Long Document Understanding · 논문 리뷰

DocSeeker

Structured Visual Reasoning with Evidence Grounding for Long Document Understanding

Analysis · Localization · Reasoning — evidence grounding으로 long-document 이해를 다시 설계한다

REVIEWED BY
Paper Review Seminar

AUTHORS
Hao Yan · Yuliang Liu · Xiang Bai et al. (HUST · Huawei)

DATE
June 2026 · Seoul

DocSeeker: Structured Visual Reasoning with Evidence Grounding for Long Document Understanding

Hao Yan[1], Yuliang Liu[1*], Xingchen Liu[1], Yuyi Zhang[1],Minghui Liao[2], Jihao Wu[2], Wei Chen[1*], Xiang Bai[1]
[1]Huazhong University of Science and Technology [2]Huawei Inc.
{haoyan, ylliu, lemuria_chen, xbai}@hust.edu.cn
https://github.com/yh-hust/DocSeeker

Paper at a Glance

DocSeeker는 long document VQA를 **Analysis · Localization · Reasoning**의 구조화된 workflow로 재정의한다.

BIBLIOGRAPHIC

DocSeeker: Structured Visual Reasoning with Evidence Grounding for LongDoc

AUTHORS · Hao Yan, Yuliang Liu*, Xingchen Liu, Yuyi Zhang, Minghui Liao, Jihao Wu, Wei Chen*, Xiang Bai
AFFIL. · Huazhong Univ. of Science & Technology · Huawei Inc.
CODE · github.com/yh-hust/DocSeeker

THREE CORE CONTRIBUTIONS

01 **ALR Paradigm** · Analysis · Localization · Reasoning
find-then-reason을 구조화된 CoT 포맷으로 명시한다

02 **Two-stage Training** · SFT + EviGRPO
Evidence Localization을 reward로 직접 학습시킨다

03 **EGRA** · Evidence-Guided Resolution Allocation
evidence page에만 high-res를 할당해 메모리를 잘라낸다

HEADLINE RESULTS · 5 multi-page DocVQA benchmarks · Qwen2.5-VL-7B backbone

+30–60%
Baseline 대비 성능 향상
5개 DocVQA benchmark 평균

SOTA
Open-source 5개 benchmark 모두 1위
closed-source 모델과도 경쟁력 확보

7B
Qwen2.5-VL-7B backbone
추가 모듈 없이 학습 절차만으로 해결

Short → Long
짧은 문서로 학습한 모델이 ultra-long document까지 일반화
≤20p 학습 데이터로 80p+ 문서에서도 안정적

Pure-vision MLLM은 왜 긴 문서에서 무너지는가

저자는 한계를 두 가지로 정리한다 — **Low SNR**과 **sparse supervision**.

P1 LOW SNR

Signal-to-Noise Ratio가 너무 낮다

수십에서 수백 페이지에 이르는 문서에서 정답에 필요한 evidence는 **극소수 페이지**에만 존재한다. 나머지는 모두 noise로 작용하면서 모델의 token budget만 소모한다.

RAG도 근본 해결책이 아니다

classic **top-k dilemma** · k가 작으면 evidence를 놓치고, 크면 noise가 다시 들어온다.
→ 외부 retriever 하나로는 근본적으로 해소되지 않는다.

P2 SPARSE SUPERVISION

중간 추론을 감독할 신호가 없다

기존 multi-page DocVQA dataset은 **"long input → short answer"** 쌍만 제공하고, Evidence Localization이나 정보 종합 단계에 해당하는 label은 없다.

그 결과

모델은 정답을 **Rote Memorization**으로 학습한다.
→ in-domain에서만 잘 동작하고 OOD와 ultra-long document에서 일반화가 무너진다.

⇒ *외부 retriever 의존을 넘어, **모델 자체의 Evidence Localization 능력**을 키워야 한다.*

III.

