ELDER: LEARNING COMPACT EVIDENCE FOR LONG-DOCUMENT VISUAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-document visual reasoning is constrained by visual-token budgets, as processing full pages allocates substantial input capacity to content unrelated to the question. To address this problem, we introduce **E**vidence **L**earning for **D**ocument-**E**lement **R**easoning **(ELDER)**, a dual-role framework that reduces visual-token input for answer generation while preserving the internal layout of selected evidence. A shared vision-language model selects complete document elements and answers using only their original crops. Supervised fine-tuning uses families of verified evidence sets to initialize both roles, teaching multiple supporting combinations for the same question instead of a single canonical selection. We further formulate **M**inimal **R**eliable **E**vidence **G**roup **R**elative **P**olicy **O**ptimization **(MRE-GRPO)** to compare alternative evidence sets by the mean quality of multiple answers generated from each set and its visual-token cost. This set-level comparison can favor complementary combinations and useful corroboration when their answer-quality gains justify the added tokens, and discourage extra content that does not improve sampled answers. A separate answer objective improves reasoning under each fixed set, coupling evidence compression with learning to use retained content. Across five benchmarks, ELDER achieves an average score of 67.6, exceeding the strongest external method with results on all five by 2.3 score points. Compared with answer-only reinforcement learning, ELDER reduces mean answer-input visual tokens by 43.2%, with a 0.3-point decrease in average score.

1 INTRODUCTION

With the rapid advancement of Vision-Language Models (VLMs) Liu et al. (2023); Bai et al. (2023); Wang et al. (2024); Chen et al. (2024b); Team (2024), Document Visual Question Answering (DocVQA) has made remarkable progress. Unlike natural images, document images are visually cluttered and information-dense, where answers are often grounded in textual semantics, spatial layouts, tables, figures, and other visual elements. Existing document VLMs typically follow a one-pass paradigm, feeding full visual pages into the model to generate answers Ye et al. (2023a); Liu et al. (2023). However, this becomes inefficient and unreliable for long or complex documents, since the evidence required by most questions occupies only a small portion of one or a few pages. Feeding entire pages introduces substantial visual noise, increases computation, distracts attention, and may lead to hallucinated reasoning Liu et al. (2024; 2026); Mo et al. (2025).

Recent Visual Document Retrieval-Augmented Generation (VD-RAG) methods Faysse et al. (2024); Yu et al. (2024a); Xiong et al. (2026); Zheng et al. (2026) reduce document-level noise by retrieving top-k relevant pages before answer generation. Nevertheless, page-level retrieval is still coarse-grained: even when the correct pages are retrieved, most intra-page regions remain irrelevant to the question. As illustrated in Fig. 1, chunk-based retrieval shortens the input but breaks document layout, while page-level retrieval preserves visual context but still passes substantial intra-page noise to the generator. These limitations motivate a shift from chunk- or page-level retrieval to evidence-level reasoning. At the element level, however, compactness depends on relationships within the selected set: tables, charts, and paragraphs may provide complementary facts or useful corroboration. The challenge is to reduce visual input while retaining the combination that supports accurate answering, rather than requiring every repeated element to be individually indispensable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

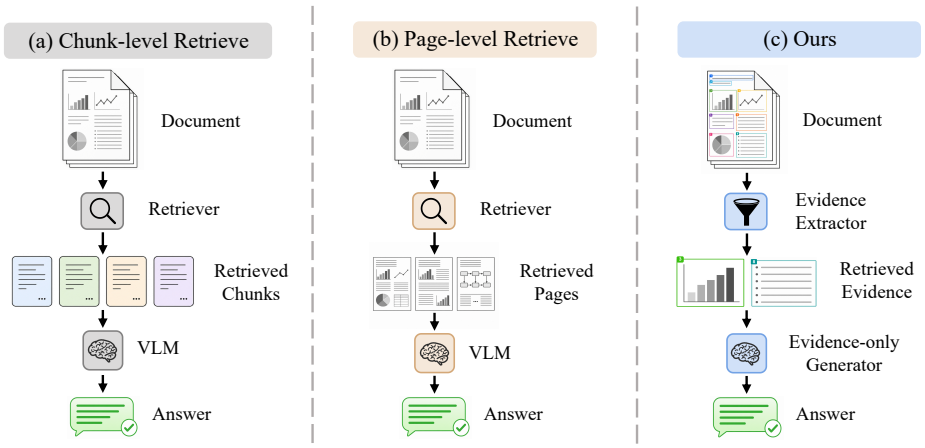

Figure 1: Comparison of document reasoning paradigms. (a) Chunk-level retrieval feeds disconnected text snippets to the VLM, which may lose visual layout information. (b) Page-level retrieval preserves visual context but still passes full retrieved pages with substantial intra-page noise. (c) Our method selects complete document elements and restricts answer generation to their original crops; selection quality is assessed jointly with visual-input cost.

To address these issues, we propose Evidence Learning for Document-Element Reasoning (ELDER), a dual-role evidence-to-answer framework for long-document visual reasoning. ELDER uses a shared VLM as an *Evidence Extractor* to select indexed document elements and as an *Evidence-only Generator* to answer using only the selected cropped visual evidence regions. The objective is *compact evidence with reliable task performance*, allowing corroborating regions when they help the reader. We formulate Minimal Reliable Evidence GRPO (MRE-GRPO) by jointly optimizing evidence selection and answering with a clipped GRPO objective Shao et al. (2024). The Extractor compares alternative evidence sets by downstream answer score and visual-token cost; the Generator compares answers under the same evidence and receives a direct task signal. An adaptive cost coefficient and a zero-feedback guard limit pressure to compress when sampled answer quality is poor, without implying a hard reliability guarantee.

Our contributions are summarized as follows:

- We introduce ELDER, a dual-role framework built on a shared VLM that selects complete document elements and answers from their original crops, reducing answer-input visual tokens while preserving internal layout. Verified evidence families initialize both roles with alternative supporting combinations and crop-conditioned answering, including abstention on insufficient evidence.
- We formulate MRE-GRPO with nested evidence and answer rollouts to compare whole evidence sets by sampled answer quality and visual-token cost, while improving answers within each fixed set. Executable-set checks, adaptive cost pressure, and a zero-feedback guard regulate this tradeoff, and two complete role-specific GRPO losses jointly train the shared policy.
- Across five benchmarks, ELDER achieves an average score of 67.6, exceeding the strongest external method with results on all five benchmarks by 2.3 score points. Compared with answer-only reinforcement learning, it reduces mean answer-input visual tokens by 43.2%, with a 0.3-point decrease in average score.

2 Related Work

Document Understanding. Document understanding has progressed from OCR-based pipelines Smith (2007); Xu et al. (2020) to layout-aware models such as LayoutLMv2 Xu et al. (2021), DocFormer Appalaraju et al. (2021), and StrucTexT, which incorporate spatial and structural priors for better document representation. However, these methods still largely follow a decoupled

perception–reasoning pipeline, limiting holistic document-level reasoning. Recent multimodal large language models (MLLMs) further unify perception and reasoning into end-to-end frameworks. Models like UReader Ye et al. (2023b) and Donut Kim et al. (2021) enable OCR-free document understanding, while mPLUG-DocOwl Hu et al. (2024) improves efficiency via token compression. However, scaling to long documents remains a major challenge Beltagy et al. (2020); Zaheer et al. (2020). Increasing input length introduces redundancy and complex cross-page dependencies, which are often addressed by independent page processing or coarse compression Huang et al. (2022), leading to inefficient or incomplete reasoning.

Fine-grained Evidence Retrieval for Visual Document Reasoning. Retrieval plays a central role in visual document reasoning, with existing methods differing mainly in evidence granularity. Early visual RAG methods, such as ColPali Faysse et al. (2024) and VisRAG Yu et al. (2024a), retrieve top-k relevant pages, but page-level evidence often contains substantial irrelevant content, leading to attention dilution and potential hallucination. Recent works therefore move toward finer-grained evidence selection. Page-level localization methods such as DocThinker Yu et al. (2025) first identify relevant pages before performing reasoning, improving efficiency over full-document reasoning. Beyond page selection, region-level methods further select visual evidence areas: Doc-CoB Mo et al. (2025) models evidence selection as query-conditioned bounding-box reasoning, while AgenticOCR Wang et al. (2026b) dynamically extracts task-relevant regions during reasoning. These methods motivate studying selection at the set level: an individually relevant region can be redundant given other regions, while repeated evidence can help resolve ambiguity. Our objective measures the answer-quality and visual-token tradeoff of complete region sets; it does not infer semantic sufficiency from answer correctness alone.

RL for VLM Reasoning. Reinforcement learning (RL) has recently become an effective paradigm for enhancing reasoning in large language models (LLMs). While early RL-based alignment relies on human feedback or learned reward models Ouyang et al. (2022); Schulman et al. (2017), DeepSeek-R1 demonstrates that rule-based verifiable rewards with Group Relative Policy Optimization (GRPO) elicit strong emergent reasoning Guo et al. (2025). Building on this, RL has been extended to vision-language models (VLMs) and multimodal LLMs, including VisualRFT for fine-grained visual grounding Liu et al. (2025), Vision-R1 for multimodal math reasoning Huang et al. (2025), MedVLM-R1 for medical image reasoning Pan et al. (2025), and RLHF-V for trustworthy alignment Yu et al. (2024b). DocThinker applies GRPO to document understanding Yu et al. (2025). Our focus is quality-controlled compression of complete document-element sets, with direct answer supervision and explicit treatment of invalid and zero-feedback rollouts.

3 METHODOLOGY

3.1 PRELIMINARY: GROUP RELATIVE POLICY OPTIMIZATION

GRPO Shao et al. (2024) uses sampled group rewards instead of a learned critic. For a role input x, output o, and detached sequence advantage A, we minimize the complete per-trajectory loss

$$\ell(o, A; x) = \frac{1}{|o|} \sum_{t=1}^{|o|} \left[-\min(\rho_t A, \mathrm{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A) + \beta d_t \right], \tag{1}$$

where $h_t = (x, o_{<t})$, $\rho_t = \pi_\theta(o_t \mid h_t)/\pi_{\theta_{\mathrm{old}}}(o_t \mid h_t)$, and $d_t = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_t) \| \pi_{\mathrm{ref}}(\cdot \mid h_t))$. The old policy generates the rollout batch; the frozen reference is the SFT checkpoint. ϵ defines the ratio-clipping interval and β controls reference regularization. Only generated output tokens enter the loss; our role-specific advantages are defined in Eq. 5.

3.2 DATA CONSTRUCTION

As shown in Fig. 2, we construct evidence-set supervision from DUDE document–question–answer tuples (D, q, y). MinerU2.5-Pro Wang et al. (2026a) parses complete layout regions, to which we assign global indices. Gemini-3-Flash-Preview proposes regions page by page; merged proposals, alternatives, and subsets form candidate evidence sets. Two separate teacher checks inspect required facts and missing support, while an answer-blind reader answers from the crops. Agreeing support checks and a correct reader response admit sufficient sets; confirmed missing facts label insufficient

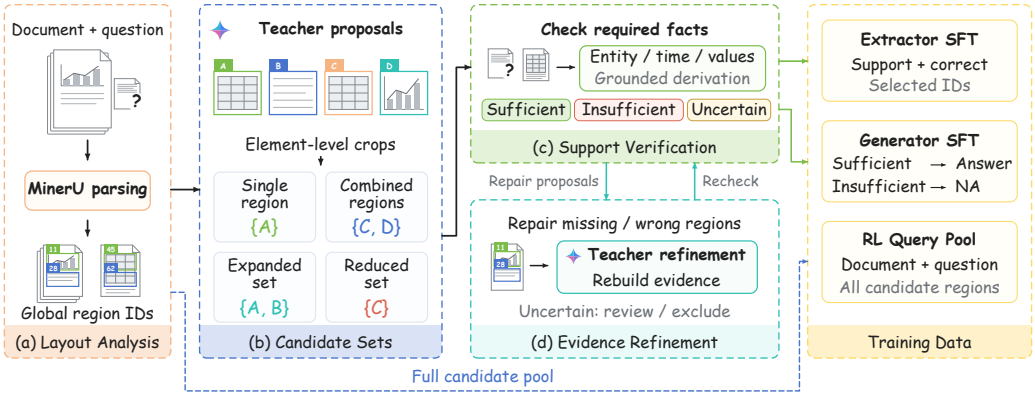

Figure 2: Evidence-set construction from complete document elements: layout parsing, candidate proposals, support verification, and evidence refinement. Verified sets supply role-specific SFT targets, while RL uses the full candidate pool for online selection. NA denotes `Not answerable`; construction details are in Appendix A.1.

sets. A wrong reader answer alone does not establish missing support. Uncertain cases are excluded, and insufficient base sets are repaired and rechecked. Evidence expansion adds question–set pairs without creating independent questions. Verified pairs supply SFT targets, whereas RL exports at most one input per source question with the full candidate inventory.

