

EviDocVLM: Evidence-Aligned Reinforcement Learning for Multi-Page Document Question Answering

Anran Liu Zongsheng Cao* Jun Xie Feng Chen Zigan Wang*

The University of Queensland

agiczsr@gmail.com, zigan.wang@uq.edu.au

Abstract

Models can answer questions from retrieved document pages while producing citations that are valid but irrelevant to the answer. This often happens when citation quality and answer correctness are optimized as separate signals, so the model can learn to cite text without relying on it. We introduce **EviDocVLM**, a training method for multi-page document question answering that links selected support to answer behavior. A structured source map provides citation addresses for page content. EviDocVLM first uses identifier validity and quotation fidelity to derive a continuous coefficient for the trajectory-level answer reward, so traceable support increases credit without making every citation error a hard failure. It then applies Evidence-Aligned GRPO, which reuses full-context rollout log-probabilities as a frozen target for one training-only evidence-view pass. The resulting sampled-token alignment encourages cited spans to support the answer without an external grader, extra rollout, or inference-time module. Experiments on document parsing and multi-page document QA benchmarks show stronger answer performance and consistent gains from the evidence-conditioned objective under complex visual layouts.

1 Introduction

Retrieval-augmented vision-language models (VLMs) (Wang et al., 2024a; Cao et al., 2026b) increasingly answer questions directly from document page images. Compared with OCR-only pipelines, keeping the visual pages preserves layout, tables, figures, and other non-linear structures that are often needed for correct answers (Yu et al., 2025; Faysse et al., 2024; Cao et al., 2026a). However, a correct answer alone gives little evidence about how the model reached it. This is especially important in multi-page settings, where retrieved pages may contain similar entities,

repeated values, or distracting content. A reliable document QA system should therefore expose concise source evidence and make the answer depend on that evidence.

Prior work has improved either document perception or evidence-aware reasoning. End-to-end parsers transform page images into structured text, and Infinity-Parser (Wang et al., 2025a) further improves layout recovery through reinforcement learning with edit-distance, element-count, and reading-order rewards. This line of work primarily evaluates whether a page is parsed correctly; it does not optimize how a QA model uses selected content from that page. Visual RAG methods take the complementary route of keeping retrieved images for multimodal reasoning. More closely related, UniDocVLM (Cao et al., 2026c) generates document units and evidence before an answer, and uses token-level reward routing to distribute separate parsing, evidence, and answer signals. These approaches establish useful mechanisms for producing and checking citations, but they do not directly test whether the answer distribution depends on the cited content.

We call this missing connection the *evidence utilization gap*. Suppose a model returns the correct answer and also quotes an unrelated but accurately transcribed page element. The citation passes identifier and quotation-to-source checks, and the answer independently passes exact-match evaluation. A training objective that adds the two scores can then assign the trajectory a high reward even though the citation was not used. This failure is difficult to observe from final accuracy or citation validity alone. At the opposite extreme, forcing the answer to depend exclusively on the cited text is also undesirable: errors in parsing or evidence selection would discard information that remains available in the original image. The central question is not only how to generate valid evidence, but also how to encourage the model to use it without creating a

*Corresponding authors.

brittle evidence bottleneck.

To address this problem, we present **EviDocVLM**, an evidence-aware training method for multi-page document QA. It assumes an evidence-producing decoder with three fields: a source map, selected support, and an answer. The source map assigns page-qualified handles to recognized content, while the support field records a small set of [page:unit] “quoted span” pairs. We use this representation to address page content and check citations, not as a new parser. It permits deterministic traceability checks and places support in the autoregressive history without removing the original page embeddings from the deployed answer path.

EviDocVLM connects support and answers in two ways. First, we introduce a *reliability-conditioned trajectory return*. Identifier validity and quotation fidelity form a citation-integrity score that scales answer credit instead of serving as a separate reward channel. A correct answer with traceable support receives full credit; a correct answer with weak support retains a reduced reward; and a citation cannot compensate for an incorrect answer. Unlike token-scope reward allocation, this produces one group-relative advantage for the complete sampled trajectory.