Analyze · Localize · Reason

find-then-reason 인지 과정을 그대로 visual reasoning에 옮긴다

ALR Visual Reasoning Paradigm은 모델이 답을 출력하기 전 반드시 세 개의 명시적 단계를 거치도록 강제한다. 각 단계의 출력은 구조화된 토큰으로 남아 학습과 검증에 모두 활용된다.

THREE-STEP WORKFLOW

Step 1

Analysis

Question Analysis
무엇을 비교·계산·찾아야 하는가를 먼저 정리한다

Step 2

Localization

Evidence Localization
“…on page 9” 형태로 page ID를 명시적으로 인용한다

Step 3

Reasoning

Reasoning Process
인용된 증거 위에서만 추론 → answer + evidence_pages 출력

이 패러다임이 제공하는 이점

01 High Interpretability
인용된 page ID로 사용자가 답을 직접 검증할 수 있다

02 Token Separation
페이지 단위로 visual token을 구분해 학습한다

03 RAG-friendly
비연속 Top-k 페이지에서도 추론이 가능하다

DocSeeker의 end-to-end 구조

backbone은 **Qwen2.5-VL-7B-Instruct**를 그대로 쓴다. 추가 모듈 없이 학습 절차만으로 ALR 능력을 주입한다.

PHASE 1 · INPUT

Page-Aware Input

문서의 모든 페이지를 image로 바꾸고,

각 페이지 앞에 **"Page {ID}:"**marker를 붙인다.

→

PHASE 2 · BACKBONE

Qwen2.5-VL 7B Backbone

ViT + VL-Adapter + LLM decoder

구조는 원본 그대로 두고, **학습 절차**만으로 능력을 올린다.

→

PHASE 3 · ALR CHAIN

Structured Reasoning

<think> Analysis · Localization · Reasoning </think>

중간 reasoning chain을 명시적 token으로 남긴다.

→

PHASE 4 · OUTPUT

Final Answer Y_{ans}

<answer> {
"evidence_pages": [...]
"answer": [...]
} </answer>

TRAINING PROCEDURE — ALR 능력을 두 단계로 주입한다

Stage I · SFT

Gemini-2.5-Flash로 distillation한 **ALR CoT** 데이터로 backbone을 fine-tuning — ALR pattern 자체를 학습시킨다.

Stage II · EviGRPO

Format · Localization · Answer 세 가지 reward로 RL을 적용해 **정답과 Evidence Localization을 동시에 최적화한다.**

+ EGRA (Evidence-Guided Resolution Allocation) — 두 단계 모두에서 적용되어 long document 학습을 가능하게 만든다.

페이지를 explicit grounding unit으로 다룬다

각 페이지 이미지 앞에 **"Page {ID}:"** 마커를 삽입해 페이지를 ALR의 기본 단위로 노출한다.

INPUT SEQUENCE — 단순화된 표현

```
Page 1:
Page 2: ⟨img tokens⟩
        ⟨img tokens⟩
Page 3: ⟨img tokens⟩
          ⋮
        ⟨img tokens⟩
Page N: Question:"Comparing telecom operators ... ?"
```

⇒ 모델 출력에서 **"...on page 9"** 형태의 명시적 page citation이 가능해진다.
Ablation (Tab. 3 · 동일 6.3k): w/ Page ID 33.8 → w/o 30.4 (–3.4pt).

WHY IT MATTERS

01. Anchoring

긴 visual token sequence 안에서 페이지 경계를 명확히 한다.

02. Citeability

출력 JSON의 `evidence_pages`를 직접 학습 가능한 신호로 활용한다.

03. RAG Compatibility

retriever가 비연속 페이지를 반환해도 ID 기반 추적이 가능하다.

04. Verifiability

사용자가 답의 근거를 직접 확인할 수 있어 실제 배포 환경에서 중요하다.

Two-Stage Training Framework

Stage I은 Data Distillation으로 ALR pattern을 올리고, Stage II는 EviGRPO로 Evidence Localization을 정래한다.