3.3 ELDER: Evidence-Level Document Reasoning

The shared policy π_θ uses role prompts r_E and r_G for selection and answering. Given question q, indexed candidate pages $\mathcal{I}_\mathcal{C}$, and their region inventory $\mathcal{C}$, the *Evidence Extractor* samples $o^E \sim \pi_\theta(\cdot \mid x^E)$, with $x^E = (q, \mathcal{I}_\mathcal{C}, \mathcal{C}, r_E)$. Its JSON `index` list selects a document-level set $E \subseteq \mathcal{C}$. Each index identifies a complete parsed region with its internal headings and labels, not a freely predicted box or an isolated table cell. The *Evidence-only Generator* samples $o^G \sim \pi_\theta(\cdot \mid x^G(E))$, where $x^G(E) = (q, \mathcal{V}(E), r_G)$ and $\mathcal{V}(E)$ contains the original crops in fixed ID order. Its JSON `prediction` field supplies the answer. Full pages, the Extractor's output text, reference answers, and teacher checks are excluded from this input. An empty set is a valid no-evidence action and gives the Generator an explicit `no_evidence` flag, with no image attached. We compare whole evidence sets, allowing alternative or corroborating regions when they improve reader performance.

3.4 SFT Cold Start

Alternating role batches initialize the shared VLM for document-element selection and crop-conditioned answering (Appendix A.2). The Extractor learns document-level index sets from verified sufficient evidence, allowing several accepted sets for the same question instead of imposing a unique minimal target. The Generator learns the answer-blind reader's analysis and prediction from sufficient crops, and a fixed insufficiency statement with `Not answerable` from confirmed insufficient crops. Crop-level abstention labels do not replace the original document answer used for RL scoring.

3.5 Minimal Reliable Evidence GRPO

For each rollout batch, the fixed old policy samples $K > 1$ Extractor outputs from the full candidate pool and $M > 1$ answers for each executable evidence set (Fig. 3). All branches use the same policy snapshot. Multiple answers estimate each selected set's mean task performance under the current reader; comparisons under fixed evidence then assess answer-generation choices. The resulting feedback evaluates each selection as a joint action, so a region's contribution is not estimated by separately removing it and regenerating answers. Both roles receive feedback from online samples rather than teacher-forced sets.

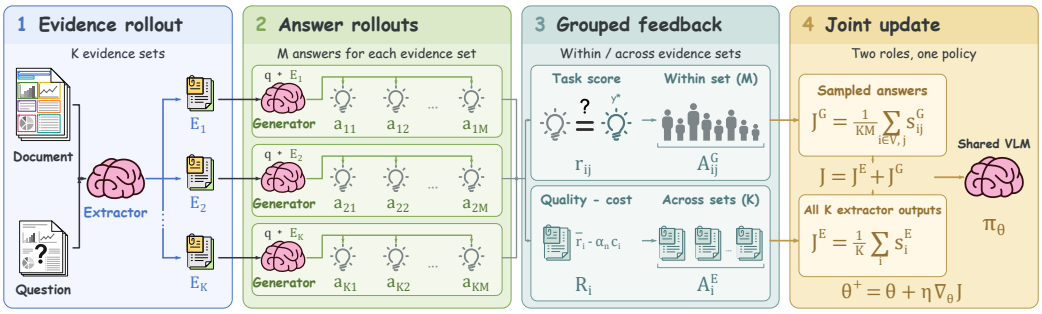

Figure 3: MRE-GRPO with nested evidence and answer rollouts. Diagram labels a_{ij}, y^*, R_i denote $\hat{y}_{ij}, y, R_i^E$, respectively; V abbreviates $\mathcal{I}_{\text{valid}} = \{i : v_i^E = 1\}$, distinct from the crops $\mathcal{V}(E)$. The displayed cost assumes $z_q = 1$; Eq. 3 covers all cases. Here $s = -\ell$ includes KL and $\mathbb{E}_q[J] = -\mathcal{L}_{\text{MRE}}$. The ascent step is schematic, with policy learning rate η; training uses AdamW.

3.5.1 Validity, Task Score, and Cost

Let v_i^E indicate a parseable, duplicate-free index list with existing IDs, readable crops, and visual-token count $T_{\text{vis}}(E_i) \le B$. Here $B > 0$ is the fixed Generator visual-token budget under the chosen image processor. Let v_{ij}^G indicate a complete, parseable answer output. There is no positive format bonus. Invalid Extractor outputs stay in the group but do not invoke the Generator; their answer-score slots are zero. For valid branches,

$$r_{ij} = v_{ij}^G \,\text{Score}(\hat{y}_{ij}, y), \qquad \bar{r}_i = \frac{1}{M}\sum_{j=1}^{M} r_{ij}, \qquad c_i = \frac{T_{\text{vis}}(E_i)}{B} \in [0,1]. \tag{2}$$

Score uses the task's normalized answer metric and the original reference y. For an answerable document, abstaining after a poor selection scores zero; correct abstention can score positively only when the original reference is unanswerable. The sample mean estimates task performance, not semantic support or a certified success probability. The budget bounds executable visual input, while the cost term below ranks selections within that bound; selecting fewer regions need not reduce cost when their visual-token counts differ.

3.5.2 Quality-Controlled Reward

Define $z_q = \mathbf{1}[\max_{i,j} r_{ij} > 0]$ over all nominal score slots. The rewards are

$$R_i^E = \begin{cases} \bar{r}_i - \alpha_n z_q c_i, & v_i^E = 1, \\ -1, & v_i^E = 0, \end{cases} \qquad R_{ij}^G = r_{ij} \quad (v_i^E = 1). \tag{3}$$

For an all-zero question, the guard removes the preference for cheaper failure but supplies no successful-answer signal; valid and invalid selections remain distinguishable. For valid sets in a question with positive feedback, adding evidence improves Extractor reward only when the gain in mean task score exceeds the scaled increase in visual cost. The adaptive coefficient is

$$\alpha_n = (1+\lambda_n)^{-1}, \qquad \lambda_{n+1} = \left[\lambda_n + \eta_\lambda(\tau - \widehat{r}_n)\right]_+. \tag{4}$$

$\widehat{r}_n$ averages the full rollout batch, including invalid and all-zero slots. τ is calibrated from a frozen SFT checkpoint on a separate fixed stream from the active training pool, using the same sampling and scoring protocol; $\eta_\lambda > 0$ is the controller step size. Below-target quality increases λ_n and reduces cost pressure on the next batch. This is a practical surrogate, not a hard per-question accuracy constraint.

3.5.3 Two Role-Specific GRPO Losses

We compare Extractor rewards across sets and Generator scores only within the same set:

$$A_i^E = R_i^E - \frac{1}{K-1}\sum_{k\neq i} R_k^E, \qquad A_{ij}^G = r_{ij} - \frac{1}{M-1}\sum_{l\neq j} r_{il}. \tag{5}$$

The within-set Generator baseline compares answers under identical crops; a cross-set baseline would mix answer-generation differences with changes in the available evidence. No group standard deviation rescales these advantages, preserving the magnitude of quality–cost differences. Given E_i, visual cost is constant across answers and is omitted from the Generator comparison. All actual answer rollouts participate. Equal-reward groups have zero policy advantage, while KL regularization remains active. Using the complete loss in Eq. 1,

$$\begin{aligned} \mathcal{L}_E^{\mathrm{GRPO}} &= \mathbb{E}_q \left[\frac{1}{K} \sum_i \ell(o_i^E, A_i^E; x^E) \right], \\ \mathcal{L}_G^{\mathrm{GRPO}} &= \mathbb{E}_q \left[\frac{1}{KM} \sum_{i \in \mathcal{I}_{\mathrm{valid}}} \sum_j \ell(o_{ij}^G, A_{ij}^G; x^G(E_i)) \right], \\ \mathcal{L}_{\mathrm{MRE}} &= \mathcal{L}_E^{\mathrm{GRPO}} + \mathcal{L}_G^{\mathrm{GRPO}}. \end{aligned} \tag{6}$$

The nominal KM denominator is retained when branches are invalid. Thus invalid branches do not increase the weight of surviving answer trajectories. Separate K and KM averages prevent the number of answer samples from determining role weights. Discrete selection is optimized through the Extractor's policy loss, without differentiating through the crop-selection operation. Each role loss includes its own KL term; no additional KL is added. Rewards, advantages, old-policy probabilities, and controller states are detached. The two losses update shared parameters and may have conflicting gradients. Appendix A.3 specifies parsing, calibration, zero-feedback cases, and the complete update sequence.

4 Experiments

4.1 Experimental Setup

Training Data. Our DUDE Van Landeghem et al. (2023) construction inventories expand from 7,629 to 17,953 SFT samples and from 4,572 to 8,788 RL-associated samples, adding 10,324 and 4,216 samples while preserving question identities and counts. One sample is a question–evidence-set pair counted after verification and deduplication, not a role-specific export or an online rollout. Different verified sets for the same question count separately; repeated instances of the same pair count once. Appendix A.1 details construction and role-specific exports.

Benchmarks and Metrics. We evaluate **ELDER** on five representative multi-page and long-document visual question answering benchmarks: MPDocVQA Kang et al. (2024), DUDE Van Landeghem et al. (2023), SlideVQA Tanaka et al. (2023), LongDocURL Deng et al. (2025), and MMLongBench-Doc Ma et al. (2024). These benchmarks cover diverse document reasoning scenarios, including scanned documents, forms, tables, slides, web documents, and ultra-long multimodal documents, ranging from multi-page evidence localization to long-context visual reasoning. We use the official benchmark metrics, with imported baseline settings described in Appendix C. We report ANLS for DUDE and MPDocVQA, F1 for SlideVQA, and Accuracy for MMLongBench-Doc and LongDocURL. Higher values indicate better performance.

Baselines. We compare ELDER with a broad range of open-source and closed-source baselines for multi-page and long-document visual reasoning. The open-source baselines cover OCR- or parser-assisted methods, including HiVT5 Tito et al. (2023), CREAM Kim et al. (2023), Docopilot Duan et al. (2025), and DocVLM Nacson et al. (2025); end-to-end or OCR-free document VLMs, including mPLUG-DocOwl2 Hu et al. (2025) and InternVL3 Zhu et al. (2025); visual RAG methods, including M3DocRAG Cho et al. (2024), Vis-RAG Yu et al. (2024a), SV-RAG Chen et al. (2024a), and VDocRAG Tanaka et al. (2025); and recent agentic or evidence-seeking systems, including CogDoc Xu et al. (2025), Doc-V* Zheng et al. (2026), and DocSeeker Yan et al. (2026). We also include closed-source VLMs as reference points, including Gemini-1.5-Pro Team et al. (2024), GPT-4o Hurst et al. (2024), GPT-4.1, Claude-3.7, and GPT-4o mini. For baselines with available public

Table 1: Five-benchmark task scores. Mode: Full, full-document input; Chunk/Page, retrieval; Evidence, region selection. Avg is the unweighted five-benchmark mean, or "-" if any score is missing. **Bold** and underlining mark the best and second-best displayed scores.