Second, we propose *Evidence-Aligned GRPO* (EA-GRPO) to optimize support sufficiency beyond surface validity. Full-context answer-token log-probabilities are retained from the frozen rollout policy. One additional teacher-forced pass then evaluates the same sampled answer from only the question and selected support. Weighted by answer quality and citation reliability, a sampled-token Monte Carlo objective aligns the evidence view with the rollout teacher. This rewards support that carries answer information, without storing full-vocabulary teacher logits or drawing another rollout. The auxiliary view is discarded after training, so inference cost is unchanged.

The two mechanisms operate at different granularities. Reliability conditioning changes the relative return of sampled trajectories, whereas alignment acts on sampled answer-token likelihoods. Their combination separates answer correctness, citation integrity, and support sufficiency. It requires neither an external reward model nor an evidence-only inference stage. Moreover, we evaluate EviDocVLM on document parsing and multi-page document QA tasks covering visually dense graphics and multi-page business documents.

In summary, our contributions are three-fold.

- We introduce a reliability-conditioned trajectory return that uses citation integrity as a soft coefficient on answer correctness, rather than assigning heterogeneous rewards to separate output-token scopes.
- We develop EA-GRPO, which augments group-relative policy optimization with a reliability-weighted, sampled-token alignment between a frozen full-context rollout and a training-only evidence view.
- We conduct extensive experiments on document parsing and multi-page document QA benchmarks, showing that EviDocVLM improves answer correctness and citation traceability without increasing inference cost.

2 Related Work

2.1 From Page Images to Addressable Context

Document QA depends on how page content is exposed to the answering model. Traditional systems first convert images into text through modular toolchains such as MinerU (Wang et al., 2024a), Marker (Paruchuri, 2024), and Docling (Livathinos et al., 2025). These systems combine region detection, recognition, and structural recovery, making intermediate results inspectable but allowing early errors to propagate. Generative parsers including Nougat (Blecher et al., 2024), Kosmos-2.5 (Lv et al., 2024), and mPLUG-DocOwl (Hu et al., 2024) instead produce structured sequences directly from page pixels. Infinity-Parser (Wang et al., 2025a) further applies reinforcement learning to this mapping, using content, element-count, and reading-order feedback to improve full-page reconstruction. Its target is faithful page reconstruction, while our focus is how selected evidence is used in downstream answering.

An alternative is to retain the page image rather than make OCR text the sole context. VisRAG (Yu et al., 2025) and ColPali (Faysse et al., 2024) retrieve visually encoded pages, preserving graphical and spatial cues that may disappear during text conversion (Zhang et al., 2024). EviDocVLM keeps the page images and uses a lightweight source map to provide stable citation addresses, so the source map does not become the main information channel.

2.2 Citation Generation and Evidence Attribution

Textual retrieval-augmented generation associates model outputs with retrieved passages and can reduce unsupported generation (Lewis et al., 2020; Asai et al., 2024). In multimodal settings, however, a retrieved page is too coarse to identify which table entry, caption, or text region supports an answer. Recent systems therefore expose finer-grained evidence or learn additional information-seeking behavior. R1-Router (Peng et al., 2025) and MMSearch-R1 (Wu et al., 2025) optimize retrieval decisions, VRAG-RL (Wang et al., 2025c) learns visual operations such as region selection, and ViDoRAG (Wang et al., 2025b) separates perception and reasoning among multiple agents. These approaches improve access to relevant content, but retrieval or localization alone does not show that the final prediction uses the reported evidence. Structured evidence generators such as UniDocVLM (Cao et al., 2026c) produce addressable page units and allocate parsing, citation, and answer rewards to different output-token scopes.

EviDocVLM instead asks whether a traceable citation contains the information needed for the answer. It uses structured output as a source map, computes a single reliability-conditioned return for each trajectory, and evaluates support sufficiency through paired answer-token likelihoods. The goal is to train support that is useful for answering, not to propose another document serialization scheme.