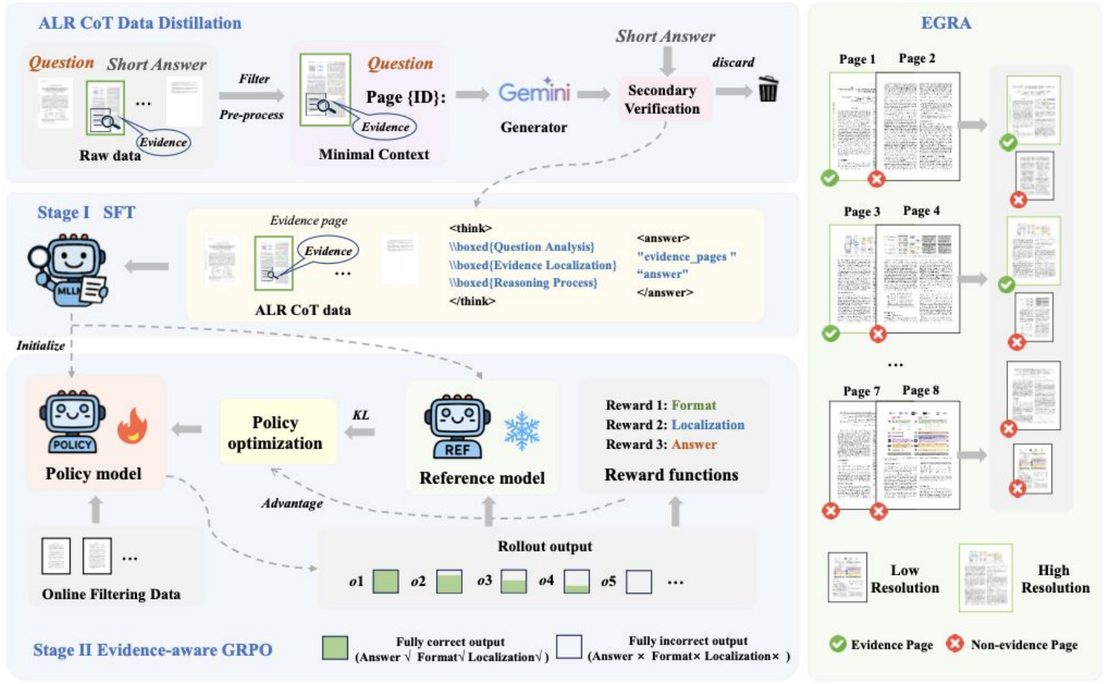


Fig. 2 (paper) · Stage I SFT (좌) — Stage II EviGRPO (중) — EGRA Resolution (우)

STAGE I · SFT

ALR CoT Distillation

teacher **Gemini-2.5-Flash**에 **Minimal Context**만 제공해 ALR CoT를 생성한다.
→ **Exact Match (EM) + GPT-4o Judge**로 Secondary Verification을 거친 13K 보관. *Role: ALR pattern을 주입한다.*

STAGE II · EVIGRPO

Evidence-aware GRPO

$R = \lambda_1 \cdot R_{format} + \lambda_2 \cdot R_{evidence} + \lambda_3 \cdot R_{answer}$(0.1, 0.3, 0.6)
Role: Evidence Localization 안정화. SFT만으로는 sub-optimal한 reasoning path를 RL이 다시 조정해 준다.

EVIGRPO REWARD — 세 reward의 결합

R_{format} ·Format
ALR template 준수 (binary)

$R_{evidence}$ ·Localization
evidence page Fβ (β>1, recall 우선)

