LILaC: Late Interacting in Layered Component Graph for Open-domain Multimodal Multihop Retrieval

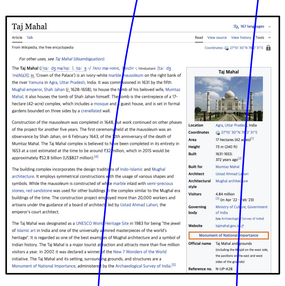
Joohyung Yun (POSTECH)
Doyup Lee (DirectorLabs)
Wook-Shin Han* (POSTECH)

My Homepage!

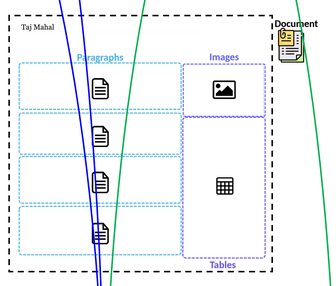
Problem Description

Multimodal Documents

- Real-world documents are often multimodal.
  - *e.g.*, webpages, PDFs, presentation slides, and way more!
- Accurate question-answering involves correctly resolving the relationships between the components.
  - *Component:* a multimodal element (i.e., text/table/image)



An example webpage

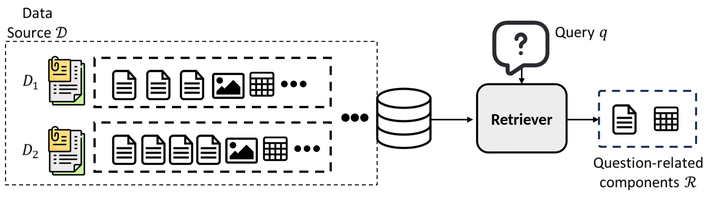


How we model a multimodal document

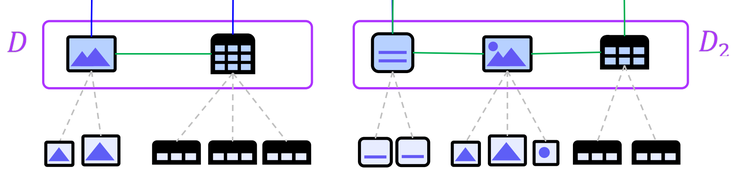
Multimodal Document Retrieval

Given a data source $\mathcal{D}$, a link mapping $\mathcal{L}$ and a query q, correctly retrieve a ranked list of components $\mathcal{R} = [C_1, \ldots, C_{n_{ret}}]$. $\mathcal{R}$ should contain the ground truth set of components $\{C_{gt_1}, \ldots, C_{gt_r}\}$.

- Data source $\mathcal{D} = \{D_1, \ldots, D_{k_{doc}}\}$
- Document $D = [C_1, \ldots, C_{k_{comp}}]$
- Component $C \in \{P, T, I\}$
  - P: passage, T: table, I: image
- Link mapping $\mathcal{L} = \mathcal{C} \rightarrow \mathcal{D}$
  - $\mathcal{C}$: A set of all components within $\mathcal{D}$