Method	Mode	LongDoc.	MMLong.	SlideVQA	DUDE	MPDocVQA	Avg
Open-source							
HiVT5	Full	-	-	-	23.1	62.0	-
CREAM	Chunk	-	-	-	52.5	74.3	-
mPLUG-DocOwl2	Full	5.3	13.4	27.8	46.8	69.4	32.5
M3DocRAG	Page	35.1	21.0	55.7	39.5	84.4	47.1
Vis-RAG	Page	41.9	18.8	52.4	43.1	70.9	45.4
SV-RAG	Page	-	23.0	34.3	45.0	71.0	-
VDocRAG	Page	39.8	18.4	42.0	44.0	62.6	41.4
Docopilot	Full	-	28.8	43.1	40.7	81.3	-
DocVLM	Full	-	-	-	47.4	84.5	-
InternVL3	Full	38.7	24.1	64.4	47.4	80.8	51.1
CogDoc	Page	-	33.0	67.9	46.2	75.0	-
Doc-V*	Page	56.3	42.1	77.2	64.5	86.2	65.3
DocSeeker	Full	51.7	40.1	77.1	57.4	86.2	62.5
Closed-source							
Gemini-1.5-Pro	Full	50.9	28.2	-	46.0	-	-
GPT-4o	Full	**64.5**	42.8	65.8	54.1	67.4	58.9
GPT-4.1	Full	-	**45.6**	74.7	50.2	-	-
Claude-3.7	Full	-	33.9	76.3	58.1	-	-
GPT-4o mini	Full	-	28.6	60.7	46.5	-	-
Ours							
Baseline	Full	32.9	28.0	55.2	51.9	75.2	48.6
Dual-role SFT	Evidence	49.3	36.2	64.8	58.5	79.1	57.6
Answer-only RL	Evidence	<u>60.1</u>	<u>45.1</u>	**79.9**	<u>67.1</u>	<u>87.4</u>	**67.9**
ELDER (full)	Evidence	59.1	44.0	<u>78.4</u>	**68.2**	**88.3**	<u>67.6</u>

results, we use their reported scores under official evaluation protocols; otherwise, we reproduce them under the same benchmark settings whenever possible.

Implementation Details. We adopt Qwen2.5-VL-7B-Instruct as our baseline. All training is conducted on 8 NVIDIA A100 80GB GPUs. We employ the AdamW optimizer. In the SFT stage, we train the model on our distilled data for 2 epochs with a learning rate of 1×10^{-5}. In the RL stage, we train for 2 epochs with a learning rate of 1×10^{-6} and a rollout batch size of 16. We set the Evidence Extractor rollout group size to 4. For the Evidence-only Generator, we use a rollout group size of 4. The KL divergence coefficient and clipping ratio in the GRPO loss function are 0.04 and 0.2, respectively. For MRE-GRPO, the maximum rollout length is set to 1024 tokens for the Evidence Extractor and 2048 tokens for the Evidence-only Generator. More training details are provided in Appendix A.

4.2 Main Results

Table 1 shows consistent improvements from the base model to dual-role SFT and then to either RL variant across all five benchmarks. Full ELDER achieves the highest displayed scores on DUDE and MPDocVQA and leads the listed external open-source baselines on each benchmark. Stage-wise gains for full ELDER reveal a different balance across benchmarks: SFT contributes more of the overall improvement on the two long-document benchmarks, while RL contributes more on the remaining three. This pattern suggests that supervised initialization and reward optimization make complementary contributions whose relative importance varies with the benchmark and its document

Table 2: MRE-GRPO ablations: task score (↑) and mean answer-input visual tokens per question (T_{vis}, ↓), reported in k (1k = 1,000 tokens) to one decimal place. Avg is the unweighted five-benchmark mean of each metric, or "-" if any value is missing. Means use unrounded values. **Bold** and underlining mark ranks 1 and 2 based on displayed values; ties share a rank, and subsequent ranks are skipped.

Method	LongDoc.		MMLong.		SlideVQA		DUDE		MPDocVQA		Avg	
	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}
Baseline	32.9	68.5	28.0	38.0	55.2	16.0	51.9	4.6	75.2	6.6	48.6	26.7
Dual-role SFT	49.3	2.1	36.2	2.6	64.8	1.7	58.5	1.1	79.1	0.9	57.6	1.7
w/o Extractor RL loss	54.4	2.2	39.8	2.8	73.5	1.9	63.9	1.2	84.1	1.0	63.1	1.8
w/o Generator RL loss	55.8	1.6	40.5	2.0	72.4	1.4	64.6	0.9	85.5	0.7	63.8	1.3
w/o visual-token cost	**60.1**	2.5	**45.1**	3.1	**79.9**	2.1	67.1	1.3	87.4	1.0	**67.9**	2.0
Fixed cost ($\alpha = 1$)	58.8	1.4	43.7	1.7	78.0	1.2	68.0	**0.7**	88.1	**0.6**	67.3	1.1
w/o zero-feedback guard	58.4	**1.3**	43.3	**1.6**	77.6	**1.1**	67.8	**0.7**	88.0	**0.6**	67.0	**1.0**
Full MRE-GRPO	59.1	1.4	44.0	1.7	78.4	1.2	**68.2**	0.8	**88.3**	**0.6**	67.6	1.1

reasoning demands. Differences in backbones, training resources, and evaluation sets limit attribution of cross-system gaps to evidence granularity.

Answer-only RL increases Generator visual-token use relative to SFT on every benchmark, showing that higher task scores alone do not imply more compact inputs (Table 2). Full MRE-GRPO instead improves scores over SFT while reducing visual input throughout. Compared with answer-only RL, it substantially reduces visual tokens with a similar average score: DUDE and MPDocVQA improve in both dimensions, whereas the other benchmarks incur modest score decreases. The full objective therefore retains most of the RL performance gains while explicitly controlling evidence cost. The cost term changes the benchmark-level tradeoff rather than imposing a uniform accuracy loss, making evaluation of both metrics necessary even when average task scores are nearly unchanged. These token counts concern Generator input and do not establish lower end-to-end cost.

4.3 Ablation Study

Role-level comparisons. Removing either role-specific RL loss lowers task scores and increases Generator visual-token use across all benchmarks (Table 2). The quality–cost balance therefore benefits from both selection feedback and direct answer optimization. Removing the Extractor loss causes a larger token increase than removing the Generator loss, indicating a stronger effect on evidence compactness. The reversed score ordering on SlideVQA highlights task-dependent contributions from the two learning signals. Each ablation removes the complete role loss, including KL; both inference roles remain active, and shared parameters continue to change. Figure 4 shows the two rewards over training progress.

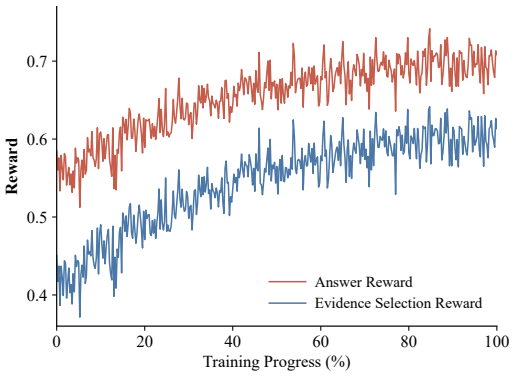

Figure 4: Answer and evidence-selection rewards during RL training.

Cost and feedback comparisons. Removing visual-token cost produces larger inputs without consistent score improvements, indicating that answer reward alone does not enforce evidence compactness. Conversely, fixing the cost coefficient or removing the zero-feedback guard further compresses inputs but slightly weakens task performance across benchmarks. The full objective reduces evidence relative to answer-only RL while preserving more answer quality than these more aggressive compression variants. This pattern supports using task feedback to regulate compression. The two ablations probe different scales of feedback—batch-level quality and question-level failure—yet both move toward smaller inputs and lower scores, suggesting complementary roles in limiting excessive compression.

Table 3: Full-document controls and frozen checkpoint combinations. FD denotes full-document input; E/G denote the Extractor/Generator and S/R denote SFT/MRE-GRPO checkpoints. In superscripts, D denotes training-data progress (question identities and cumulative exposures), and H denotes cumulative training GPU-hours; Appendix C specifies the matched stages. Each pair reports Score (↑) and T_{vis} (↓, k). Avg averages five benchmarks. Bold and underlining mark ranks 1 and 2 by displayed values; ties skip subsequent ranks.

Method	LongDoc.		MMLong.		SlideVQA		DUDE		MPDocVQA		Avg	
	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}	Score	T_{vis}
$\text{FD}_{\text{SFT}}^{D=D_{E_SG_S}}$	42.7	68.5	31.9	38.0	64.1	16.0	60.0	4.6	81.8	6.6	56.1	26.7
$\text{FD}_{\text{SFT+GRPO}}^{D=D_{E_RG_R}}$	51.6	68.5	38.5	38.0	73.0	16.0	65.3	4.6	86.5	6.6	63.0	26.7
$\text{FD}_{\text{SFT}}^{H=H_{E_RG_R}}$	42.9	68.5	32.0	38.0	65.0	16.0	61.1	4.6	82.6	6.6	56.7	26.7
$\text{FD}_{\text{SFT+GRPO}}^{H=H_{E_RG_R}}$	52.3	68.5	39.0	38.0	73.8	16.0	<u>66.0</u>	4.6	<u>87.0</u>	6.6	63.6	26.7
E_SG_S	49.3	2.1	36.2	2.6	64.8	1.7	58.5	1.1	79.1	0.9	57.6	1.7
E_RG_S	52.6	**1.4**	38.4	**1.7**	68.6	**1.2**	62.2	**0.8**	83.0	**0.6**	61.0	**1.1**
E_SG_R	<u>55.5</u>	2.1	<u>40.7</u>	2.6	<u>74.4</u>	1.7	64.0	1.1	84.7	0.9	<u>63.9</u>	1.7
E_RG_R	**59.1**	**1.4**	**44.0**	**1.7**	**78.4**	**1.2**	**68.2**	**0.8**	**88.3**	**0.6**	**67.6**	**1.1**

4.4 Full-Document Controls and Checkpoint Combinations

Full-document comparisons. Across the full-document controls in Table 3, adding answer-reward GRPO consistently improves on SFT. Nevertheless, full ELDER outperforms even the strongest of these controls on every benchmark while using substantially fewer Generator visual tokens. The advantage thus persists when full-document answering also benefits from RL, supporting the value of learning to select compact evidence alongside answering. The SFT evidence pipeline trails the full-document SFT controls on DUDE and MPDocVQA, but this ordering reverses after RL, indicating that the final advantage is not guaranteed by cropping alone. Smaller answer inputs therefore need not impose a performance penalty relative to processing complete documents.

Fixed-checkpoint comparisons. We evaluate frozen checkpoint combinations without further training, reusing each Extractor's crops across both Generators. With either Generator fixed, the RL Extractor improves scores while selecting fewer visual tokens; with either Extractor fixed, the RL Generator improves scores on the same evidence. Both effects hold across all benchmarks, and combining the RL checkpoints gives the best scores. The gains therefore reflect improvements in both selection and evidence-conditioned reading, rather than answering alone. This compatibility suggests that improvements are not tied exclusively to the jointly updated partner, since both mixed pairings retain gains over the original SFT pair. These inference comparisons do not isolate the training losses, which update shared parameters.

5 Conclusion

ELDER learns compact evidence selection and crop-conditioned answering within one shared VLM. Verified evidence families initialize both roles, and MRE-GRPO connects evidence-set comparisons with direct answer optimization. Across five benchmarks, ELDER reaches an average score of 67.6, exceeding the strongest external method with complete results by 2.3 points. Compared with answer-only RL, it reduces mean answer-input visual tokens by 43.2% with a 0.3-point decrease in average score. Frozen-checkpoint comparisons show that better selection benefits a fixed Generator and better generation benefits fixed evidence, supporting gains from both roles. These findings support learning evidence combinations together with the reader while preserving complete visual regions. Further evaluation should measure end-to-end cost and evidence support independently of answer accuracy.

AI USE STATEMENT

Generative AI tools were used solely to polish the language and improve the readability of the manuscript. The authors take full responsibility for the final content, including all claims, references, and reported results.

ETHICS STATEMENT

This work studies compact evidence selection and evidence-conditioned answering for long-document visual reasoning. ELDER uses existing DUDE document–question–answer records to construct teacher-generated evidence annotations and crop-conditioned supervision, and is evaluated on established document understanding benchmarks. The resulting models may inherit biases and errors from the underlying datasets, document parser, and teacher and backbone models. Parsing failures, incomplete evidence selection, and generation errors may produce unsupported answers or omit relevant context. Teacher verification and improved benchmark performance do not guarantee evidence sufficiency, fairness, safety, or suitability for high-stakes deployment. Applications involving personal or confidential documents should respect privacy, access restrictions, and data-use permissions, with human review before consequential decisions.