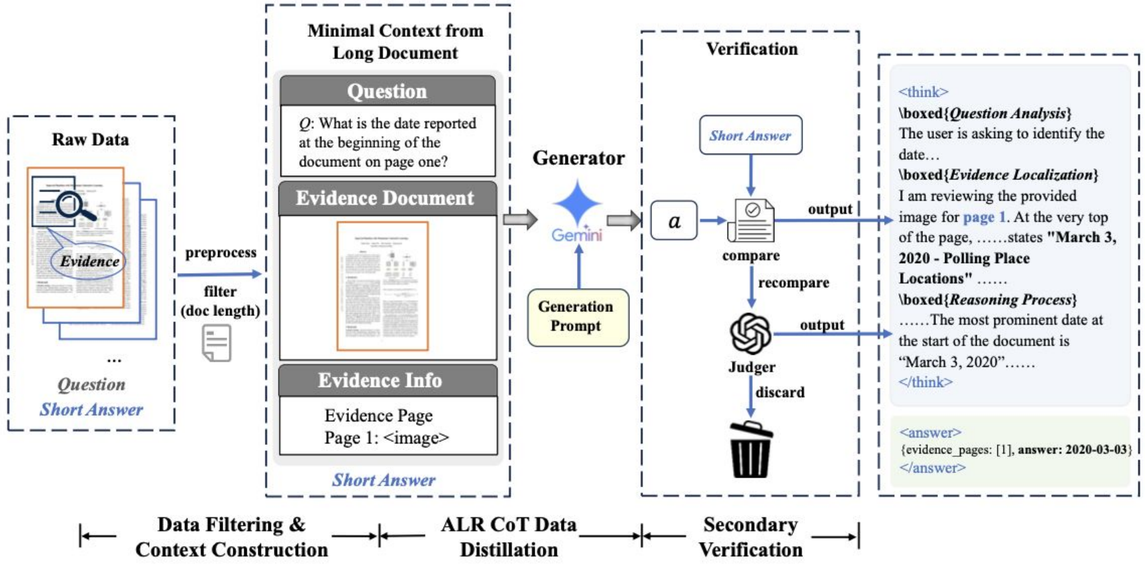
R_{answer} ·Answer
ANLS (Levenshtein 기반)

왜 "Evidence-aware"인가 · Tab. 5

SFT 38.6 → Vanilla GRPO (answer-only) 38.7 → **EviGRPO 40.1**
localization reward 없는 GRPO는 +0.1로 사실상 무효 — λ_2가 실질 +1.5를 만든다.

Data Distillation · Minimal Context & Verification

ALR CoT 데이터를 어떻게 생성하고 어떻게 검증하는지 — Appendix A의 두 가지 핵심 장치.



THREE PHASES — appendix A.1

01 Filtering & Context
짧은 문서 제거 후 Minimal Context 구성

02 ALR CoT Distillation
Gemini-2.5-Flash가 ALR 양식으로 생성

03 Secondary Verification
EM filter + GPT-4o Judge

MINIMAL CONTEXT

teacher에게는 evidence pages·소수 distractor·질문만 전달한다.

full-context는 noise·길이 때문에 teacher가 ALR 양식 생성에 실패한다. evidence만 주면 teacher가 집중하면서 success rate가 급등한다.

Distillation Success 20.4% →**67.3%**

SECONDARY VERIFICATION

Noise는 버리고, paraphrase는 살린다.

①Exact Match (EM) filter
답·evidence pages가 ground-truth와 일치하는지 자동 검사

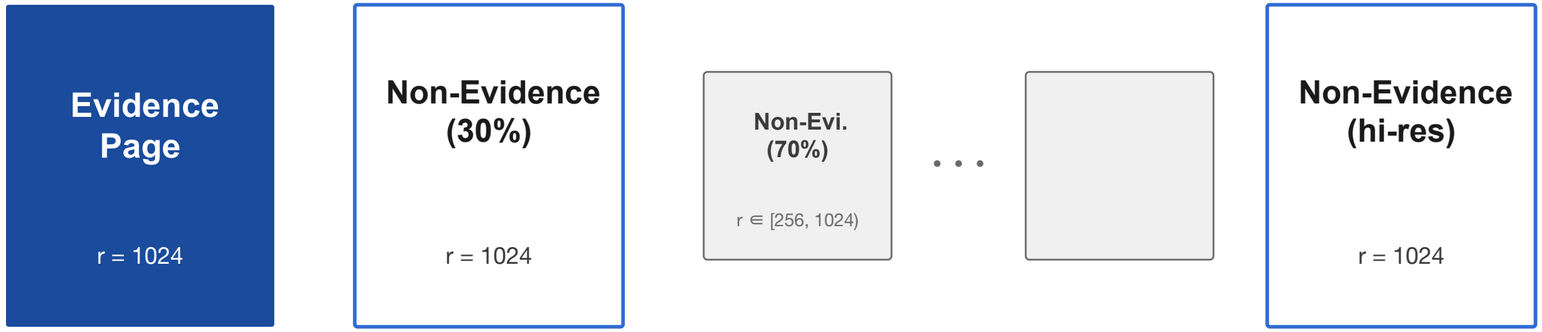
②GPT-4o semantic Judge
"17th February 1916" vs "1916-02-17" 같은 표기만 다른 정답을 회수

