & inputs hese images to LVLM

M3DocRAG

Experiments - MP-DocVQA (Closed-domain)

Method	Answer Accuracy ANLS	Page Retrieval R@1
Multimodal LMs		
Arctic-TILT 0.8B [10]	0.8122	50.79
GRAM [9]	0.8032	19.98
GRAM C-Former [9]	0.7812	19.98
ScreenAI 5B [3]	0.7711	77.88
Text RAG		
ColBERT v2 + Llama 3.1 8B	0.5603	75.33
M3DocRAG		
ColPali + Qwen2-VL 7B (Ours)	**0.8444**	**81.05**

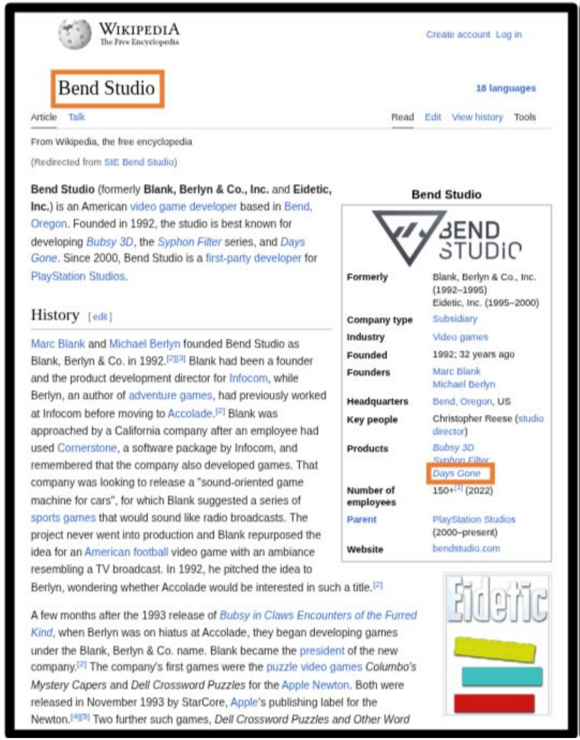
MP-DocVQA

1. **Closed-domain**
2. **Models must handle a long PDF document (up to 20 pages)**
    a. Concatenation strategy that combines all screenshot pages into either 1 or 5 images & inputs hese images to LVLM
3. Existing Entries are fine-tuned specifically for MP-DocVQA

```
Question: “SIE Bend Studio's 2019 game cover has man leaning on what?”

ColPali + Qwen2-VL 7B: “motorcycle”
```

Top 2 pages retrieved by ColPali



WIKIPEDIA The Free Encyclopedia — Create account Log in

Bend Studio

18 languages

Article Talk — Read Edit View history Tools

From Wikipedia, the free encyclopedia

(Redirected from SIE Bend Studio)

Bend Studio (formerly **Blank, Berlyn & Co., Inc.** and **Eidetic, Inc.**) is an American video game developer based in Bend, Oregon. Founded in 1992, the studio is best known for developing *Bubsy 3D*, the *Syphon Filter* series, and *Days Gone*. Since 2000, Bend Studio is a first-party developer for PlayStation Studios.

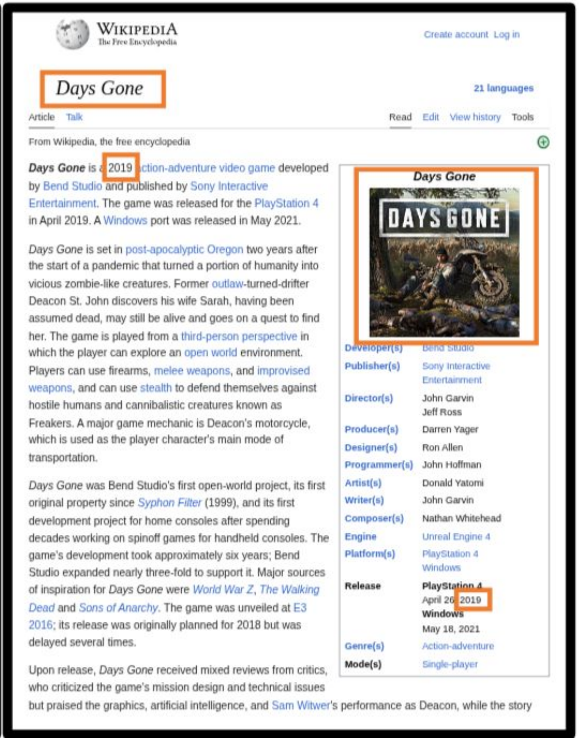
History [edit]