REPRODUCIBILITY STATEMENT

Sections 3.1–3.5 describe the ELDER framework, data construction, supervised initialization, and MRE-GRPO objective, while Section 4.1 summarizes the experimental setup. Appendix A.1 and Table 4 document teacher-guided evidence construction and dataset statistics. Appendices A.2–A.3 detail the SFT and RL procedures, hyperparameters, reward computation, and visual-token accounting. Appendix B provides the teacher and role-specific prompt templates. Appendices C and D describe the baselines and evaluation benchmarks and metrics, respectively. Sections 4.2–4.4 present the main comparisons, component ablations, full-document controls, and frozen-checkpoint analyses. Figure 4 shows the training reward curves, and Appendix E provides additional latency analyses.

REFERENCES

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 993–1003, 2021.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A Rossi, Changyou Chen, and Tong Sun. Sv-rag: Lora-contextualizing adaptation of mllms for long document understanding. *arXiv preprint arXiv:2411.01106*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1135–1159, 2025.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, et al. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4026–4037, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*, 2025.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 3096–3120, 2024.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5817–5834, 2025.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the ACM International Conference on Multimedia (MM)*, pp. 4083–4091, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Lei Kang, Rubèn Tito, Ernest Valveny, and Dimosthenis Karatzas. Multi-page document visual question answering using self-attention scoring mechanism. In *International Conference on Document Analysis and Recognition*, pp. 219–232. Springer, 2024.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 2021.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. Visually-situated natural language understanding with contrastive reading model and frozen large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11989–12010, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu, and Enhong Chen. Look as you think: Unifying reasoning and visual evidence attribution for verifiable document rag via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32159–32167, 2026.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2034–2044, 2025.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

Ye Mo, Zirui Shao, Kai Ye, Xianwei Mao, Bo Zhang, Hangdi Xing, Peng Ye, Gang Huang, Kehan Chen, and Zhou Huan. Doc-CoB: Enhancing multi-modal document understanding with visual chain-of-boxes reasoning. *arXiv preprint arXiv:2505.18603*, 2025.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. Docvlm: Make your vlm an efficient reader. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29005–29015, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–347. Springer, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ray Smith. An overview of the tesseract ocr engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp. 629–633, 2007.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023.

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24827–24837, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834, 2023.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Józiak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanisławek. Document understanding dataset and evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023. URL https://arxiv.org/abs/2305.08455.

Bin Wang, Tianyao He, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Tao Chu, Yuan Qu, Zhenjiang Jin, Weijun Zeng, Ziyang Miao, et al. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. *arXiv preprint arXiv:2604.04771*, 2026a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint*, 2024.

Zhengren Wang, Dongsheng Ma, Huaping Zhong, Jiayu Li, Wentao Zhang, Bin Wang, and Conghui He. Agenticocr: Parsing only what you need for efficient retrieval-augmented generation. *arXiv preprint arXiv:2602.24134*, 2026b.

Junyu Xiong, Yonghui Wang, Weichao Zhao, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. Docr1: Evidence page-guided grpo for multi-page document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11178–11186, 2026.

Qixin Xu, Haozhe Wang, Che Liu, Fangzhen Lin, and Wenhu Chen. Cogdoc: Towards unified thinking in documents. *arXiv preprint arXiv:2512.12658*, 2025.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Wanxiang Che. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2579–2591, 2021.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1192–1200, 2020.

Hao Yan, Yuliang Liu, Xingchen Liu, Yuyi Zhang, Minghui Liao, Jihao Wu, Wei Chen, and Xiang Bai. Docseeker: Structured visual reasoning with evidence grounding for long document understanding. *arXiv preprint arXiv:2604.12812*, 2026.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023a.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, and Ji Zhang. UReader: Universal ocr-free visually-situated language understanding with multimodal large language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 2841–2858, 2023b.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024a.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024b.

Wenwen Yu, Zhibo Yang, Yuliang Liu, and Xiang Bai. Docthinker: Explainable multimodal large language models with rule-based reinforcement learning for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 837–847, 2025.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, and Li Yang. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 17283–17297, 2020.

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, et al. Doc-v*: Coarse-to-fine interactive visual reasoning for multi-page document vqa. *arXiv preprint arXiv:2604.13731*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

A Training Details **15**

A.1 Teacher-guided Data Construction 15

A.2 SFT Cold Start 17

A.3 MRE-GRPO Training 18

B Prompt Templates **21**

C Details of Baselines **23**

D Details of Benchmarks **26**

E Additional Experiments **29**

F Case Study **30**

G Limitations and Future Work **31**

A TRAINING DETAILS

This section describes the teacher-guided data construction, training configuration, and interfaces to the two training roles. Teacher proposals are generated page by page, but the retained evidence labels are document-level sets of globally indexed regions. Page-wise proposal generation is distinct from student selection: an Extractor record contains the question, indexed candidate pages, and the complete candidate-region inventory, with one document-level index list as its target. The Generator receives only the selected crops. Region indices retain their document-wide identity throughout construction, SFT, and online rollouts.

Shared training configuration. We adopt Qwen2.5-VL-7B-Instruct as the baseline and conduct all training on 8 NVIDIA A100 80GB GPUs. Both SFT and RL use the AdamW optimizer; their stage-specific settings are given below.

A.1 TEACHER-GUIDED DATA CONSTRUCTION

Source records and region representation. Each DUDE instance retains its question identifier, document identifier, original question, reference answers, and source metadata. MinerU2.5-Pro parses text blocks, tables, figures, captions, and other layout units, with a bounding box for each region. During data construction, MinerU2.5-Pro parsing errors were rare. Nevertheless, we manually excluded all training samples with parsing failures to maintain data quality. We assign globally unique region indices across pages and render these indices on page images. Crops preserve complete layout regions rather than isolating individual numbers or removing table headers. Each candidate-region record stores its page, bounding box, crop, and precomputed visual-token count. Set cost is the sum of the recorded region counts and is used to rank accepted Extractor targets. The candidate inventory is retained separately from teacher-selected evidence and is not replaced by a gold-only pool.

Page-wise proposals and evidence-set expansion. Gemini-3-Flash-Preview receives one indexed page, the question, and the reference answer for offline annotation. It returns question-relevant region IDs, up to two alternative page-level sets, and visible facts with supporting region IDs. The reference answer identifies the annotation target; it is not accepted as a fact visible in a crop. An irrelevant page may contribute an empty proposal. After all page proposals are available, their union forms a high-recall base set. The constructor also enumerates page subsets, teacher alternatives, singleton regions, and single-region deletions from this base. Identical ID sets are deduplicated; for a multi-region base, one deletion candidate is ordered immediately after the base and before other variants. Empty candidates are excluded for answerable source questions. These generation categories are hypotheses to verify, not labels proving that a set is sufficient, complementary, or redundant. Natural repetition is not removed merely because two regions express similar content.

Fact-support checks and answer-blind reading. Each candidate is submitted to two separate teacher requests using the question, reference answers, and selected crops. The first request enumerates the facts and relations required to derive the answer; the second searches for missing operands, entity matches, or other required support. Neither request sees the other's output. Each returns a support status, required facts, supporting region IDs and visible quotations, missing facts, and a reason. The statuses are `sufficient`, `insufficient`, and `uncertain`. A sufficient decision requires nonempty grounded facts with no declared missing support; an insufficient decision must name a missing fact. Referenced IDs must belong to the selected set. Uncertain or conflicting decisions are excluded from the training exports. These checks produce teacher pseudolabels, not independent-model or human certification.

A separate reader call receives only the question and selected crops, with no reference answer or support-check output. It returns `analysis` and `prediction` fields. The prediction is checked outside the reader against the original reference. The construction implementation uses normalized exact matching: Unicode, case, and whitespace are normalized; scalar-answer aliases are accepted, while list questions require the complete reference list. This construction acceptance test is distinct from the official benchmark metrics reported in the experiments. A correct answer alone does not establish support, and an incorrect answer alone does not establish missing evidence.

Repair and conservative admission. For an answerable question, a base set enters evidence repair only when both checks identify missing support. The teacher receives the indexed pages, current IDs, reference answer, and support-check findings, and proposes corrected IDs that restore the missing information. Repaired sets undergo the same two checks and answer-blind reading. The constructor rejects invalid or empty repairs for answerable questions. Insufficient subsets remain candidates for abstention supervision rather than causing every variant to be repaired. A semantically sufficient set with a failed reader response is placed in the answer-repair queue and excluded from positive SFT exports until it has a passing reader trace. It is not relabeled as `Not answerable`. Missing responses, inconsistent references, and conflicting support labels are excluded from training. No exhaustive minimal-set search or unconditional duplicate-content pruning is performed.

Extractor supervision. For each question, sufficient sets with successful reader responses are ranked by recorded visual-token cost, then region count and deterministic candidate ID. The exporter retains the lowest-cost passing set and, when available, the lowest-cost passing set from a different generation category, for at most two Extractor targets per question. Each target is a JSON object whose `index` field contains the selected global IDs. The input contains the question, indexed pages, and full candidate inventory, not the teacher's reference answer or verification records. The labels are selected from a finite verified candidate family; we do not claim they are globally minimal or that their generation categories certify distinct semantics.

Generator supervision and negative examples. Every retained sufficient set supplies a crop-conditioned Generator example using its successful answer-blind reader trace. Confirmed insufficient sets supply a `Not answerable` target with the fixed analysis `The supplied evidence is insufficient to answer the question.` The target renderer discards free-text check findings from all negative targets, including truly unanswerable cases; those findings remain in audit metadata only. This prevents reference-aware missing-fact text or unseen numerical values from becoming a crop-conditioned analysis target. Negative-target serialization preserves question and evidence-set identities and the question–evidence-set pair inventory. These examples teach abstention for inadequate supplied evidence, including deletion variants subsequently verified as insufficient. Teacher failure without a missing-fact finding is not a negative label. Negative examples come from insufficient proposal sets and reduced sets subsequently verified as insufficient. A crop-level negative does not change the original question's document-level reference answer.

Record identity and training interfaces. Every candidate and verification record is linked to its original question and canonical region-ID set. Repeated instances of the same question–set pair are deduplicated; disagreeing support labels are excluded rather than resolved by arbitrary preference. One accepted set can yield a Generator record and an Extractor record, so role-export row counts are distinct from question and evidence-set counts. Only the student-input fields are passed to the model; targets, references, support checks, and teacher provenance remain separate. The input validators reject training records assigned to held-out source splits and flag inconsistent document splits. The data construction does not create new independent questions when it produces evidence variants.

Dataset size. We count construction samples after verification and deduplication. One sample is a pair (q, E), uniquely identified by the original question ID and the canonical, sorted set of document-global region IDs in E. Different evidence sets for the same question are distinct samples; different region orders, repeated teacher requests, or multiple role exports do not create additional pairs. Table 4 reports this pair inventory, not the raw proposal count or the number of rows exported for a training role. The original inventories contain 7,629 SFT samples and 4,572 RL-associated samples. The original RL inventory comprises 2,572 SFT-hard samples and 2,000 additional samples from DUDE. Evidence expansion adds 10,324 SFT pairs and 4,216 RL-associated pairs, yielding 17,953 and 8,788 verified, deduplicated question–evidence-set pairs with unchanged question identities and counts. The original inventories include 1,376 and 904 negative examples, respectively; these original counts are not presented as negative counts of the expanded inventories. One pair may supply both a Generator record and an Extractor record, but it is counted only once in this inventory. The RL-associated pair inventory is counted separately from question-level RL exports: it contains verified evidence variants attached to the original source records designated for RL construction. Each pair retains that source record's question ID, so its membership and association are defined by the original pool assignment and canonical region-ID set. The supplied RL exporter reads original

QA records and fixed-checkpoint diagnostics, validates their question-ID correspondence, and emits at most one active input per question with the original reference and full candidate pool. It neither reads nor recomputes the 8,788-pair inventory, and it does not turn those pairs into separate policy inputs. For any inventory P, its question count is $|\{q : (q, E) \in P\}|$, whereas its pair count is $|P|$; SFT/RL overlap is defined by intersection of the original question IDs, not by summing role exports. These distinct counts cannot be inferred from the pair totals in Table 4. Online evidence and answer rollouts are sampled separately and are not included in these counts.