⇒ Minimal Context로 teacher의 입력 noise를 잘라낸 뒤, Secondary Verification으로 paraphrase까지 회수해 13K 학습 데이터를 확보한다.

EGRA · evidence page에만 고해상도를 할당한다

Evidence-Guided Resolution Allocation — GPU 메모리 한계 안에서 긴 문서 학습을 가능하게 하는 단순한 장치.

SCHEME — page-wise resolution allocation



MECHANISM

학습 시: non-evidence의 70%를 무작위 down-sample, 30%는 high-res 유지.
추론 시: 모든 페이지를 native high-res로 처리.

WHY THIS DESIGN

토큰 예산을 증거 페이지로 재분배해 학습 시 SNR을 끌어올린다. 동시에 일부 non-evidence를 high-res로 남겨 두면서 train-infer 간 해상도 차이를 줄여 distractor robustness까지 확보한다.

NET EFFECT

evidence fidelity는 유지하면서 불필요한 visual token만 줄인다.

EGRA의 정량적 ablation은 Results 슬라이드에서 함께 다룬다.

Experimental Setup & Benchmarks

multi-page DocVQA benchmark 5개 — in-domain 2개, OOD 3개로 일반화 성능까지 함께 측정한다.

SETUP

Experimental Setup

BACKBONE
Qwen2.5-VL-7B-Instruct

TEACHER (distillation)
Gemini-2.5-Flash

TRAIN DATA
ALR-CoT 13K (≤ 20 pages)

RL ALGORITHM
EviGRPO (rollout group size 16, online filter)

BENCHMARKS — 5 multi-page DocVQA datasets

DATASET	TYPE	METRIC	CHARACTER
DUDE	IN-DOMAIN	ANLS	general document QA
MP-DocVQA	IN-DOMAIN	ANLS	multi-page document QA
MMLongBench-doc	OOD	Accuracy	hundred-page scale, visual
LongDocURL	OOD	Accuracy	understanding + locating
SlideVQA	OOD	F1	slide-style multi-page

COMPARED WITH

Open-source: HiVT5 · CREAM · mPLUG-DocOwl2 · M3DocRAG · Vis-RAG · SV-RAG · VDocRAG · Docopilot · DocVLM · InternVL3
Closed-source: GPT-4V · GPT-4o · Gemini-1.5-Pro · Qwen-VL-Max

5개 benchmark에서 일관된 SOTA

Baseline과 동일 구조의 모델 대비 평균 +30 ~ +60% 성능 향상을 보인다. (Table 1)

TABLE 1 — Comparison on five document understanding benchmarks

METHOD	PARAM.	DUDE	MP-DocVQA	MMLong.	LongDoc.	SlideVQA
Open-Source MLLMs						
mPLUG-DocOwl2	8B	46.8	69.4	13.4	5.3	—
M3DocRAG	9B	—	84.4	21.0	35.1	55.7
Vis-RAG	4B	—	70.9	18.8	41.9	50.7
VDocRAG	8B	44.0	62.6	18.4	39.8	42.0
InternVL3	8B	47.4	80.8	24.1	38.7	54.4
Closed-Source Commercial						
Gemini-1.5-Pro	—	46.0	—	28.2	50.9	—
GPT-4o	—	54.1	67.4	42.8	64.5	—
Ours						
Baseline	7B	35.2	70.1	25.4	37.8	59.8
Baseline-SFT (short-answer)	7B	56.0	82.9	28.8	42.7	67.4
DocSeeker-SFT	7B	56.8	82.1	38.6	49.1	75.2
DocSeeker (full)	**7B**	**57.4**	**86.2**	**40.1**	**51.7**	**77.1**