Marc Blank and Michael Berlyn founded Bend Studio as Blank, Berlyn & Co. in 1992.[2][3] Blank had been a founder and the product development director for Infocom, while Berlyn, an author of adventure games, had previously worked at Infocom before moving to Accolade.[2] Blank was approached by a California company after an employee had used Cornerstone, a software package by Infocom, and remembered that the company also developed games. That company was looking to release a "sound-oriented game machine for cars", for which Blank suggested a series of sports games that would sound like radio broadcasts. The project never went into production and Blank repurposed the idea for an American football video game with an ambiance resembling a TV broadcast. In 1992, he pitched the idea to Berlyn, wondering whether Accolade would be interested in such a title.[2]

A few months after the 1993 release of *Bubsy in Claws Encounters of the Furred Kind*, when Berlyn was on hiatus at Accolade, they began developing games under the Blank, Berlyn & Co. name. Blank became the president of the new company.[2] The company's first games were the puzzle video games *Columbo's Mystery Capers* and *Dell Crossword Puzzles* for the Apple Newton. Both were released in November 1993 by StarCore, Apple's publishing label for the Newton.[4][5] Two further such games, *Dell Crossword Puzzles and Other Word*

Bend Studio

Formerly	Blank, Berlyn & Co., Inc. (1992–1995) Eidetic, Inc. (1995–2000)
Company type	Subsidiary
Industry	Video games
Founded	1992; 32 years ago
Founders	Marc Blank Michael Berlyn
Headquarters	Bend, Oregon, US
Key people	Christopher Reese (studio director)
Products	*Bubsy 3D* *Syphon Filter* *Days Gone*
Number of employees	150+[1] (2022)
Parent	PlayStation Studios (2000–present)
Website	bendstudio.com



WIKIPEDIA The Free Encyclopedia — Create account Log in

Days Gone

21 languages

Article Talk — Read Edit View history Tools

From Wikipedia, the free encyclopedia

Days Gone is a 2019 action-adventure video game developed by Bend Studio and published by Sony Interactive Entertainment. The game was released for the PlayStation 4 in April 2019. A Windows port was released in May 2021.

Days Gone is set in post-apocalyptic Oregon two years after the start of a pandemic that turned a portion of humanity into vicious zombie-like creatures. Former outlaw-turned-drifter Deacon St. John discovers his wife Sarah, having been assumed dead, may still be alive and goes on a quest to find her. The game is played from a third-person perspective in which the player can explore an open world environment. Players can use firearms, melee weapons, and improvised weapons, and can use stealth to defend themselves against hostile humans and cannibalistic creatures known as Freakers. A major game mechanic is Deacon's motorcycle, which is used as the player character's main mode of transportation.

Days Gone was Bend Studio's first open-world project, its first original property since *Syphon Filter* (1999), and its first development project for home consoles after spending decades working on spinoff games for handheld consoles. The game's development took approximately six years; Bend Studio expanded nearly three-fold to support it. Major sources of inspiration for *Days Gone* were *World War Z*, *The Walking Dead* and *Sons of Anarchy*. The game was unveiled at E3 2016; its release was originally planned for 2018 but was delayed several times.

