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

香港中文大學 The Chinese University of Hong Kong
arXiv GitHub
30th ANNIVERSARY EMNLP 2025 Suzhou, China November 4-9 中国苏州 11月4日-9日

Motivation

Document Question Answering (DocQA) Answer a question based on the content of a document

- Interpreting medical reports
- Assisting with academic literature
- Supporting financial decision-making

Existing LLM-based Methods First convert the document into text using OCR, and then retrieve relevant paragraphs from text to feed into LLM

× *Inevitable **multi-modal information loss** like tables, figures, document layouts, etc*

Existing LVLM-based Methods

- **Direct:** Directly feeding all image snapshots of the document to an LVLM for question answering × *Exceed **LVLM context***
- **Retrieval-based:** Use a document encoder to encode pages and retrieve relevant ones based on vector similarity × *Only **semantic relevance***

Precise question answering requires pages that are **logically relevant** to the query, e.g., providing clues for the derivation of the answer

Methodology

Graph-based Index

Construct a page graph to represent the dependencies between pages

$$E_{p_i} = \mathrm{DocEncoder}(p_i)$$

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

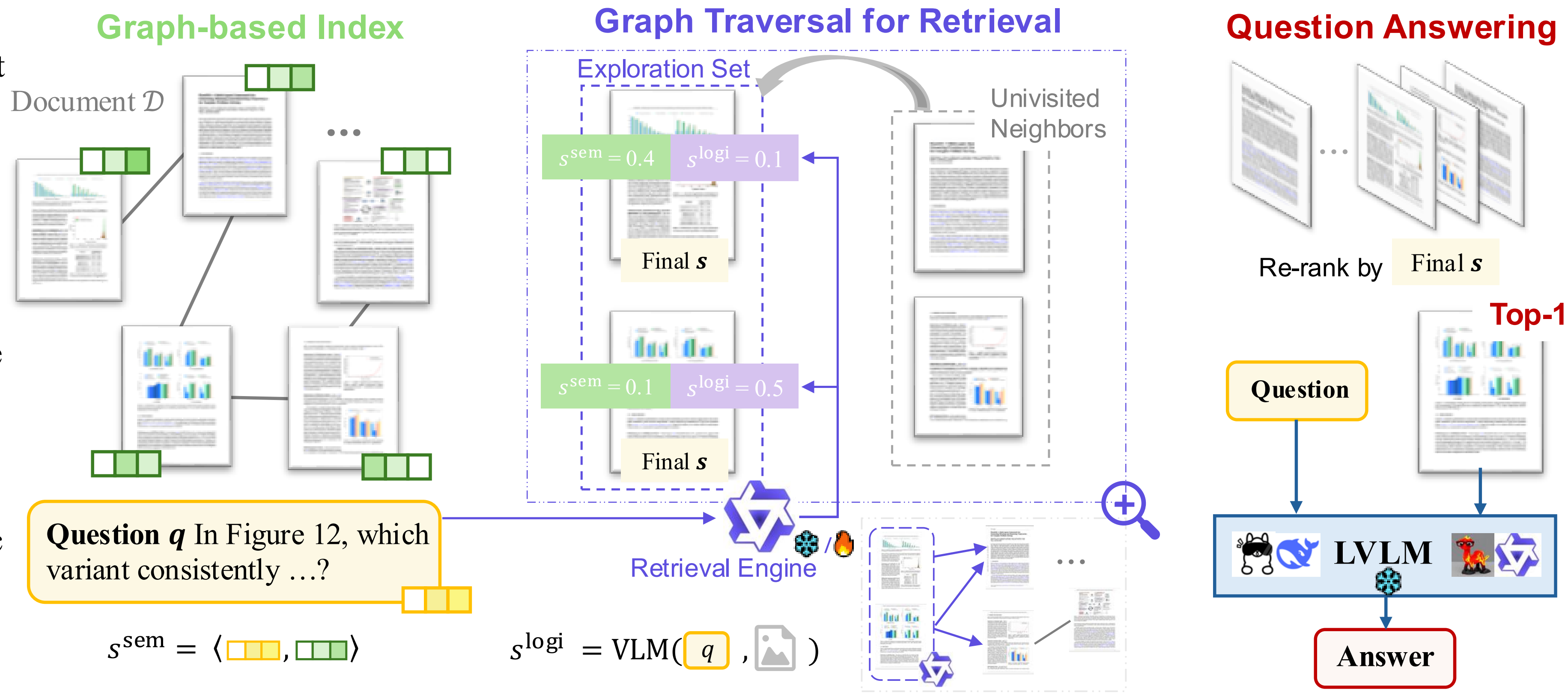
Graph Traversal for Retrieval

Leverage a VLM to serve as the retrieval engine, reasoning over the graph through traversal to identify logically relevant pages