*Best in **accent**. DUDE / MP-DocVQA = ANLS · MMLong. / LongDoc. = Accuracy · SlideVQA = F1.*

THREE TAKEAWAYS

01.

Open-source SOTA

InternVL3-8B 등 open-source 모델을 모두 능가한다. **MP-DocVQA 86.2**로 GPT-4o(67.4)도 추월한다.

02.

ALR > Short-answer SFT

short-answer SFT는 **Rote Memorization**으로 끝나 OOD 향상이 없다.
ALR CoT는 OOD에서도 일반화된다. (appendix Tab. A1)

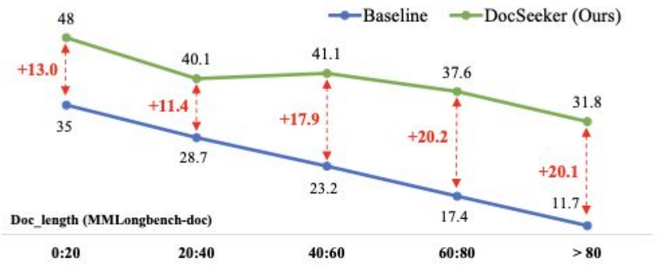
03.

EviGRPO refines SFT

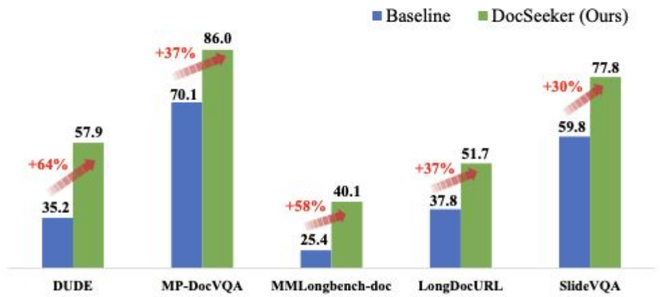
SFT만으로도 큰 향상을 얻지만, EviGRPO를 더하면 **모든 benchmark에서 추가 향상 (최대 +4pt)**이 나타난다.

Document Length에 강건하다

Baseline은 페이지 수가 늘수록 성능이 계단으로 무너지지만, DocSeeker는 전 구간에서 평탄한 곡선을 유지한다. (Fig. 1a · Fig. 3)



(a) Model performance across different document page ranges



(c) Model performance on 5 multi-page document VQA datasets

Fig. 1(a) · Acc by document page range — Baseline (gray) vs DocSeeker (blue)

FULL-DOC vs EVIDENCE-ONLY · Tab. 2

Evidence-only 41.1 (완벽 localization 상한) 기준으로, Baseline은 full-doc에서 −15.7pt 하락하지만 DocSeeker는 **−1.0pt**에 그친다 — **localization이 곧 성능**이라는 핵심 증거.

KEY NUMBERS — by document length (MMLongBench-doc)

PAGES	BASELINE	DOCSEEKER	Δ
0 – 20	35.0	**48.0**	**+13.0**
20 – 40	28.7	**40.1**	**+11.4**
40 – 60	23.2	**41.1**	**+17.9**
60 – 80	17.4	**37.6**	**+20.2**
> 80	11.7	**31.8**	**+20.1**

OBSERVATION

문서가 길어질수록 **격차가 오히려 벌어진다**.
80p+ 구간에서 Baseline은 11.7로 추락, DocSeeker는 31.8을 안정적으로 유지한다.

