M3DocDep

Dependency-first chunking

M3DocDep: Multi-modal, Multi-page, Multi-document Dependency Chunking with Large Vision-Language Models

Joongmin Shin[1], **Jeongbae Park**[1], **Jaehyung Seo**[2]‡, **Heuiseok Lim**[1,3]‡

[1]*Human-inspired AI Research, Korea University* [2]*CSE, Konkuk University* [3]*CSE, Korea University* ‡*Corresponding authors*

CVPR JUNE 3-7, 2026 DENVER COLORADO

Problem

Industrial corpus RAG fails when chunks lose their parent section, cross-page continuation, or figure–caption binding. Existing parsers ignore visual cues or produce fragile token sequences that miss document structure.

Our idea

Recover a global document dependency tree first, then build chunks from section subtrees — preserving page spans and figure/caption bindings.

Contributions

1. **Multi-modal dependency scoring.** Biaffine head over LVLM SoftROI block embeddings.

2. **Global MST tree decoding.** Single-root, acyclic — with cross-page parent–child links.

3. **Tree-guided chunking.** Section-subtree DFS + figure–caption binding.

Takeaways

1. Industrial corpus RAG is a dependency-recovery problem.

2. Hierarchy + retrieval + QA all improve under shared blocks.

3. DP, OCR, LVLM, retrievers swap without changing chunks.

Paper
arxiv.org/abs/...

Code
github.com/...

Contact
tlswndals13@korea.ac.kr

Method

(a) SharedDet (DP + OCR) → (b) LVLM multi-modal block embedding → (c) biaffine + MST tree decoder → (d) structure-aware chunking.

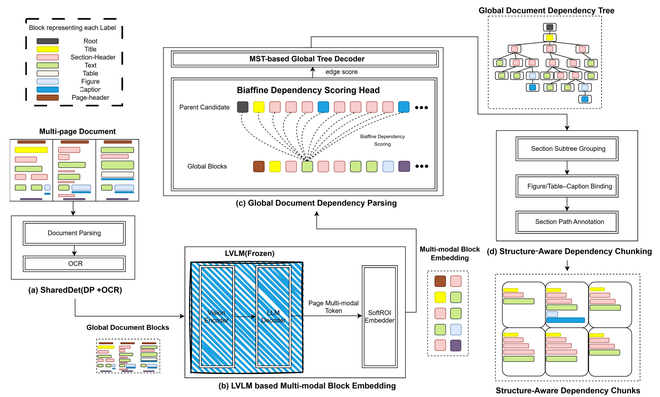
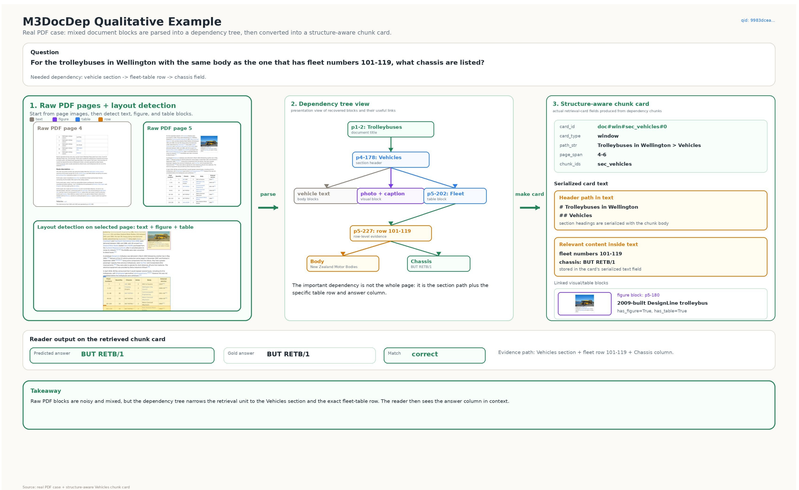


Figure 1. M3DocDep pipeline overview.



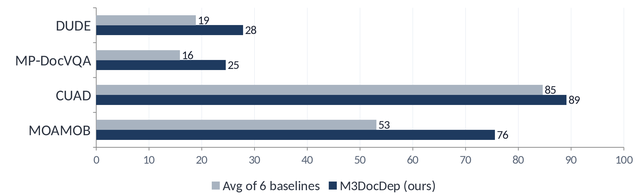
Results

+10.6%
avg retrieval nDCG gain
range +1.1 to +15.3% across 4 corpora

+9.8%
avg QA ANLS gain
range +4.5 to +15.3% across 4 corpora

Relative improvement over the strongest chunking baseline (MultiDocFusion) on DUDE · MP-DocVQA · CUAD · MOAMOB.

Retrieval nDCG — ours vs. baseline avg



Downstream QA — ANLS (%)

Dataset	Ours	Δ vs. MultiDocFusion
DUDE	**21.4**	+2.84
MP-DocVQA	**18.2**	+2.02
CUAD	**29.3**	+1.87
MOAMOB	**27.1**	+1.18

Document Hierarchical Parsing (DHP) — F1 / STEDS (%)

	HRDS	HRDH	DocHieNet
Qwen2.5-VL (LVLM only)	*28.4 / 14.5*	*27.6 / 20.0*	*18.2 / 9.6*
DSPS	65.3 / 59.6	54.1 / 38.4	35.6 / 23.8
DSHP-LLM	44.9 / 29.5	61.3 / 51.3	64.3 / 53.5
M3DocDep (ours)	**82.9 / 76.5**	**77.8 / 71.7**	**76.0 / 70.8**

Ablation — STEDS drop

No cross-page edges	**−9.26**
MST → local argmax	**−6.70**

Robustness — top-1 under all component swaps

3 DP backbones (DETR · DIT · VGT) · **3 OCR** engines (Tesseract · EasyOCR · TrOCR) · **4 embedders** (BGE · E5 · BM25 · MM-Embed)