Name : Yiming Xu

Computer Vision Center

UAB Universitat Autònoma de Barcelona

Agentic Multimodal Document Understanding with Large Language Models

Motivation

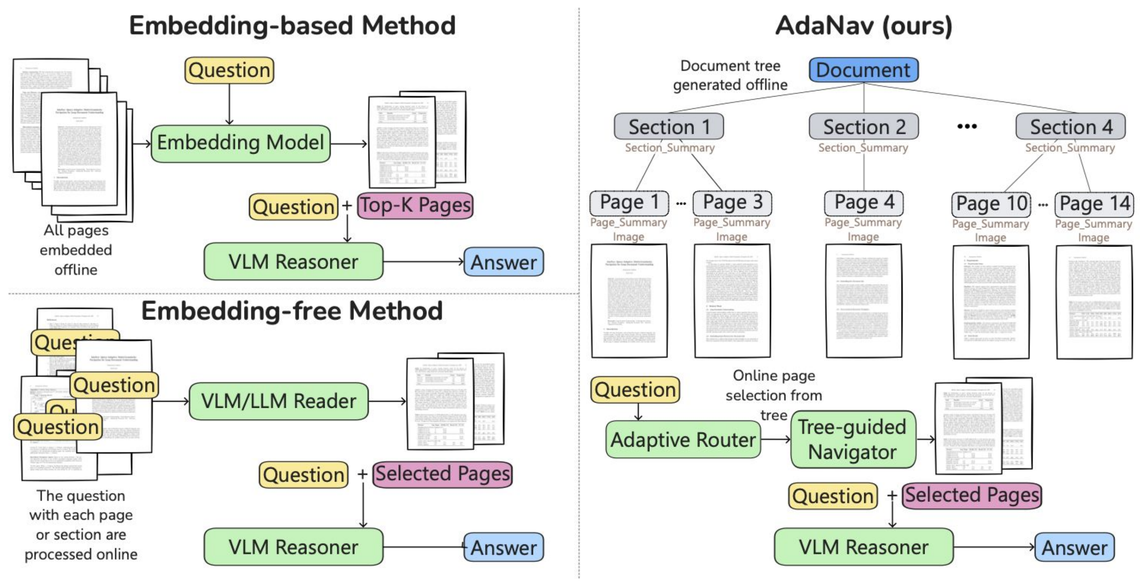
Long-document VQA: evidence is scattered, context is limited.

Two paradigms, two failure modes:

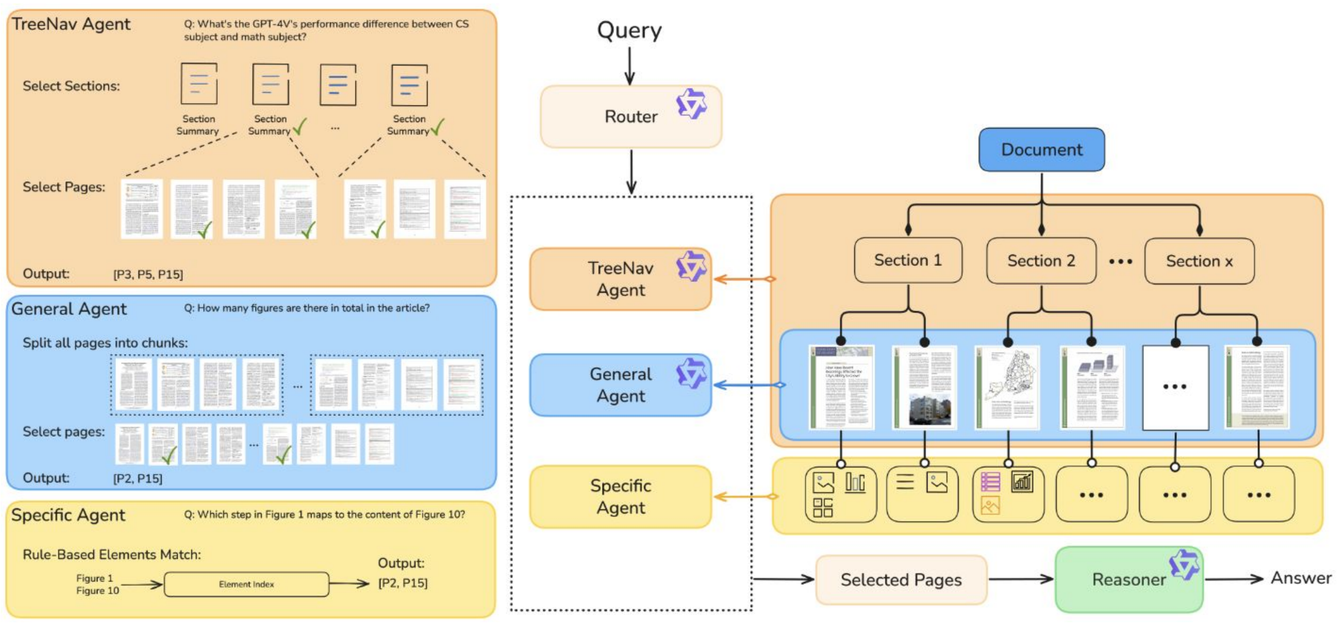
- **Embedding retrieval**: scalable, but misses compositional (multi-hop, cross-page) queries.
- **Embedding-free reading**: no retrieval bias, but prohibitive token cost.