Upon release, *Days Gone* received mixed reviews from critics, who criticized the game's mission design and technical issues but praised the graphics, artificial intelligence, and Sam Witwer's performance as Deacon, while the story

Days Gone

Developer(s)	Bend Studio
Publisher(s)	Sony Interactive Entertainment
Director(s)	John Garvin Jeff Ross
Producer(s)	Darren Yager
Designer(s)	Ron Allen
Programmer(s)	John Hoffman
Artist(s)	Donald Yatomi
Writer(s)	John Garvin
Composer(s)	Nathan Whitehead
Engine	Unreal Engine 4
Platform(s)	PlayStation 4 Windows
Release	**PlayStation 4** April 26, 2019 **Windows** May 18, 2021
Genre(s)	Action-adventure
Mode(s)	Single-player

M3DocRAG

Conclusion

1. **M3DocRAG** - RAG Framework that flexibly accommodates various ***document contexts*** *(open & closed-domain)*, ***question hops*** *(single & multi)*, and ***evidence modalities*** *(text, chart, figure, etc.)*
2. **M3DocVQA** - the first benchmark that evaluates open-domain multi-modal document understanding capabilities
3. Robust performance in three datasets: M3DocVQA, MP-DocVQA, MMLongBench-Doc

MoLoRAG: Bootstrapping Document Understanding via Multi-modal Logic-aware Retrieval

Xixi Wu[1], Yanchao Tan[2], Nan Hou[1], Ruiyang Zhang[3], Hong Cheng[1](✉)
[1]The Chinese University of Hong Kong
[2]Fuzhou University [3]University of Macau
{xxwu, nhou, hcheng}@se.cuhk.edu.hk
yctan@fzu.edu.cn, yc47931@um.edu.mo

2025 EMNLP

Main

MoLoRAG

BootStrapping Document Understanding via Multi-modal Logic-aware Retrieval

Framework

1. RAG methods rely solely on **Semantic Relevance**
    a. Ignoring logical connections between pages & query → *Essential for reasoning*

→*Page graph that captures contextual relationships/dependencies between pages*
→*Combination of semantic & Logical Relevance to deliver more accurate retrieval*

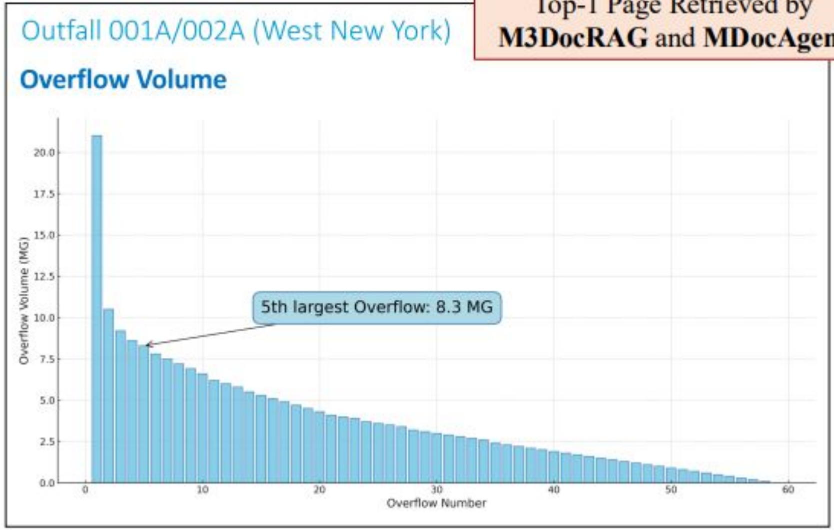
⇒ **MoLoRAG**
Multi-m**o**dal **Lo**gic-aware Document **R**etrieval-**A**ugmented **G**eneration