2.3 Optimizing Evidence Use

Reinforcement learning has improved reasoning in multimodal models beyond supervised trajectory imitation. Vision-R1 (Zhan et al., 2025), MM-Eureka (Meng et al., 2025), and OpenVLThinker (Deng et al., 2025) optimize sampled reasoning outcomes, while VLM-R1 (Shen et al., 2025) and Mixed-R1 (Xu et al., 2025) add perception-oriented signals to answer feedback. PPO (Schulman et al., 2017) estimates advantages with a learned value function; GRPO (Shao et al., 2024) instead normalizes rewards among responses sampled for the same prompt. GRPO is convenient when correctness can be checked automatically, but it does not specify how intermediate evidence should affect the scalar return.

One option is to add format, perception, citation, and answer scores or assign them to separate token groups. Such decompositions provide

local supervision, but they can let answer correctness and citation validity be optimized independently. EviDocVLM instead uses citation integrity as a continuous trajectory-level coefficient on the answer reward: evidence cannot compensate for an incorrect answer, and traceable support distinguishes otherwise equally correct trajectories. This changes the relative return of complete samples rather than selecting which output positions receive each reward.

Outcome calibration alone still cannot determine whether a cited span contains enough information for the answer. EA-GRPO therefore uses the answer-token log-probabilities recorded by the frozen full-context rollout as a teacher for one evidence-view forward pass. A reliability-weighted Monte Carlo estimator aligns the two views without retaining full-vocabulary logits. This training-only comparison connects selected support to answer behavior without a learned evaluator, a separate evidence model, or an evidence-only deployment path.

3 Methodology

Given a question q and retrieved page images $\mathcal{D}_R = \{d_1, \dots, d_k\}$, the model emits $y = (P, E, A)$. Here P is an auxiliary source map, E is the selected support, and A is the answer. This interface can be instantiated by an existing structured document decoder; our contribution is to couple E and A during optimization. Producing a citation and using it are different properties: a syntactically valid citation may be appended to an answer without affecting how the answer is obtained. We therefore supervise evidence at two levels. At the outcome level, evidence reliability determines the value of a correct answer. At the distribution level, an auxiliary conditioning view tests whether the cited content can reproduce the model’s full-context answer behavior. The first signal selects better trajectories; the second acts directly on answer-token distributions.

3.1 Source Map and Support Interface

The source map P serializes each page into addressable records $u_j = (h_j, \ell_j, c_j)$, where h_j is a page-qualified handle, ℓ_j is a coarse layout label, and c_j is recognized content. For example, the handle `p2:e7` denotes the seventh unit on page two. We use these records only to resolve citations. Neither their serialization nor the underlying

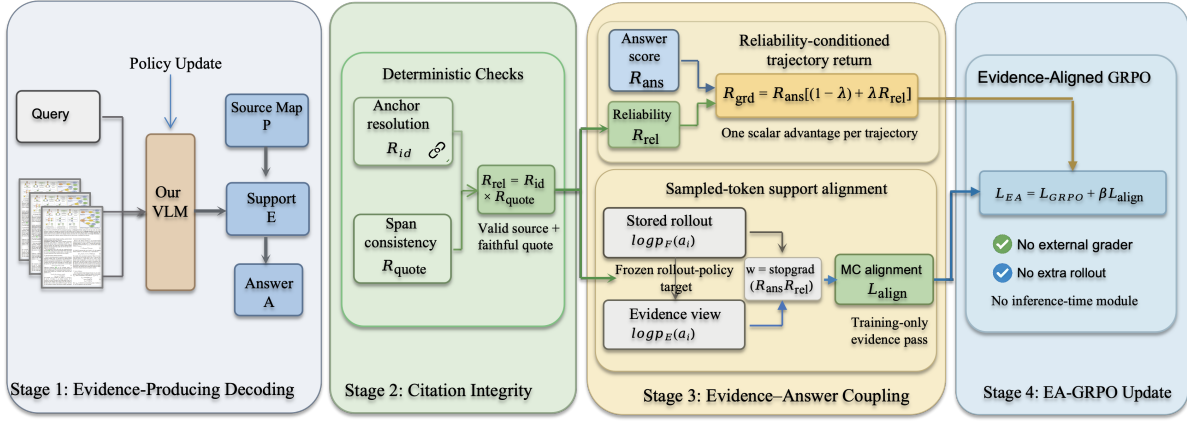


Figure 1: EviDocVLM trains an evidence-producing document model with two complementary signals. Citation integrity scales the trajectory-level answer reward, while a training-only alignment term asks the selected support to reproduce full-context answer behavior. The source map provides citation addresses, and no visual input is removed at inference time.

page-recognition objective is part of our contribution; ordinary sequence fine-tuning supplies this interface before EA-GRPO.