Insight: documents are hierarchical (Section → Page → Element). **Navigate coarse-to-fine**, filter branches early.

AdaNav: query-adaptive tree navigation. No embedding model, fewer pages, higher accuracy.



Architecture



Offline: MinerU builds a **4-level tree** (Document, Section, Page, Element). Pages keep screenshots + structured summaries; tables, figures, charts are cropped and indexed.

Online: Route → Navigate → Reason.

- **Router**: rule-based first, LLM fallback.
- **Specific**: direct index lookup, zero LLM cost (10.5% of queries).
- **General**: chunked full-document scan (12.8%).
- **TreeNav**: coarse-to-fine section-to-page navigation (76.7%).

Reasoner sees only the raw retrieved pages.

Results

Method	Pg. Ret.	MMLB	PaperTab	FetaTab	Avg.
Oracle (GT pages)	–	67.65	–	–	–
M3DocRAG	6	41.80	60.10	79.80	60.57
MDocAgent	12	55.30	**64.90**	**84.50**	68.23
SimpleDoc	3.2	59.55	64.38	80.31	68.08
AdaNav (Qwen2.5-VL)	4.17	59.46	61.06	82.89	67.80
AdaNav (Qwen3-VL)	4.17	**65.73**	63.61	83.75	**71.02**

Method	Table	Figure	Chart	Multi-page	ACC
Oracle (GT pages)	65.7	63.6	66.3	59.9	67.7
M3DocRAG	39.2	43.8	45.4	30.6	41.8
MDocAgent	56.1	52.8	54.7	42.9	55.3
SimpleDoc	51.2	51.2	54.9	39.2	59.6
AdaNav (Base)	59.4	47.3	57.1	43.5	61.4
AdaNav (Agentic)	**68.7**	**57.7**	**60.9**	**54.9**	**65.7**

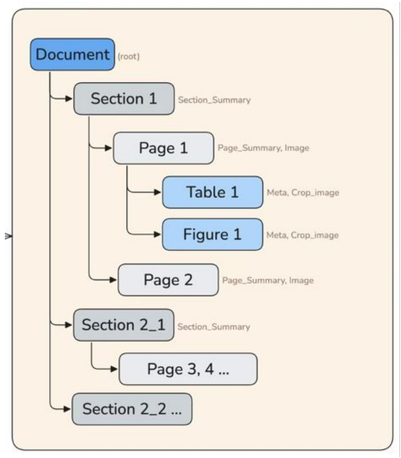
Accuracy (%) on MMLongBench-Doc / PaperTab / FetaTab (top) and per-category on MMLongBench-Doc (bottom).

65.7%
MMLB accuracy, +6.2 vs best baseline

4.17
Pages retrieved per query

18.3×
Smaller index than embedding RAG, 78 MB vs 1.4 GB

DocTree Construction



- **Document**: tree root, one per PDF
- **Section**: queryable index cards (topics, entities, page provenance)
- **Page**: full screenshot + structured summary
- **Element**: cropped tables, figures, charts with captions

Future Work

RL-trained retrieval routing:

Learn when to invoke text-based vs. visual retrieval, and when to stop, via RL for corpus-scale document QA.

Memory-augmented agents:

Explore cross-question episodic memory for long-horizon document QA.

Acknowledgements:
Eric López, Artemis Llabrés, Maximiliano Hormazábal, Dimosthenis Karatzas & Ernest Valveny