MoLoRAG

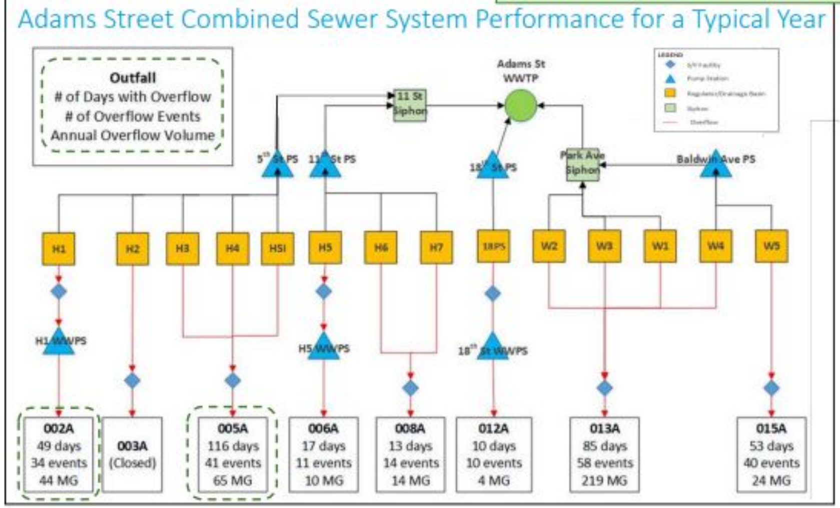
BootStrapping Document Understanding via Multi-modal Logic-aware Retrieval

Question How many days with overflow do Outfall 002A (Southwest Hoboken) and Outfall 005A (Central Hoboken) have in total?