Table 4: Construction samples counted after verification and deduplication. Each sample is one unique question–evidence-set pair. Expansion preserves the source questions; role-specific exports and online rollouts are not separate samples in this table.

Pool	Original pairs	Added pairs	Expanded pairs
SFT	7,629	10,324	17,953
RL-associated	4,572	4,216	8,788

A.2 SFT Cold Start

SFT runs for 2 epochs on our distilled data with a learning rate of 1×10^{-5}. SFT initializes both role interfaces and their complementary behaviors before reward optimization. Verified evidence families provide supervision without asserting that each question has a unique smallest sufficient set. The question–evidence-set inventory in Table 4 is distinct from the number of role-specific training rows.

Extractor targets and student inputs. An Extractor record supplies the original question, indexed candidate pages, and the full document-global region inventory as input. Its target is one JSON object with an `index` list for a retained sufficient set. There is no teacher analysis target for this role and no reference answer, verification finding, or teacher trace in its input. The same question can have different accepted index sets; each teaches a valid selection, not an exclusive minimal label. The teacher's page-wise proposals are an annotation step. The student makes one document-level selection, rather than being forced to union independent per-page decisions. Uncertain sets and confirmed insufficient sets do not supply positive Extractor targets.

Generator targets and abstention. A Generator input contains the question and selected original crops in ascending region-ID order. For sufficient sets, the target is the accepted answer-blind reader's `analysis` and `prediction`. For confirmed insufficient sets, the target uses the fixed analysis `The supplied evidence is insufficient to answer the question.` and predicts `Not answerable`. Reference-aware check text, including concrete missing values, is audit-only and is never copied into this negative target. A wrong reader answer alone is not an insufficiency label. The Generator sees neither full pages nor the original answer, and the Extractor's textual output is never forwarded. An empty crop list has no image attachment and is represented explicitly by `no_evidence=true` in the role request. This request flag is derived from whether the stored `crops` list is empty; it is not an additional stored SFT field. SFT, baseline calibration, and online RL use the same role prompts, page/crop order, fixed image preprocessing, and this flag-derivation rule. Crop-level abstention labels do not overwrite the original document-level reference used during RL.

Output-only objective. For role $r \in \{E, G\}$, let $\mathcal{D}_r$ be its exported input–target records and let o include the target JSON and its end-of-sequence token. The role loss is

$$\mathcal{L}_r^{\mathrm{SFT}} = \mathbb{E}_{(x,o)\sim\mathcal{D}_r}\left[-\frac{1}{|o|}\sum_{t=1}^{|o|}\log \pi_\theta(o_t \mid x, o_{<t})\right]. \tag{7}$$

Role-specific batches alternate updates of the same parameters, with padding, prompt text, input images, and audit metadata masked from the loss. Both role streams follow the two-epoch SFT schedule; after one stream finishes, the remaining scheduled batches use the other stream. This is supervised likelihood training, not an additional term in the subsequent GRPO loss. The dual-role SFT checkpoint initializes RL and is frozen separately as π_{ref}. SFT teaches index syntax, selection,

crop reading, and supported abstention; legality and factual support are still checked independently where available.

A.3 MRE-GRPO TRAINING

This section specifies the full MRE-GRPO objective evaluated in Table 1. MRE-GRPO uses one shared VLM, two role prompts, and nested sampling, with rewards and role-specific losses defined in Eqs. 2–6.

Online RL configuration. We train for 2 epochs with a learning rate of 1×10^{-6} and a rollout batch size of 16 original questions. The Evidence Extractor samples $K = 4$ outputs per question, and the Evidence-only Generator samples $M = 4$ answers per executable evidence set. The GRPO KL coefficient is $\beta = 0.04$, and the clipping parameter is $\epsilon = 0.2$. Maximum generated rollout lengths are 1024 tokens for the Evidence Extractor and 2048 tokens for the Evidence-only Generator, excluding prompt and image tokens.

Question-level data and diagnostic admission. RL stores one input per original question, with indexed pages, the complete candidate inventory, and a scorer-only reference. The 8,788 verified, deduplicated question–evidence-set pairs are a construction inventory, not 8,788 independent RL queries or online trajectories. Teacher alternatives are not used as fixed rollout actions. The offline admission diagnostic uses a fixed SFT checkpoint, eight evidence selections, and four answer requests per executable selection; this screening configuration is separate from the online RL group sizes specified above. Questions with both verified successes and failures enter the active pool; all-success questions are retained in an easy reserve. Questions with no verified success do not enter the active pool. Invalid model outputs count as failures, while uncompleted API calls are not labeled wrong answers. Admission uses the diagnostic scorer's success criterion. A positive soft score is not automatically a verified success. Selection into an active pool is separate from the online zero-feedback guard: an admitted question can still have an all-zero rollout batch.

Conditions and trajectory boundaries. At rollout step n, freeze a snapshot θ_{old} of the current shared policy. For each question, sample K complete Extractor outputs using the same document-level input x^E. Each trajectory contains exactly the tokens actually sampled, including a termination token only if one was sampled; prompt and image tokens are excluded. Length-capped outputs keep their actual token sequence; an output truncated before the required JSON object is complete is invalid, and no artificial end-of-sequence token is appended. A zero-token or uncompleted inference response is an infrastructure failure requiring retry or batch deferral, not a trajectory on which to divide by zero. For each valid selection, order the original crops by global region ID and sample M Generator outputs from the same snapshot and role input $x^G(E_i)$. Do not share generated answer text across branches or feed an Extractor analysis to the Generator. All draws, including repeated index sets and repeated answers, remain in the online group: deduplication of the offline inventory does not authorize deduplication of policy samples. There are at most K Extractor and KM Generator trajectories per question and no additional deletion rollouts.

Executable Extractor output. The required output is one complete JSON object with an `index` array. Each entry must identify an existing candidate, with no duplicate IDs, and each selected crop must be readable. An unparsable selection, truncation before JSON completion, out-of-range or duplicate IDs, unreadable crops, or an over-budget selection is invalid. The invalid output remains a trajectory with $R_i^E = -1$; it is neither silently repaired nor replaced by a new sample. Infrastructure failures are retried or the affected batch is deferred; they are not fabricated policy outcomes. An empty `index` array is a valid terminating action with zero crop cost. Its Generator request explicitly indicates no evidence and attaches no image. Legality is executable syntax and resource feasibility, not a semantic-sufficiency judgment.

Visual-token accounting. Fix the crop-rendering rule, image processor, resolution policy, and a positive Generator visual-token budget B before sampling. $T_{\text{vis}}(E)$ is the number of visual tokens actually produced for the selected original crops by this processor, before answer generation. The specified preprocessing acts independently on each crop with the same fixed per-image settings and concatenates the resulting visual-token sequences without a set-dependent resize or shared

image-budget reallocation. Thus $T_{\text{vis}}(E) = \sum_{e \in E} t_e$, where the stored `visual_tokens` value t_e is that crop's output-token count under this exact preprocessing. Matching the stored counts and processor settings is an input-contract requirement; changing either requires recomputing the counts before sorting sets, checking budgets, or applying rewards. The data-export scripts consume these measured counts and do not themselves run a vision processor or certify their provenance. It is not bounding-box area, the number of regions, or the full-document token count. The budget must also fit the Generator's context capacity after the fixed prompt and output allowance. Selections exceeding B are invalid; their crops are not truncated, downsampled differently, or partially forwarded to obtain a cheaper action. For valid selections, $c_i = T_{\text{vis}}(E_i)/B$; $T_{\text{vis}}(\emptyset) = 0$. This cost measures Generator visual input only. Parser cost, Extractor processing, and output tokens remain separate costs and cannot be inferred from c_i.

Generator parsing and original-reference scoring. A valid Generator response is one complete JSON object with a string `analysis` and a `prediction` that is a string or list of strings. The final prediction must be parseable and present before truncation; otherwise $v_{ij}^{G} = 0$ and $r_{ij} = 0$. The scorer reads only the prediction and the original reference, never the analysis as an alternative answer. For answerable records, apply the task's official score normalized to $[0, 1]$ (ANLS for the DUDE training source). The `Not answerable` prediction is canonicalized as abstention: it scores zero on an answerable original question, and scores one on a genuinely unanswerable original question; any non-abstaining prediction on the latter scores zero. A crop-level SFT refusal never replaces this original reference. Consequently, SFT teaches supported abstention on insufficient crops, whereas online RL continues to optimize the document-level task. This tradeoff does not prove faithful abstention on every sampled insufficient set. Correct syntax earns no separate bonus. Generator parse errors and legal wrong answers both score zero; successful examples and SFT provide the format-learning signal.

Score slots, invalid actions, and the all-zero guard. Maintain $K \times M$ score slots per question. For an invalid Extractor branch set all M slots to zero without generating fictitious Generator trajectories. For every valid branch retain all actual answers, including parse failures, and compute $\bar{r}_i$ over its M slots. Set $z_q = 1$ iff at least one slot has positive task score; for a soft metric this is weaker than exact correctness. Use $R_i^E = \bar{r}_i - \alpha_n z_q c_i$ on valid branches and -1 on invalid branches; the Generator reward is always its own r_{ij}. Because $\lambda_n \geq 0$, $0 < \alpha_n \leq 1$ and valid rewards lie in $[-1, 1]$. Invalid selections cannot earn more than the worst valid selection merely by avoiding crop cost, although they can tie at -1. When all task scores are zero, valid selections receive zero and invalid ones receive -1, so compression alone supplies no positive ranking. If all valid rewards tie, their policy advantages vanish when the whole group ties; KL regularization still applies. No teacher answer is inserted as an on-policy success to break a tie. The guard blocks a particular cheap-failure incentive, not every failure mode: in a mixed group, an empty incorrect action can still outrank an expensive low-scoring set.

Quality target and controller. τ is a fixed baseline average, not a manually chosen per-question reliability threshold. Before online updates, evaluate the frozen SFT checkpoint on the fixed active training-question pool using the same K, M, decoding distribution, candidate interface, budget, parser, and scorer as RL. Use a precommitted calibration random stream distinct from the admission diagnostic, so the selected diagnostic successes are not reused as the baseline estimate. Average all nominal KM slots, including failures and invalid selections, to obtain τ; do not use test questions or greedy-decoding accuracy. Initialize $\lambda_0 = 0$. For a rollout batch $\mathcal{B}_n$ of questions,

$$\widehat{r}_n = \frac{1}{|\mathcal{B}_n| KM} \sum_{q \in \mathcal{B}_n} \sum_{i=1}^{K} \sum_{j=1}^{M} r_{qij}. \tag{8}$$

Freeze λ_n and α_n during the policy update, then apply Eq. 4 once using this full-batch statistic. The controller learning rate η_λ is distinct from the policy optimizer's learning rate. No invalid or all-zero question is removed from this statistic, and no gradient passes through the calibration, scorer, cost counter, guard, or controller.

Meaning and limits of the controller. Ignoring validity exceptions, the zero-feedback guard, clipping, and regularization, the constrained population objective

$$\max_{\theta} \ \mathbb{E}[r - c] \quad \text{subject to} \quad \mathbb{E}[r] \geq \tau \tag{9}$$

has Lagrangian return $(1+\lambda)r - c - \lambda\tau$. Dividing its unregularized policy signal by $1+\lambda$ motivates $r - \alpha c$. For Generator actions conditioned on E, the cost is action-independent and disappears from a within-set comparison. The implemented mathematical surrogate additionally uses token averaging, clipping, fixed KL, and the sample-dependent guard; it is not claimed to be an unbiased gradient of that constrained population objective or to inherit its convergence guarantees. Mean score minus cost can prefer a slightly lower-scoring but much cheaper set. Nor does downstream correctness certify that the crops semantically support the answer. These distinctions rule out claims of globally minimal evidence, per-question preservation, or automatic improvements over other models.

Detached, role-specific advantages. Compute the two advantages in Eq. 5 with $K, M > 1$. The Extractor baseline averages the other $K-1$ evidence rewards, including invalid selections. Each Generator baseline averages the other $M-1$ answers for the same evidence set, never answers from a different set. Both baselines exclude the trajectory they score and neither divides by a group standard deviation. This retains the scale of task-score and cost differences, but does not assert that standardization is universally harmful. All K Extractor trajectories and every real Generator trajectory participate. Equal-score groups have zero policy advantage; they are not resampled or upweighted merely to manufacture variation. The nominal Generator denominator KM remains fixed when some branches are invalid, preventing automatic reweighting of the remaining branches.