The support block is a short list of handles paired with quotations,

$$E = \{e_m\}_{m=1}^M, \quad e_m = [h_m] \text{ “} v_m \text{”}, \quad (1)$$

where v_m is a short span copied from the record identified by h_m . This compact format avoids a free-form rationale whose source would require a separate semantic check. For example, [p2:e7] “FY 2019: 123.3” identifies both the source and the exact content being used. On samples whose retrieved context is labeled insufficient, the permitted evidence block is the single symbol NONE.

Support is placed before A , so its tokens are already in the decoding history when the answer is produced. This ordering does not restrict the answer to the support block. The query, page embeddings, source map, and support remain available in the normal answer path:

$$p_\theta(y | q, \mathcal{D}_R) = p_\theta(P | q, \mathcal{D}_R) p_\theta(E | P, q, \mathcal{D}_R) \cdot p_\theta(A | E, P, q, \mathcal{D}_R). \quad (2)$$

Thus, a missed source record does not make its visual information inaccessible. The source map supplies addresses without becoming a text-only bottleneck at inference.

3.2 Reliability-Conditioned Trajectory Return

The restricted syntax supports two deterministic checks. Let $\mathcal{U}(P)$ be the handles instantiated in the current source map. Identifier validity, called

Algorithm 1 Evidence-Aligned GRPO for EviDocVLM

Require: SFT policy π_θ , training set $\mathcal{T}$, group size G , coefficients λ, β

```

1: for each training iteration do
2:   Freeze rollout snapshot  $\pi_{\text{old}} \leftarrow \pi_\theta$ 
3:   Sample  $(q, \mathcal{D}_R, A^*)$  from  $\mathcal{T}$ 
4:   Sample  $\{o^i = (P_i, E_i, A_i)\}_{i=1}^G$  from  $\pi_{\text{old}}$ 
5:   for each trajectory  $o^i$  do
6:     Compute the parseability gate  $R_{\text{fmt}}^i$ 
7:     if  $R_{\text{fmt}}^i = 1$  then
8:       Compute  $R_{\text{ans}}^i(A_i, A^*)$  and  $R_{\text{rel}}^i$  (Eq. 5)
9:       Compute  $R_{\text{grd}}^i$  (Eq. 6) and set  $R_i = R_{\text{grd}}^i$ 
10:      Set  $w_i = \text{sg}(R_{\text{ans}}^i R_{\text{rel}}^i)$ 
11:      Record  $\log p_{i,t}^F(a_{i,t})$  from the rollout
12:      Evaluate  $\log p_{i,t}^E(a_{i,t})$  by one evidence-view forward pass
13:     else
14:       Set  $R_i \leftarrow 0$  and  $w_i \leftarrow 0$ 
15:     end if
16:   end for
17:   Normalize  $R_{1:G}$  within the group to obtain  $\hat{A}_{1:G}$ 
18:   Compute clipped  $\mathcal{L}_{\text{GRPO}}$  against  $\pi_{\text{old}}$ 
19:   Compute  $\hat{\mathcal{L}}_{\text{align}}$  from the paired log-probabilities (Eq. 9)
20:   Update  $\theta$  using  $\mathcal{L}_{\text{EA-GRPO}}$  (Eq. 10)
21: end for

```

anchor resolution in Figure 1, measures the fraction of citations that resolve:

$$R_{\text{id}} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[h_m \in \mathcal{U}(P)]. \quad (3)$$

Quotation fidelity, called span consistency in Figure 1, compares v_m with the content $c(h_m)$ of the resolved unit. We use normalized token F1 for text and tolerance-aware equality for numbers, denoted

by $s(\cdot, \cdot) \in [0, 1]$:

$$R_{\text{quote}} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[h_m \in \mathcal{U}(P)] s(v_m, c(h_m)). \quad (4)$$