Top-1 Page Retrieved by **M3DocRAG** and **MDocAgent**



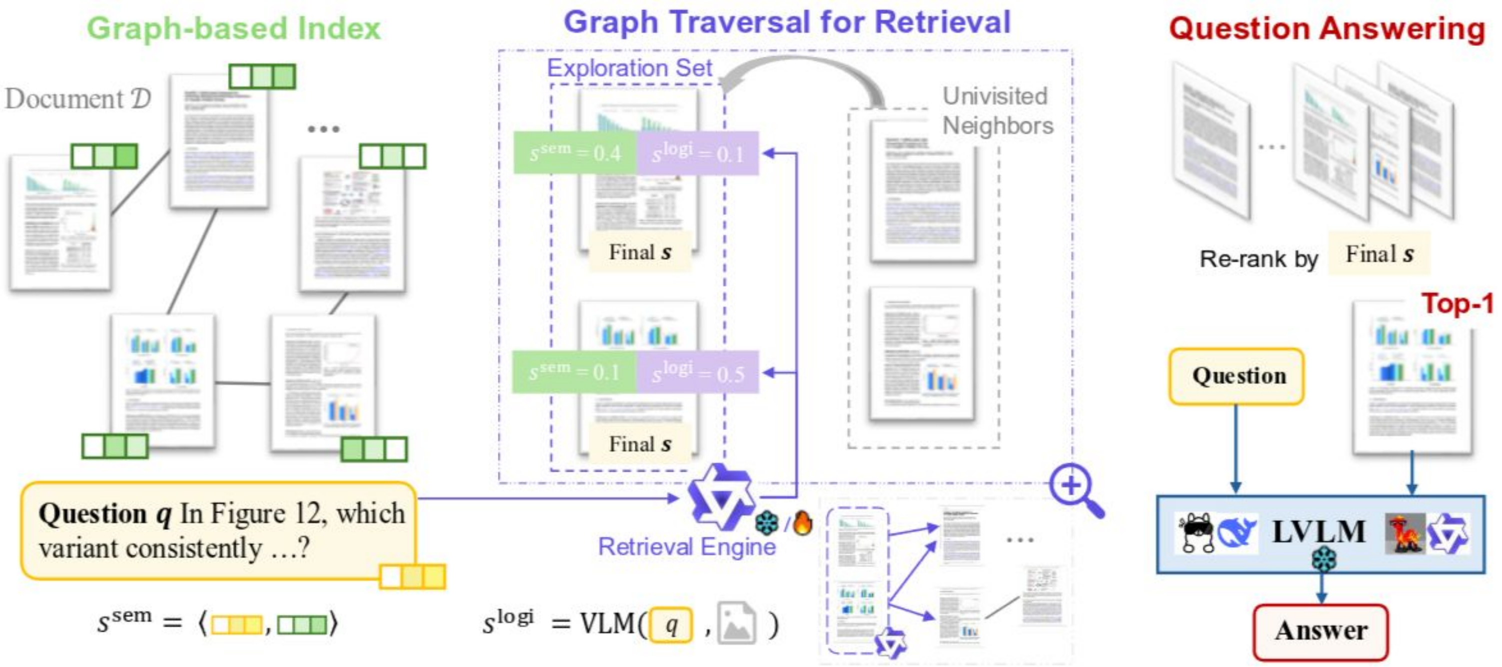
Ground-truth Evidence Page



Answer 49 days + 116 days = 165 days

MoLoRAG

BootStrapping Document Understanding via Multi-modal Logic-aware Retrieval



MoLoRAG

BootStrapping Document Understanding via Multi-modal Logic-aware Retrieval

Graph-based Index

1. Page Graph Construction
   a. Node - page
   b. Edge - based on the similarity b/w pages

$$G(\mathcal{V}, \mathcal{E})$$

$$p_i \in \mathcal{V}$$

$$\mathcal{E} = \{(p_i, p_j) | \langle E_{p_i}, E_{p_j} \rangle \geq \theta\}$$

Graph Traversal for Retrieval

1. Initialization
   c. Semantic score based-selection →***Initial Exploration Set***
2. Relevance Scoring
   a. VLM assigns a Logical Relevance Score (page - query) to each page
   b. **Final Relevance Score = Logical relevance score + Semantic score**
3. Iterative Traversal
   a. Once completed, all visited nodes are **re-ranked** based on their final relevance score

MoLoRAG

Experiments - Overall Performance (top-3 retrieval)

Type	Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
LLM-based	Mistral-7B	Text RAG	24.47	25.06	11.45	41.14	25.53
	Qwen2.5-7B	Text RAG	25.52	27.93	12.72	40.06	26.56
	LLaMA3.1-8B	Text RAG	22.56	29.80	13.49	45.96	27.95
	GPT-4o	Text RAG	27.23	32.74	14.25	50.20	31.11
	DeepSeek-V3	Text RAG	**29.82**	**34.73**	**17.05**	**52.36**	**33.49**
LVLM-based	LLaVA-Next-7B	Direct	7.15	10.78	3.05	11.61	8.15
		M3DocRAG	**10.10**	**13.85**	5.34	**13.98**	**10.82**
		MoLoRAG	9.37	13.49	4.83	13.78	10.37
		MoLoRAG+	9.47	13.58	**5.60**	13.48	10.53
	DeepSeek-VL-16B	Direct	8.40	14.72	6.11	16.14	11.34
		M3DocRAG	18.12	29.60	7.89	27.07	20.67
		MoLoRAG	20.43	29.98	9.67	38.98	24.77
		MoLoRAG+	**25.47**	**37.21**	**10.94**	**41.54**	**28.79**
	Qwen2.5-VL-3B	Direct	26.65	24.89	25.19	51.57	32.08
		M3DocRAG	29.11	44.40	24.68	53.25	37.86
		MoLoRAG	32.11	**45.79**	24.43	57.68	40.00
		MoLoRAG+	**32.47**	45.27	**27.23**	**58.76**	**40.93**
	Qwen2.5-VL-7B	Direct	32.77	26.38	29.77	64.07	38.25
		M3DocRAG	36.18	49.03	28.50	63.78	44.37
		MoLoRAG	39.28	51.71	**32.32**	69.09	48.10
		MoLoRAG+	**41.01**	**51.85**	31.04	**69.19**	**48.27**
Multi-agent	MDocAgent (LLaMA3.1-8B+Qwen2.5-VL-7B)		38.53	46.91	30.03	66.34	45.45

1. **LLMs struggle with document understanding compared to LVLM-based methods**
2. **MoLoRAG consistently boosts LVLM performance**