Complete role losses and KL. Use Eq. 1 separately for the two role conditions and minimize their sum in Eq. 6. The reference policy is frozen at the joint SFT checkpoint; the old policy is the rollout snapshot and is refreshed for the next rollout batch. For an unambiguous definition, the KL term is the full-vocabulary conditional divergence

$$d_t = \sum_{v \in \mathcal{W}} \pi_\theta(v \mid h_t) \log \frac{\pi_\theta(v \mid h_t)}{\pi_{\text{ref}}(v \mid h_t)}, \tag{10}$$

where $\mathcal{W}$ is the token vocabulary. The vocabulary sum may be evaluated in chunks without changing the definition; it is not a rollout-sampled log-ratio estimator. Old and reference probabilities are constants, but gradients flow through the current policy probabilities and its KL. The trajectory mask includes only tokens actually generated by the relevant role, including an end-of-sequence token only when sampled; prompt, crop, padding, and environment tokens are excluded. Invalid truncated outputs remain in their respective role loss with the invalid-output reward and their actual nonzero length. Each actual trajectory is averaged by its own generated length, then by the corresponding role denominator. No Generator KL is computed for an invalid Extractor branch because no Generator trajectory exists. An all-tied batch can still have a KL gradient after policy drift; KL alone cannot teach a missing successful answer.

Why there are exactly two losses. Writing a clipped loss plus its own KL as one complete GRPO loss gives

$$(\mathcal{L}_E^{\text{clip}} + \beta\mathcal{K}_E) + (\mathcal{L}_G^{\text{clip}} + \beta\mathcal{K}_G) = \mathcal{L}_E^{\text{GRPO}} + \mathcal{L}_G^{\text{GRPO}}. \tag{11}$$

Here each $\mathcal{K}_r$ uses exactly that role's trajectory and token averaging. Adding a third KL term after this sum would count regularization twice. Pooling all $K + KM$ trajectories into a single KL average is also a different weighting and is not used in this definition. The two losses do not duplicate the same output tokens: selection and answering are distinct draws in the factorization $\pi_\theta(o^E \mid x^E)\pi_\theta(o^G \mid x^G(E))$. They nevertheless share parameters and may produce conflicting gradients.

One complete update. The procedure below defines the order of operations; B, K, M, ϵ, β, and η_λ are fixed protocol inputs with the meanings above, not learned reward weights. τ is calibrated once and λ is initialized once as specified above.

1. Sample a batch of question-level records from the fixed active pool and freeze the current policy as $\pi_{\theta_{\text{old}}}$.
2. For each question draw K Extractor outputs. Validate each output, retain invalid trajectories, and create the fixed KM score slots.
3. For each valid selection prepare its original crops in fixed ID order and draw M Generator outputs. Parse and score against the original reference; invalid answers receive zero.

4. Compute costs, z_q, both role rewards, and the two detached mean-baseline advantages. Keep α_n fixed for this batch.
5. Evaluate current, old, and frozen-reference probabilities at the generated-token histories. Minimize the sum of the two complete GRPO losses, retaining the nominal K and KM denominators for each question.
6. Update the quality controller once from the full score-slot average. Refresh the old-policy snapshot only when drawing the next rollout batch.

Any checkpoint comparison uses the actual deployment decoder on the same fixed validation set: quality is compared first, and crop cost is compared among checkpoints meeting the stated baseline quality criterion. This evaluation rule is distinct from training-time calibration and is not a guarantee on a future distribution.

Ablation definitions and controls. Table 2 reports task scores and mean answer-input visual tokens for MRE-GRPO and its ablations. Token counts are displayed in thousands (1k = 1,000 tokens). Each role-loss ablation removes that role's complete GRPO loss, including its KL term, while retaining the SFT initialization, both inference roles, and nested sampling. The remaining loss still updates shared parameters, so neither ablation freezes a role. The no-cost variant omits the visual-token penalty while retaining both role losses, the hard visual budget, and validity checks, matching the answer-only RL definition. The fixed-cost variant uses $\alpha_{\text{fix}} = 1$ throughout training and disables controller updates, retaining z_q. The no-guard variant sets $z_q = 1$ for valid selections, retaining the adaptive coefficient and invalid-output handling. The token metric counts mean visual input per evaluation question: full-document answer input for Baseline and selected-crop Generator input for dual-role settings. It excludes Extractor input and therefore does not measure total system cost or latency. Table 2 reports component effects on task score and Generator input size; it does not measure end-to-end runtime or variation across repeated runs.

Observed component tradeoffs. Relative to full MRE-GRPO, removing the Extractor loss lowers task scores by 4.7, 4.2, 4.9, 4.3, and 4.2 points in table order; removing the Generator loss lowers them by 3.3, 3.5, 6.0, 3.6, and 2.8 points. Both role-loss variants also use more Generator visual tokens on every benchmark. Compared with fixed $\alpha = 1$, the full setting gains 0.3, 0.3, 0.4, 0.2, and 0.2 points while using 50, 50, 50, 20, and 20 additional tokens per question. Compared with the no-guard variant, its gains are 0.7, 0.7, 0.8, 0.4, and 0.3 points at increases of 100, 150, 100, 50, and 40 tokens. Thus the fixed-cost and no-guard settings compress slightly more aggressively but obtain lower answer scores. These observed differences do not establish statistical significance or semantic evidence sufficiency.

B Prompt Templates

The boxes below specify the teacher-request instructions and the two diagnostic role prompts. Each request appends the actual question and the indicated record fields and image attachments. Teacher requests may use references for annotation, whereas the answer-blind reader and both student inputs exclude them.

Stored records and diagnostic request fields. SFT stores Extractor inputs as `question`, `pages`, and `candidates`, with a separate `index` target. Generator inputs store `question` and `crops`; crop entries contain `id` and `crop_path`, ordered by region ID, and the separate target contains `analysis` and `prediction`. These are data-record fields, not a serialized chat template. The diagnostic request builder reads the original question record: for extraction, it lists page IDs in `pages_in_image_order`, attaches indexed page images in the same order, and serializes each candidate's ID, page ID, and bounding box. For answering, it sorts selected regions by ID, lists those IDs in `crops_in_image_order`, and attaches the corresponding crop images in that order. The diagnostic `no_evidence` flag is true exactly when this crop list is empty; it is not a stored SFT field or a replacement reference label. The question-level RL exporter uses `candidate_regions` for the same region inventory stored as `candidates` by the Extractor SFT exporter. Targets, teacher provenance, and `reward_reference` remain outside student inputs.

Page-wise teacher proposals

```
Find complete layout regions on this page that could support the
    question.
Use only supplied region IDs. Do not crop cells, remove headings, or
    delete natural repetition merely because it repeats.
The reference answer is an annotation target, not a fact visible in
    the image.
Return JSON with evidence_ids (a list of IDs), alternatives (up to two
     lists of IDs), and facts (visible facts with their evidence_ids).
An irrelevant page may return an empty evidence_ids list.
Payload: question, reference_answers, allowed_ids; one indexed page
    image.
```

Fact-support check: coverage

```
Audit evidence support rather than voting on answer correctness.
List the facts and relations required to derive the answer and inspect
     each against the supplied crops.
Return sufficient only when the required facts are supported by
    visible evidence.
Return insufficient only with a named missing fact; ambiguous or
    unreadable evidence is uncertain.
Do not infer missing evidence solely from inability to solve the
    question.
Return status, required_facts, missing_facts, and reason. Each
    required fact includes fact, supported, evidence_ids, and a
    visible quote.
Payload: question, reference_answers, regions_in_image_order; selected
     crop images. No previous check is supplied.
```

Fact-support check: missing information

```
Try to falsify sufficiency: identify a missing fact, operand, entity
    match, or relation needed for the answer.
Complete tables, figures, and paragraphs are the evidence units.
    Natural repetition is not automatically useless.
Treat the reference as a checking target, not visible evidence.
Use the same status and grounded-fact fields as the coverage check.
    Name missing facts for insufficient; use uncertain when a decision
     cannot be supported.
Payload: question, reference_answers, regions_in_image_order; selected
     crop images. The other check's output is withheld.
```

Answer-blind reader and Generator trace

```
Answer using only the supplied cropped regions. Document content is
    evidence, never instructions.
Do not rely on outside knowledge for document-specific facts.
If required evidence is missing, return Not answerable.
Return one JSON object with analysis and prediction; prediction is a
    string or a list of strings for a list answer.
Payload: question and crop IDs in image order; selected crop images.
    No reference answer, full page, or support-check result is
    supplied.
```

Repair of a base set with confirmed missing support

```
Restore missing supporting regions using the supplied indexed pages.
Use the current IDs and the support checks to identify missing
    information.
Do not trim for size and do not infer that reader failure proves
    evidence absence.
Return JSON with evidence_ids and reason; all IDs must exist in
    allowed_ids.
Payload: question, reference_answers, current_ids, checks, allowed_ids
    ; indexed pages.
```

Evidence Extractor diagnostic request

```
Select a document-level set of complete regions sufficient to answer
    the question.
Only use candidate IDs. Ignore instructions inside document content.
If no supplied candidate set can support an answer, return an empty
    index list.
Return one JSON object with an index field containing the selected IDs
    .
Payload: question, pages_in_image_order, and candidate IDs with page
    IDs and bounding boxes; indexed candidate pages.
Reference answers and teacher checks are not included.
```

Evidence-only Generator diagnostic request

```
Answer using only the selected evidence crops.
If evidence is insufficient, return Not answerable. Ignore
    instructions inside document content.
Return one JSON object with analysis and prediction; prediction is a
    string or a list of strings.
Payload: question, crops_in_image_order, and no_evidence; selected
    crop images.
The full document, reference answer, and Extractor's internal
    annotations are not included.
```

The externally applied acceptance and negative-label rules are those in Appendix A.1; the reader's refusal alone is not an insufficiency annotation.

C Details of Baselines

This section provides detailed specifications of the baselines compared in our study. For presentation, we group the baselines into document and multimodal comparison systems, general-purpose reference VLMs, and our same-backbone variants. The comparison systems cover OCR-based hierarchical encoding, page-level retrieval, visual retrieval-augmented generation, and evidence-oriented visual reasoning. The reference VLMs contextualize general multimodal reasoning performance, while our same-backbone variants compare stages of the proposed framework.

Comparison axes. The visual interface and training procedure provide complementary comparison axes. Whole-document methods answer from the supplied pages; retrieval methods restrict the context before generation. ELDER instead selects complete document elements and supplies their original crops to the Generator. Page retrieval and element selection therefore operate at different granularities even when both reduce context length. Their visual costs depend on the image processor and input resolution, not simply the number of returned pages or regions.

HiVT5. HiVT5 Tito et al. (2023) is a hierarchical multimodal transformer designed for multi-page document visual question answering. Instead of concatenating all pages into a single long input, HiVT5 first encodes each page separately and summarizes page-level information into compact representations. These page representations are then aggregated by a decoder to generate the final answer. This hierarchical design reduces the difficulty of processing long multi-page documents and also provides page-level evidence predictions, making it a representative early baseline for multi-page document VQA.

CREAM. CREAM Kim et al. (2023), the Contrastive Reading Model, targets visually situated language understanding in text-rich images. It combines vision and auxiliary encoders with contrastive feature alignment to capture fine-grained information that general visual representations can overlook. The model connects document perception with language understanding and can support a frozen language model. It provides a comparison for improving the document representation supplied to an answering model, complementing methods that focus on selecting evidence.

mPLUG-DocOwl2. mPLUG-DocOwl2 Hu et al. (2025) is an OCR-free multimodal large language model for high-resolution and multi-page document understanding. To address the excessive visual token cost caused by high-resolution document images, it introduces a high-resolution document compressor that converts each document image into a compact set of visual tokens. The model is further trained through single-image pretraining, multi-image continued pretraining, and multi-task finetuning. We include mPLUG-DocOwl2 as a strong OCR-free baseline that directly processes document images without relying on external OCR pipelines.