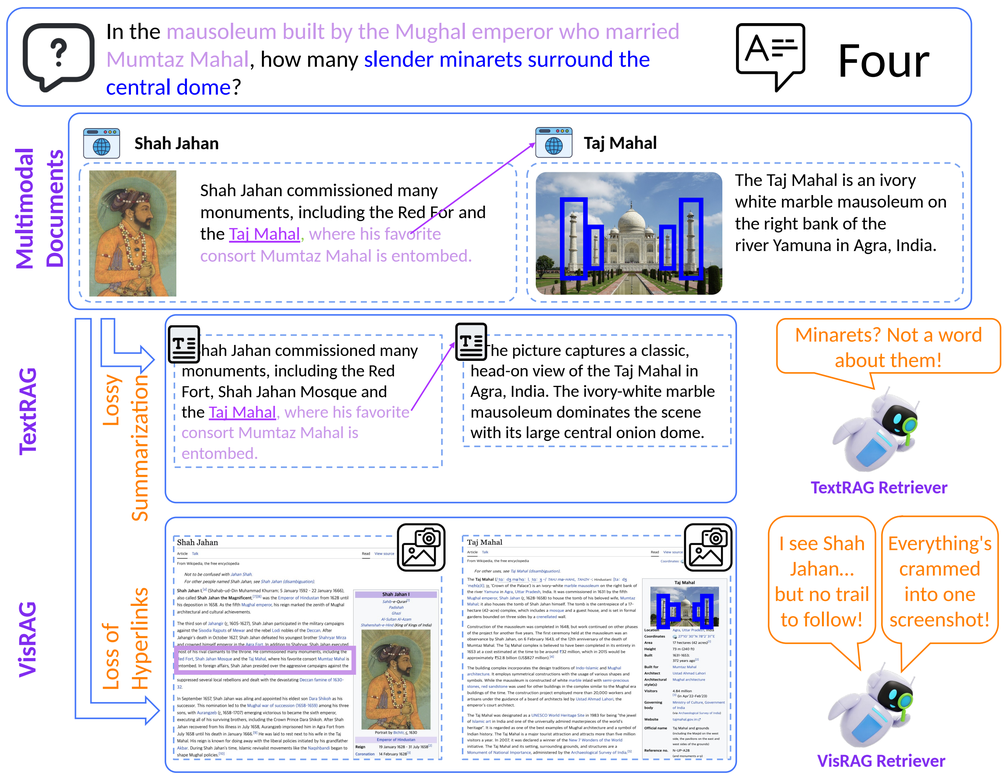


Keys to Accurate Multimodal Document Retrieval

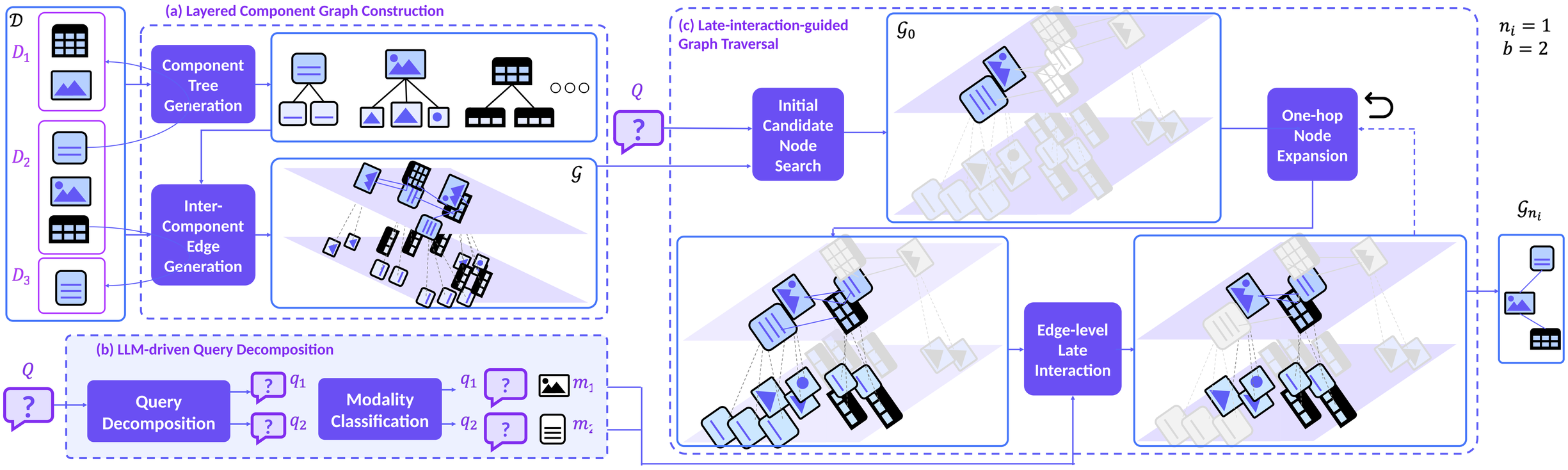


- **Modality preservation**
- **Consideration of fine-grained granularity**: Paragraph -> sentences, Table -> rows, Image -> detected objects
- **Multimodal multihop reasoning**: Exploit both *inter-document links* and *intra-document relationships*

Limitations of Previous Works



The LILaC Algorithm



Brief Explanation about LILaC

Layered Component Graph (LCG)

- Builds a two-layer graph of each document: coarse nodes (paragraphs, tables, images) and fine nodes (sentences, rows, detected images).
- Adds edges for (i) hierarchical containment (coarse → fine) and (ii) navigational links (within a document clique + cross-document hyperlinks) to support multihop reasoning.

LLM-driven Query Decomposition

- Splits the query into modality-specific subqueries (text/table/image) and embeds each.