LATENCY vs ACCURACY TRADE-OFF · appendix C

ALR은 output token이 거의 2배(202→401)로 늘어나지만, end-to-end **latency는 19s → 25s (+31.6%)**에 그친다. 대신 **accuracy +14.7pt**를 얻는다.
bottleneck이 text decoding이 아닌 visual pre-fill에 있기 때문이다. 실용 관점에서 합리적인 trade-off.

Qualitative Trace & RAG Integration

DocSeeker를 retriever 위에 그대로 올린다. top-k dilemma가 완화되고 k 값에 둔감해진다. (Fig. 4)

QUALITATIVE — ALR output trace

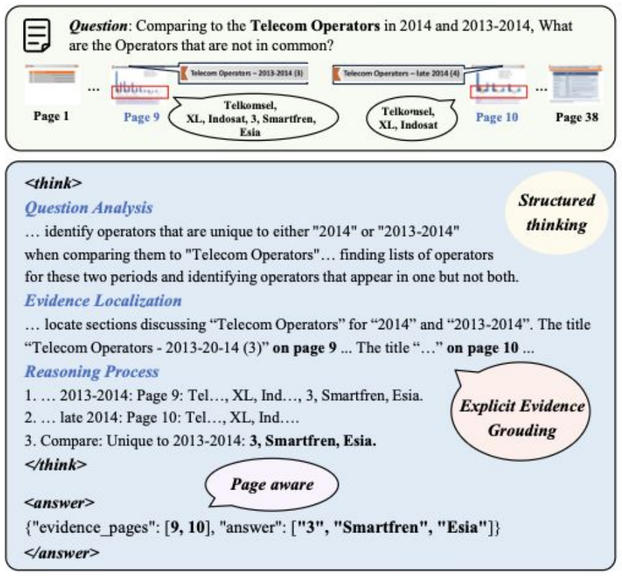


Fig. 1(b) · <think> Analysis → Localization → Reasoning </think> → <answer>

RAG INTEGRATION — robustness across k

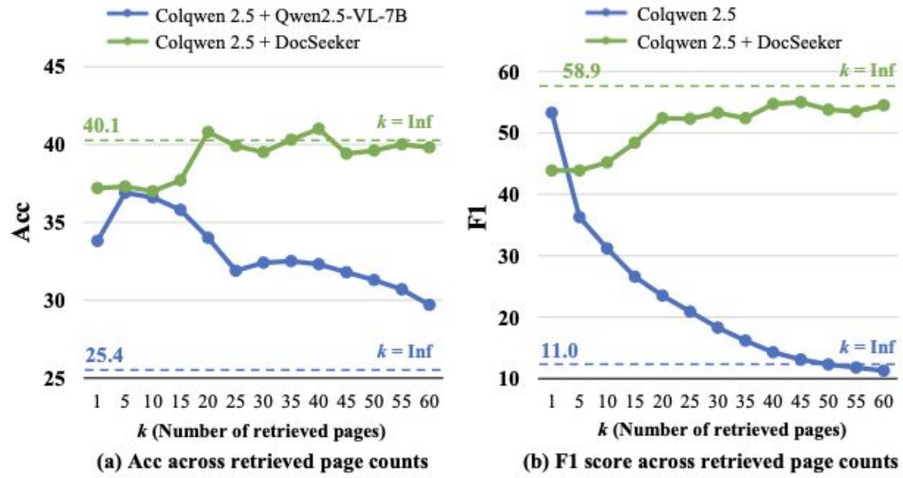


Fig. 4 · Acc / F1 vs retrieved page count k — DocSeeker stays flat

NATURAL SYNERGY

Baseline은 k>5 구간에서 무너지지만, DocSeeker는 k=1–60 전 구간에서 성능이 유지된다. 그대로 RAG reader로 투입 가능한 실용 구조이다.