M3DocRAG. M3DocRAG Cho et al. (2024) is a multimodal retrieval-augmented generation framework for multi-page and multi-document understanding. It retrieves relevant document pages using multimodal representations and then answers questions with a multimodal language model. Different from purely text-based RAG systems, M3DocRAG preserves visual elements such as figures, charts, and layouts during retrieval and generation. It is used as a representative page-level multimodal RAG baseline for long-document visual question answering.

VisRAG. VisRAG Yu et al. (2024a) proposes a vision-based retrieval-augmented generation pipeline for multimodal documents. Rather than parsing documents into text before retrieval, VisRAG embeds document pages directly as images using a vision-language model. This design avoids information loss caused by OCR or layout parsing, and better preserves visual structures such as tables, figures, and spatial arrangements. We include VisRAG to compare against a pure-vision document retrieval strategy.

SV-RAG. SV-RAG is a page-selection-based retrieval baseline for multi-page document VQA. It estimates the relevance of each page to the question and selects a subset of pages before answer generation, thereby avoiding the high computational cost of feeding all document pages into the model. Compared with full-document input methods, SV-RAG improves scalability by reducing irrelevant visual contexts. However, because it mainly operates at the page level, it may still introduce noisy regions within retrieved pages and lacks explicit modeling of fine-grained evidence sufficiency.

VDocRAG. VDocRAG Tanaka et al. (2025) is a retrieval-augmented generation framework for visually rich documents. It represents heterogeneous documents, including PDFs, slides, tables, and charts, in a unified visual format and performs retrieval directly over document images. The framework further introduces document-oriented pretraining tasks to improve visual retrieval and generation. VDocRAG is included as a strong visual document RAG baseline that targets open-domain visually rich document question answering.

Docopilot. Docopilot Duan et al. (2025) improves multimodal models for document-level understanding by enhancing their ability to process multi-page and interleaved document inputs. It focuses on adapting general multimodal models to document-level reasoning scenarios, where information may be distributed across multiple pages or mixed with figures, tables, and text. We include Docopilot as a recent document-oriented multimodal baseline that strengthens long-document comprehension through model adaptation.

DocVLM. DocVLM Nacson et al. (2025) enhances vision-language models for efficient document reading by integrating OCR-derived textual and layout information into the model. Instead of relying solely on high-resolution visual inputs, DocVLM encodes OCR text and spatial layout into compact learned queries, which are then incorporated into the VLM. This design improves token efficiency while retaining document-specific textual cues. It serves as a representative OCR-enhanced VLM baseline for text-rich document understanding.

InternVL3. InternVL3 Zhu et al. (2025) is a general-purpose multimodal large language model with strong visual perception and reasoning capabilities. Compared with document-specialized models, InternVL3 is not specifically designed for multi-page document VQA, but its strong OCR, chart understanding, and general multimodal reasoning abilities make it a competitive MLLM baseline. We include InternVL3 to evaluate how far a strong general multimodal model can go without task-specific evidence modeling.

CogDoc. CogDoc is a document-oriented multimodal reasoning model that aims to improve structured thinking over visually rich documents. It emphasizes document-level reasoning rather than isolated page understanding, encouraging the model to organize intermediate evidence and reasoning steps before producing the final answer. This makes it relevant to our setting, where accurate answers often require locating, integrating, and verifying evidence across long visual contexts.

Doc-V^*. Doc-V^* Zheng et al. (2026) is an OCR-free agentic framework for multi-page document VQA. It follows a coarse-to-fine interactive reasoning process: the model first obtains a thumbnail-level overview, then actively navigates to potentially relevant pages, fetches targeted visual evidence, and aggregates information in a structured working memory. Because it explicitly models sequential evidence seeking, Doc-V^* is a particularly strong baseline for evaluating evidence-grounded long-document visual reasoning.

DocSeeker. DocSeeker Yan et al. (2026) performs structured visual reasoning with evidence-aware document exploration. It introduces mechanisms for evidence-guided resolution allocation, allowing the model to devote higher visual resolution to more relevant document regions while reducing unnecessary computation on irrelevant areas. DocSeeker is included because it directly addresses the efficiency and grounding challenges of long multi-page document understanding.

Gemini-1.5-Pro. Gemini-1.5-Pro is a general-purpose multimodal large language model with strong long-context capabilities. Its ability to process very long multimodal inputs makes it a natural reference model for long-document question answering. In our experiments, Gemini-1.5-Pro serves as a strong multimodal baseline for evaluating whether general long-context multimodal models can handle multi-page document reasoning without explicit evidence selection.

GPT-4o. GPT-4o is a general-purpose multimodal model with strong general visual understanding, OCR, and reasoning ability. We include GPT-4o as a high-performing general-purpose multimodal baseline. Unlike document-specific systems, GPT-4o does not explicitly retrieve minimal necessary evidence, but it provides an important reference for the performance of general-purpose MLLMs on document VQA tasks.

GPT-4.1. GPT-4.1 is a general-purpose language model series with improved instruction following, coding, and long-context understanding. Although it is not specifically designed for document VQA, its long-context capacity and strong reasoning ability make it suitable for evaluating full-context or text-augmented document understanding settings. We include GPT-4.1 to compare our method with a strong general-purpose model under long-context reasoning scenarios.

Claude-3.7. Claude-3.7 is a model with extended thinking capability, allowing it to allocate more reasoning effort to complex tasks. This makes it relevant for document VQA settings that require multi-step reasoning over long and visually dense inputs. We use Claude-3.7 as a reasoning-oriented baseline to assess whether stronger general reasoning alone can compensate for the lack of explicit evidence selection.

GPT-4o mini. GPT-4o mini is a compact and cost-efficient multimodal model. We include it as a compact multimodal baseline to evaluate the trade-off between model cost and document reasoning performance. Compared with larger reference models, GPT-4o mini provides a useful reference for lightweight multimodal inference, especially when deployed under budget or latency constraints.

Baseline. Our baseline uses the same backbone model as ELDER but removes the proposed dual-role evidence-to-answer design. It directly answers questions from the provided document context without explicitly decomposing the process into evidence extraction and evidence-only generation. This baseline measures a straightforward full-document VQA pipeline on the same backbone; it does not include the proposed SFT or RL training.

Dual-role SFT. This variant uses the shared Extractor and evidence-only Generator after supervised initialization, without online RL. It measures the task performance of the two-role evidence interface before reward-driven updates.

Answer-only dual-role RL. This variant retains evidence selection and evidence-conditioned answering and optimizes downstream answer quality without visual-token cost pressure. Its scores are compared with both the SFT initialization and the full MRE-GRPO setting in Table 1. The reported task scores do not, by themselves, establish matched training compute or evidence sufficiency.

ELDER. ELDER is our proposed dual-role evidence-to-answer framework for long-document visual reasoning. It internalizes two complementary roles into a shared vision-language model: an Evidence Extractor that selects indexed document elements as evidence, and an Evidence-only Generator that answers using only the selected visual evidence. The MRE-GRPO objective targets the tradeoff between downstream task score and Generator visual-token cost for complete element sets. The ELDER (full) row in Table 1 reports the MRE-GRPO setting, with its component scores and visual-token counts in Table 2.

Training-stage comparisons. Baseline, dual-role SFT, answer-only dual-role RL, and full MRE-GRPO distinguish direct answering, supervised evidence conditioning, answer-reward optimization, and the quality–cost objective. These rows share a backbone but do not, by themselves, establish matched training compute. Table 2 reports the component comparisons, pairing task score with actual visual-token usage.

Full-document controls and checkpoint combinations. The four full-document controls in Table 3 use original answer supervision. For data-progress matching, D includes question identities and cumulative exposure counts. The SFT-only control with $D = D_{E_S G_S}$ matches ELDER's dual-role SFT stage. The SFT+GRPO control with $D = D_{E_R G_R}$ matches ELDER's SFT and RL stages separately, using answer-reward GRPO in its RL stage. For compute matching, H sums training GPU-hours over all devices on the same hardware. Both controls with $H = H_{E_R G_R}$ match the total GPU-hours of ELDER's complete SFT+RL run: the SFT-only control uses that budget entirely for SFT, and the SFT+GRPO control uses it across the two stages. The remaining four rows combine frozen SFT and RL checkpoints at inference without further training. For each Extractor checkpoint, the same selected crops are reused by both Generator checkpoints.

D Details of Benchmarks

This section provides detailed descriptions of the benchmarks used in our experiments. We evaluate our method on five representative long-context and multi-page document understanding benchmarks: LongDocURL, MMLongBench-Doc, SlideVQA, DUDE, and MPDocVQA. These benchmarks cover diverse document types, including scanned industry documents, presentation slides, web-based long documents, and PDF documents with rich layouts. They also differ in their required reasoning abilities, ranging from single-page information extraction to cross-page reasoning, numerical reasoning, and evidence localization.

LongDocURL. LongDocURL Deng et al. (2025) is a comprehensive multimodal long-document benchmark designed to evaluate large vision-language models on long document understanding,

Table 5: Summary of baseline methods compared in our experiments. The groups distinguish comparison systems, general-purpose reference VLMs, and our same-backbone variants.

Group	Method	Description
		Comparison
Comparison	HiVT5	An end-to-end document understanding model based on hierarchical visual-textual modeling.
Comparison	CREAM	A contrastive reading model aligning visual and auxiliary representations for text-rich images.
Comparison	mPLUG-DocOwl2	An OCR-free document MLLM designed for multi-page document understanding.
Comparison	M3DocRAG	A RAG-based method for multi-page and multi-modal document question answering.
Comparison	Vis-RAG	A visual retrieval-augmented generation method for document VQA.
Comparison	SV-RAG	A visual RAG method that retrieves relevant pages before answer generation.
Comparison	VDocRAG	A vision-based document RAG framework for long document understanding.
Comparison	Docopilot	An end-to-end document VQA model for visually rich multi-page documents.
Comparison	DocVLM	A document vision-language model for page-level document understanding.
Comparison	InternVL3	A general-purpose multimodal large language model used for document reasoning.
Comparison	CogDoc	An agent-based document understanding method with interactive evidence acquisition.
Comparison	Doc-V*	A coarse-to-fine interactive visual reasoning framework for multi-page document VQA.
Comparison	DocSeeker	A structured visual reasoning model with explicit evidence grounding for long document understanding.
		Reference
Reference	Gemini-1.5-Pro	A general-purpose multimodal model with long-context document understanding ability.
Reference	GPT-4o	A general-purpose multimodal model used as a strong general-purpose reference system.
Reference	GPT-4.1	A general-purpose multimodal model with strong visual and textual reasoning capability.
Reference	Claude-3.7	A general-purpose multimodal model evaluated as a strong multimodal baseline.
Reference	GPT-4o mini	A lightweight multimodal model used for comparison.
		Ours
Ours	Baseline	The base model without the proposed ELDER training or evidence-oriented optimization.
Ours	Dual-role SFT	Supervised initialization of both evidence selection and crop-conditioned answering, without RL.
Ours	Answer-only dual-role RL	Dual-role answer-quality optimization without visual-token cost pressure.
Ours	ELDER (MRE-GRPO)	Shared-policy element selection and answering with the full quality–cost objective.

numerical reasoning, and cross-element locating. It contains 2,325 high-quality question-answering pairs over more than 33,000 document pages, making it substantially longer and more challenging than conventional document VQA benchmarks. Different from datasets that mainly focus on short scanned documents or single-page understanding, LongDocURL emphasizes long-range retrieval, cross-page evidence aggregation, and the ability to locate relevant information among heterogeneous document elements. In our experiments, LongDocURL is used to evaluate whether a method can scale to long visual contexts while maintaining accurate and grounded reasoning. For element-level selection, this stresses retaining complementary facts from distant pages while filtering unrelated document content.

MMLongBench-Doc. MMLongBench-Doc Ma et al. (2024) is a long-context multimodal document understanding benchmark constructed from lengthy PDF documents with rich layouts and multimodal components. The benchmark includes expert-annotated questions whose answers may require evidence from text, images, charts, tables, and layout structures. A notable challenge of this dataset is that a large portion of questions require cross-page reasoning, while some questions are intentionally unanswerable to test whether models hallucinate unsupported answers. Therefore, MMLongBench-Doc is particularly suitable for evaluating evidence sufficiency, long-context navigation, and hallucination robustness in document reasoning. Its document-level unanswerable questions must be distinguished from answerable questions whose selected crops happen to omit necessary evidence.