Question Answering

Combine both logical and semantic relevance into a unified similarity score to re-rank pages

$$s_i = \mathrm{Combine}(s_i^{\mathrm{sem}}, s_i^{\mathrm{logi}})$$

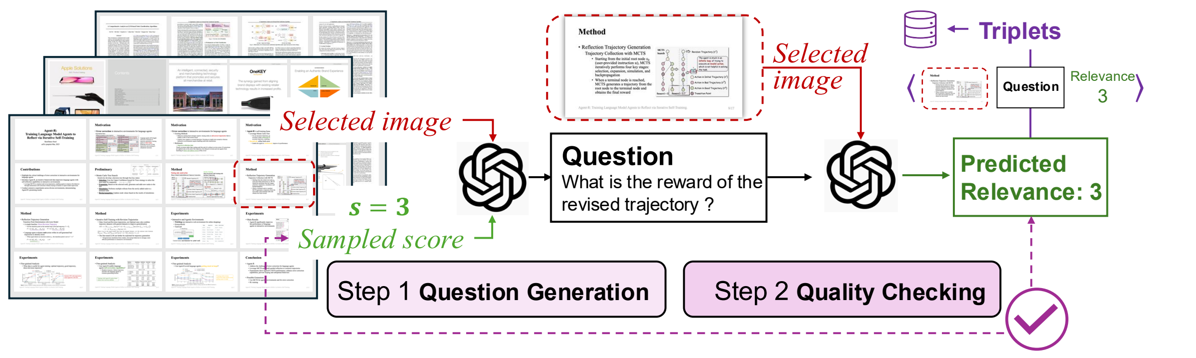


✓ Compatibility with arbitrary LVLMs ✓ Enhanced retrieval accuracy ✓ Efficiency of controlled graph traversal

MoLoRAG+: Fine-tuned Retrieval Engine

Replace the pre-trained VLM retrieval engine with a fine-tuned version, acquiring the specialized logical relevance score checking capability via SFT using curated <Question, Image, Relevance Score> triplets

Backbone: Qwen2.5-VL-3B



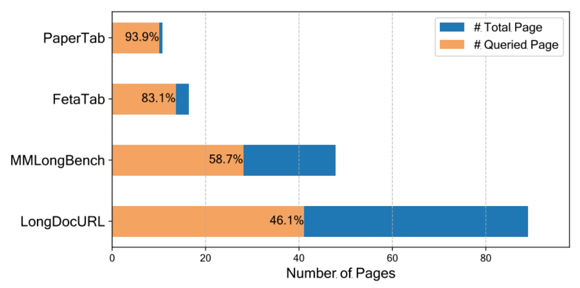
Experiments

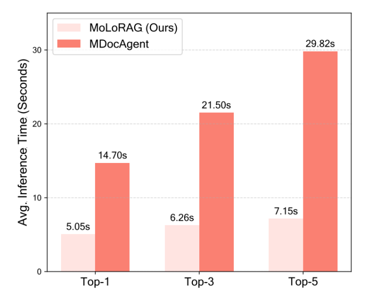
Type	Model	Method	MMLongBench	LongDocURL	PaperTab	FetaTab	Avg.
LLM-based	Mistral-7B	Text RAG	24.47	25.06	11.45	41.14	25.53
	Qwen2.5-7B	Text RAG	25.52	27.93	12.72	40.06	26.56
	LLaMA3.1-8B	Text RAG	22.56	29.80	13.49	45.96	27.95
	GPT-4o	Text RAG	27.23	32.74	14.25	50.20	31.11
	DeepSeek-V3	Text RAG	**29.82**	**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

DocQA Performance
MoLoRAG consistently boosts diverse LVLM's performance

Top-K	Method	MMLongBench Recall	MMLongBench Precision	MMLongBench NDCG	MMLongBench MRR	LongDocURL Recall	LongDocURL Precision	LongDocURL NDCG	LongDocURL 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**

Retrieval Accuracy MoLoRAG identifies relevant pages





Retrieval scalability and inference efficiency of MoLoRAG