MoLoRAG

Experiments - Retrieval Performance

Top-K	Method	MMLongBench				LongDocURL			
		Recall	Precision	NDCG	MRR	Recall	Precision	NDCG	MRR
1	M3DocRAG	43.31	56.67	56.67	56.67	46.84	64.66	64.66	64.66
	MDocAgent (Text)	29.30	38.99	38.99	38.99	42.03	58.37	58.37	58.37
	MDocAgent (Image)	43.79	57.49	57.49	57.49	46.80	64.57	64.57	64.57
	MoLoRAG	45.46	59.95	59.95	59.95	48.98	67.71	67.71	67.71
	MoLoRAG+	**51.32**	**66.86**	**66.86**	**66.86**	**50.82**	**70.08**	**70.08**	**70.08**
3	M3DocRAG	64.17	31.62	54.13	65.36	67.00	33.78	58.23	72.51
	MDocAgent (Text)	43.21	20.77	37.13	45.26	58.53	29.33	54.12	65.28
	MDocAgent (Image)	64.74	31.97	54.75	66.12	66.67	33.62	58.26	72.47
	MoLoRAG	67.22	40.81	57.34	68.56	**70.04**	36.41	61.56	75.78
	MoLoRAG+	**68.87**	**48.67**	**64.49**	**73.50**	68.92	**47.53**	**64.90**	**77.14**
5	M3DocRAG	72.00	22.58	54.06	66.92	74.32	23.34	58.05	73.83
	MDocAgent (Text)	50.60	15.48	37.19	46.98	65.41	20.41	53.97	66.55
	MDocAgent (Image)	71.45	22.37	54.58	67.53	74.60	23.50	58.06	73.90
	MoLoRAG	**74.13**	35.83	57.29	69.63	**77.14**	26.13	61.30	76.88
	MoLoRAG+	72.37	**45.34**	**64.36**	**73.97**	73.69	**42.47**	**64.74**	**77.89**

MoLoRAG

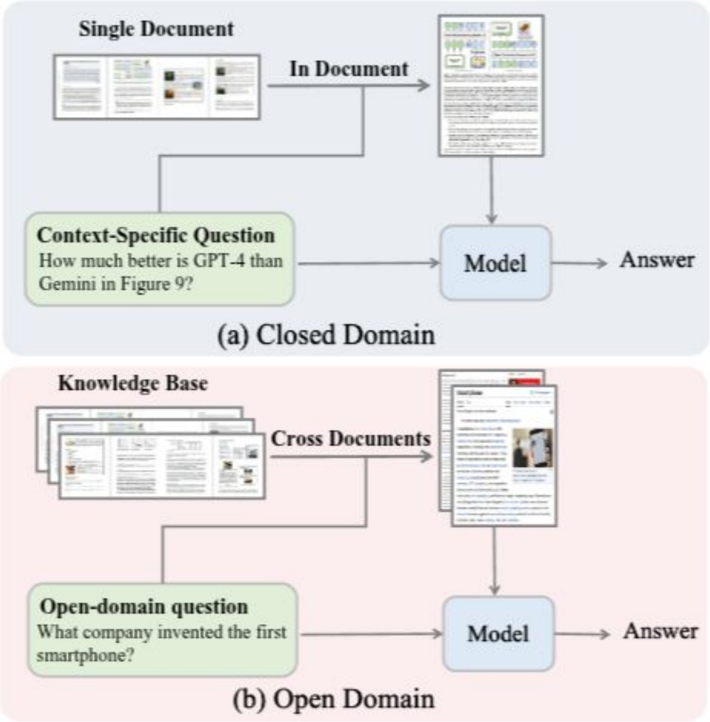
Conclusion & Limitations

Conclusion

1. Overcame the reliance solely on semantic relevance for retrieval
   ⇒ Incorporating Logical Relevance via Page Graph
2. Multi-hop Reasoning over page graph

Limitation

1. Primarily focused on closed-domain document understanding
2. Extension to **Open-Domain** setting remains as a challenge



Thank you
Q&A