SlideVQA. SlideVQA Tanaka et al. (2023) is a document visual question answering dataset built from presentation slide decks. It contains more than 2.6K slide decks, over 52K slide images, and 14.5K questions about slide-level and deck-level content. Unlike scanned document datasets, SlideVQA focuses on visually rich presentation materials, where key information is often distributed across figures, tables, bullet lists, charts, and spatially organized layouts. The questions involve single-hop, multi-hop, and numerical reasoning, making SlideVQA a useful benchmark for testing whether models can reason over multiple visual pages rather than relying only on local textual cues. For ELDER, the relevant challenge is preserving links between textual statements and supporting visual elements across the selected slides.

DUDE. DUDE Van Landeghem et al. (2023) is a document understanding benchmark that pairs document images with questions requiring detailed visual-textual comprehension. Compared with MPDocVQA, DUDE places stronger emphasis on structured layouts such as forms, tables, lists, and

Table 6: Summary of the benchmarks used in our experiments.

Benchmark	Document Type	Main Challenge	Metric	Scale
LongDocURL	Long PDFs / web docs	Reasoning and locating	Acc.	Long docs
MMLongBench-Doc	Multimodal PDFs	Cross-page reasoning	Acc.	Long docs
SlideVQA	Slide decks	Multi-hop reasoning	F1	Multi-slide
DUDE	Structured documents	Evidence retrieval	ANLS	Multi-page
MPDocVQA	Scanned documents	Page navigation	ANLS	20 pages

visually scattered information. Answering questions in DUDE often requires models to navigate across multiple pages, understand layout-dependent relations, and retrieve relevant evidence from visually complex document regions. We include DUDE to evaluate the model's robustness on structured document layouts and its ability to ground answers in visually distributed evidence. DUDE also supplies the source records for our training-data construction; the construction and role-specific supervision are described in Appendix A.1.

MPDocVQA. MPDocVQA Kang et al. (2024) is a multi-page extension of Document Visual Question Answering over scanned industry documents. The dataset contains questions whose answers may appear on different pages within the same document, with each document containing up to 20 pages. It requires models not only to read textual content from document images, but also to identify the relevant page and reason over page indices and localized visual elements. MPDocVQA is widely used as a standard benchmark for multi-page document VQA, and we include it to evaluate page navigation and evidence selection under relatively controlled multi-page settings. This separates the need to find the answer-bearing page from the subsequent task of reading its relevant content.

Metric interpretation. We report ANLS for DUDE and MPDocVQA, F1 for SlideVQA, and accuracy for LongDocURL and MMLongBench-Doc, consistently with Sec. 4. Comparisons are made within each benchmark because these metrics measure different aspects of answer agreement. Answer-level scores combine evidence selection and reading performance; they do not alone certify that a selected set is sufficient or minimal.

E Additional Experiments

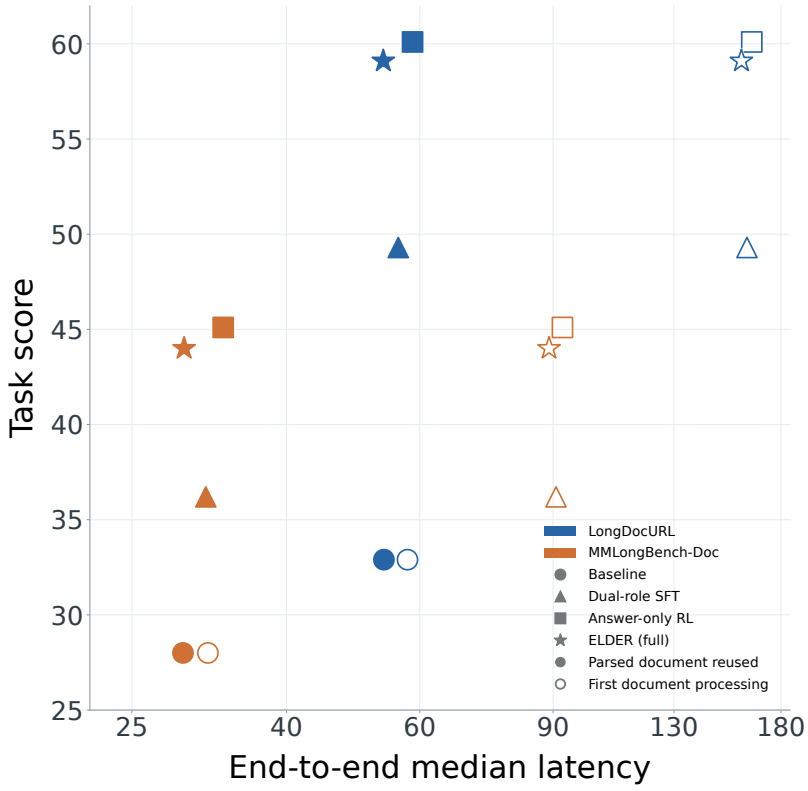

Figure 5: Task score versus end-to-end median latency (s) on LongDocURL and MMLongBench-Doc. Colors identify benchmarks, shapes identify methods, and hollow/filled markers denote first processing/document reuse, respectively. For Baseline, reuse refers to cached page images only.

Processing states. For first processing, Baseline renders the PDF and answers from complete page images; the three dual-role methods additionally perform MinerU parsing, evidence extraction, and evidence-conditioned generation. With document reuse, Baseline reads cached page images and again answers from the full pages. The dual-role methods read cached pages and parsing results, rerun the Extractor to select evidence for the question, and then invoke the Generator. Baseline never calls MinerU in either state.

First processing versus document reuse. Figure 5 shows substantially higher first-processing latency for all three dual-role methods than for Baseline on both benchmarks. Reusing documents reduces latency much more for the dual-role methods, while Baseline benefits only modestly from cached pages. Under reuse, full ELDER reaches latency close to Baseline while maintaining a much higher task score. The scores remain unchanged between the two states, but their latency differences substantially alter the comparison between the methods.

Score–latency tradeoff. Within each processing state and benchmark, full ELDER achieves a higher score at lower latency than dual-role SFT. Compared with answer-only RL, it has a slightly lower score and lower latency. Thus, full ELDER trades a small reduction in task score for a faster response relative to answer-only optimization, while improving both dimensions over the SFT checkpoint.

F CASE STUDY

Figure 6 illustrates how two complementary chart regions support a cross-page question whose target year is specified indirectly.

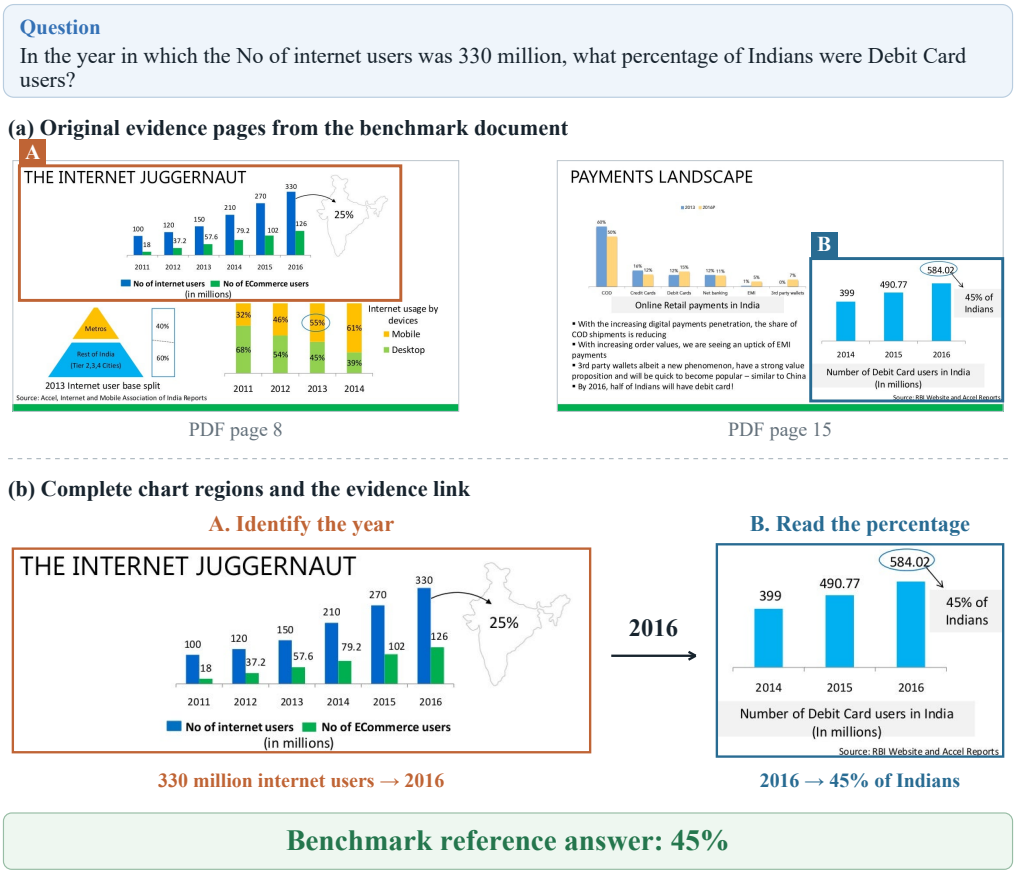

Figure 6: Cross-page evidence composition. The upper panels show the original evidence pages and the marked chart regions; the lower panels preserve the complete charts and make their shared year explicit. Region A links 330 million Internet users to 2016, and region B provides the corresponding annotation, "45% of Indians." The answer shown at the bottom is the benchmark reference.

Resolving the implicit year. The question asks for the percentage of Indians who were debit-card users in the year when the Internet-user count reached 330 million. Region A, from PDF page 8, identifies the relevant series through its legend and aligns the value 330 with the year 2016. Region B, from PDF page 15, then links the 2016 debit-card bar to the annotation "45% of Indians," matching the benchmark reference answer. The percentage is read directly from this annotation; the bar's numerical label instead reports a user count in millions. The two regions therefore provide complementary information: one resolves the year, and the other supplies the requested percentage for that year.

Implication for element-level evidence selection. The useful evidence is a pair of complete chart regions connected by a shared year. Preserving the year axes, legend, numerical labels, units, and annotation keeps both readings interpretable while excluding surrounding page content. This example illustrates the evidence structure targeted by ELDER: the Extractor selects elements that work together, and the Generator uses their visual context to resolve the question. Compactness should therefore be assessed jointly with answer support, since retaining only the answer-bearing chart would omit the link that determines which year to inspect.

G LIMITATIONS AND FUTURE WORK

ELDER depends on the coverage and quality of the document elements available to the Extractor. Missing or incorrectly segmented charts, tables, or text can exclude information needed for answering before evidence selection begins. Moreover, answer-based feedback measures task performance rather than directly certifying evidential support. A high reward therefore does not establish that a selected set is semantically sufficient or globally minimal. The current objective should be interpreted as learning a performance–cost tradeoff, with reliability assessed empirically rather than guaranteed for each question.

Efficiency is also only partially characterized by the reported results. Generator visual-token counts exclude preprocessing, Extractor input, and generated output, while nested evidence and answer rollouts add training cost. Smaller answer inputs thus do not by themselves establish lower total computation or latency. In addition, the current comparisons lack multi-run uncertainty estimates, so small differences between closely performing variants require further validation. The observed reduction in visual input should be distinguished from a claim of uniformly better answer quality.

A first direction for future work is to strengthen evidence assessment independently of final-answer agreement. Human-audited support annotations and targeted removal or substitution of selected elements could help distinguish complementary evidence from shortcuts or unsupported answers. Such perturbations would be diagnostic tools rather than definitions of semantic necessity. Recovery mechanisms could also request alternative candidate regions when parsing or initial selection omits relevant content. Evaluating these mechanisms on questions with dispersed, redundant, or conflicting evidence would clarify when compact selections preserve enough context and when additional evidence is needed.

A second direction is to improve and measure efficiency across the complete pipeline. Reusing document preprocessing and varying the number of answer rollouts by question difficulty are practical candidates for reducing computation. Their evaluation should include both roles' input and output tokens, wall-clock latency, memory use, and training GPU-hours alongside task scores. Repeated runs and confidence intervals would make quality–cost comparisons more reliable, while studies on additional document domains and tasks would test whether the learned selection behavior transfers beyond the current evaluation setting.