Late-Interaction Subgraph Retrieval

- **Seed**: conduct a kNN search to find top-b query-relevant components.
- **Traverse**: iteratively perfom beam search
  1. Perform one-hop expansions from an existing subgraph.
  2. Leave the top-b query-relevant edges.
     - Late-interaction edge score calculation: $s(e; Q_{sub}) = \sum_{q \in Q_{sub}} \max_{c \in e_c} sim(c, q)$

Edge-level Late Interaction

Evaluation

Retrieval Accuracy

Algorithm	Embedder Type	MP-DocVQA		SlideVQA		InfoVQA		MultimodalQA		MMCoQA	
		R@3	MRR@10	R@3	MRR@10	R@3	MRR@10	R@3	MRR@10	R@3	MRR@10
NV-Embed-v2	Text	67.85	61.91	88.49	79.55	86.21	80.86	60.19	67.86	46.16	41.45
VisRAG-Ret	Image	83.25	75.55	91.55	84.30	92.76	86.22	50.08	55.08	27.63	23.75
ColPali	Image	80.71	74.86	89.39	81.55	88.30	82.76	58.73	65.05	36.24	32.33
LILaC (w/ mmE5)	Multimodal	61.25	55.30	77.52	68.80	75.09	69.86	54.79	59.02	48.88	40.30
LILaC (w/ UniME)	Multimodal	77.83	71.42	84.35	77.93	82.28	75.27	58.52	61.44	49.63	42.97
LILaC (w/ MM-Embed)	Multimodal	**83.59**	**78.75**	**92.81**	**84.43**	**93.17**	**86.83**	**69.07**	**75.28**	**55.80**	**50.77**

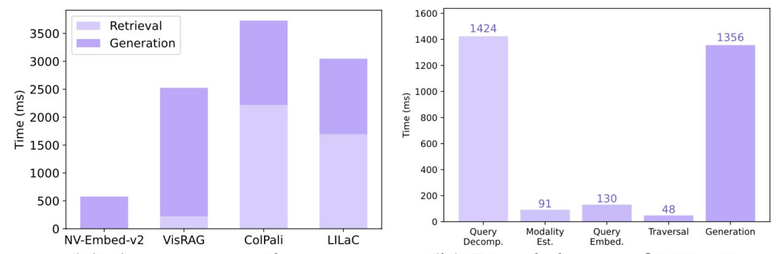
Question Answering Accuracy

Algorithm	MLLM	MP-DocVQA		SlideVQA		InfoVQA		MultimodalQA		MMCoQA	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
NV-Embed-v2	Qwen2.5-VL 7B	56.51	63.16	53.77	64.41	60.72	63.40	37.23	43.85	28.05	34.67
VisRAG-Ret	MiniCPM V2.6	54.31	68.86	43.88	62.37	50.83	57.55	28.18	34.01	21.51	27.87
VisRAG-Ret	Qwen2.5-VL 7B	65.34	72.24	55.03	66.13	60.16	61.93	22.24	25.55	16.69	20.90
ColPali	Qwen2.5-VL 7B	64.46	71.16	53.77	64.54	58.07	60.38	23.59	27.37	18.07	22.30
LILaC (w/ mmE5)	Qwen2.5-VL 7B	52.96	59.53	50.89	59.07	50.12	53.12	40.72	47.46	33.90	40.38
LILaC (w/ UniME)	Qwen2.5-VL 7B	62.43	69.40	53.05	62.89	53.39	56.86	43.42	49.72	33.39	40.12
LILaC (w/ MM-Embed)	Qwen2.5-VL 7B	**65.48**	**72.42**	**55.57**	**66.32**	**60.91**	**62.87**	**44.57**	**51.97**	**36.31**	**43.22**

Ablation Study

Embedder Model	Variant	MP-DocVQA		SlideVQA		InfoVQA		MultimodalQA		MMCoQA	
		R@3	MRR@10	R@3	MRR@10	R@3	MRR@10	R@3	MRR@10	R@3	MRR@10
mmE5	LILaC (w/o LCG & QD)	48.90	43.97	75.91	68.13	65.60	58.55	42.99	46.92	41.22	34.51
	LILaC (w/o QD)	60.81	55.02	74.14	67.58	67.70	60.01	45.15	51.12	44.18	36.62
	LILaC	**61.25**	**55.35**	**76.80**	**68.99**	**68.91**	**61.18**	**54.78**	**59.32**	**48.54**	**40.22**
UniME	LILaC (w/o LCG & QD)	52.12	45.31	81.47	71.22	83.57	77.07	47.68	49.06	45.78	38.41
	LILaC (w/o QD)	77.83	71.27	83.45	75.70	85.11	78.01	52.18	54.01	47.11	39.85
	LILaC	**77.83**	**71.39**	**84.35**	**77.93**	**85.53**	**78.81**	**58.43**	**61.32**	**49.45**	**42.91**
MM-Embed	LILaC (w/o LCG & QD)	75.80	69.09	92.80	82.19	90.39	83.71	61.10	67.35	47.94	43.75
	LILaC (w/o QD)	82.23	77.75	92.27	83.20	92.17	85.53	63.19	69.91	50.18	45.59
	LILaC	**83.59**	**78.75**	**92.81**	**84.43**	**93.17**	**86.83**	**69.07**	**75.28**	**55.80**	**50.77**

Runtime Analysis



(a) Average runtime (b) Breakdown of LILaC