Failure Analysis · 세 가지 실패 모드

MMLongBench-Doc 1,082 sample을 Recall × Accuracy 조합으로 나눠 stage-wise로 분류한 결과. (Tab. A3)

TABLE A3 — Stage-wise Error Breakdown

	Acc ≥ 0.5	Acc = 0	Total
Recall = 1	300	288	588
Recall < 1	121	373	494
Total	421	661	1,082

Recall = Evidence Localization 정확도 · Acc = final answer 정확도

THREE FAILURE MODES — appendix D

01 Reasoning Failures

Recall = 1 · Acc = 0 — evidence는 제대로 찾았지만 **Reasoning Process에서 오류**를 낸 case (288건).
예: chart 수치는 정확히 읽었으나 산술 단계에서 잘못 계산.

02 Localization Failures

Recall < 1 · Acc = 0 — **evidence page를 찾지 못해** 추론에 필요한 context도 확보하지 못한 case (373건, 가장 큰 비중).
→ 향후 개선 여지가 가장 큰 지점.

03 Formatting Mismatches

답은 의미상 맞지만 **format이 다르다는 이유**로 Exact Match에 탈락 (예: “17th February 1916” vs “1916-02-17”).
→ 학습 단계에서는 Secondary Verification으로 회수되지만, 평가 metric의 구조적 한계이기도 하다.

⇒ *전체 오답 **661건 (61%)** 중 **Localization Failures (373건)**가 최대 비중 — **Grounding 강화가 곧 다음 우선순위**.*

IV.

Strengths

이 논문이 특히 잘한 다섯 가지 지점

01 문제 정의가 명확하다

long-document VQA의 본질을 Low SNR과 sparse supervision 두 가지로 명확히 분리해 다뤄 잘 읽힌다.

02 아키텍처를 건드리지 않는다

Qwen2.5-VL backbone을 그대로 두고 학습 절차만으로 능력을 주입하므로 재현과 확장이 쉬움.

03 해석 가능성을 설계에 내장했다

출력에 evidence page가 명시되므로 사용자가 답을 직접 검증할 수 있다.

04 학습 효율이 뛰어나다

13K 규모의 학습 데이터만으로 short → long 일반화를 달성. EGRA로 GPU 메모리 한계 안에서 long document 학습이 가능하다.

05 RAG 시스템과 자연스럽게 결합된다

k에 둔감한 reader이기 때문에 multi-document·ultra-long document 설정에 바로 투입 가능하다.

*"**SFT가 ALR pattern을 가르치고, EviGRPO가 Evidence Localization을 안정화한다**" — 두 단계의 역할이 깔끔하게 나뉘어 있다.*

V.

Limitations · Open Questions

LIMITATION · 01

Teacher 의존성 (Gemini-2.5-Flash)

ALR CoT 데이터의 품질이 외부 LLM에 종속된다. teacher가 약해지면 ALR 신호 자체가 흔들리게 된다.

LIMITATION · 02

Evidence가 page 단위에 머무른다

한 페이지 안의 다중 영역·표·caption 단위 fine-grained grounding은 아직 다루지 못한다.

LIMITATION · 03

Backbone 일반성 검증이 부족하다

실험이 Qwen2.5-VL-7B에 국한되어 InternVL3·GLM-4.5V 등 다른 backbone에서도 ALR이 같은 폭으로 작동하는지는 미확인.

OPEN QUESTION

Closed-source와의 격차는 남아 있다

MMLong. 40.1 vs GPT-4o 42.8 · LongDoc. 51.7 vs 64.5 — ALR을 더 큰 모델에 올렸을 때 성능 상한이 어디인지는 미지수.

Take-home Message

"Find first.
Then reason."

DocSeeker는 long-document visual reasoning을 **“evidence를 먼저 찾고, 그 위에서 추론한다”**는
구조화된 paradigm으로 재정의한다.

PARADIGM

ALR

Analysis · Localization · Reasoning
find-then-reason을 CoT로 내장한다

TRAINING

SFT + EviGRPO

Data Distillation + Evidence-aware RL
+ EGRA로 메모리 효율화

OUTCOME

SOTA + Robust

5개 benchmark SOTA · 80p+ 문서 안정성
RAG reader로 즉시 투입 가능

감사합니다 · Thank you · github.com/yh-hust/DocSeeker