The averages above apply when $M \geq 1$. For an answerable sample, we define $R_{\text{rel}} = 0$ when the evidence block is empty. For an insufficient-context sample, we define $R_{\text{rel}} = 1$ for NONE and zero for any citation. Otherwise, the reliability score is

$$R_{\text{rel}} = R_{\text{id}} R_{\text{quote}}. \quad (5)$$

Multiplication prevents a fluent quotation with an invalid handle, or a valid handle with altered content, from receiving high credit. R_{rel} measures citation integrity rather than logical entailment. Whether the cited content is useful for answering is addressed separately by the support-alignment constraint below.

Let $R_{\text{ans}} \in [0, 1]$ be the dataset-specific answer score. We turn citation integrity into a continuous routing coefficient $g_i = (1 - \lambda) + \lambda R_{\text{rel}}^i$. Instead of adding R_{rel} as an independent bonus, this coefficient controls how much answer reward is assigned to trajectory i . We refer to this mechanism as reward routing:

$$R_{\text{grd}} = R_{\text{ans}} [(1 - \lambda) + \lambda R_{\text{rel}}], \quad 0 < \lambda < 1. \quad (6)$$

This design keeps three properties. Evidence cannot compensate for a wrong answer; reliable citations distinguish trajectories with equal answer quality; and an otherwise correct trajectory retains $(1 - \lambda)R_{\text{ans}}$ when its citation is imperfect. We use $R_i = R_{\text{fmt}}^i R_{\text{grd}}^i$ as the QA trajectory reward, where R_{fmt} is a binary parseability check, and compute the ordinary group-relative advantage from R_i . This is trajectory-level rather than token-scope routing: the same scalar advantage is applied to the generated sequence, and no reward channel is assigned to a particular output span. Citation integrity therefore changes the relative value of complete answers instead of introducing separate advantages for parsing, citations, and answers.

3.3 Evidence-Aligned GRPO

Reward calibration favors traceable support but does not show that the support contains the information needed for the answer. We therefore construct a training-only second view. The full-view teacher is the frozen rollout policy and retains the images, source map, and support; the evidence view uses

the current policy with only the question, support, and teacher-forced answer prefix:

$$p_{i,t}^F = p_{\text{old}}(\cdot \mid q, \mathcal{D}_R, P_i, E_i, a_{i,<t}), \quad (7)$$

$$p_{i,t}^E = p_{\theta}(\cdot \mid q, E_i, a_{i,<t}). \quad (8)$$

The second view is never used for rollout generation or deployment. It tests whether the selected support contains enough information for the answer while the deployed path continues to use the complete visual context.

The same reliability principle weights the trajectories used for support alignment. Each trajectory contributes in proportion to both answer quality and citation reliability. Let $w_i = \text{sg}(R_{\text{ans}}^i R_{\text{rel}}^i)$. Because answer tokens are already sampled from the full-view rollout policy, we optimize the sampled-token Monte Carlo estimator of the forward KL:

$$\hat{\mathcal{L}}_{\text{align}} = \frac{\sum_{i=1}^G w_i \frac{1}{|A_i|} \sum_{t=1}^{|A_i|} [\log p_{i,t}^F(a_{i,t}) - \log p_{i,t}^E(a_{i,t})]}{\sum_{i=1}^G w_i + \epsilon}. \quad (9)$$

The first log-probability is detached and already recorded during rollout; only the evidence-view term receives gradients. Alignment therefore requires one additional teacher-forced forward pass, but neither full-vocabulary teacher logits nor an additional rollout. Weighting by w_i suppresses low-quality answers and corrupted citations during distillation. Our final objective is

$$\mathcal{L}_{\text{EA-GRPO}} = \mathcal{L}_{\text{GRPO}} + \beta \hat{\mathcal{L}}_{\text{align}}, \quad (10)$$

with λ controlling trajectory-level reliability conditioning and β controlling evidence alignment. This changes neither rollout generation nor the inference architecture. The first term prefers correct answers paired with traceable citations; the second encourages those citations to retain the answer information available in the full visual context. Algorithm 1 summarizes the complete training loop, including rollout scoring, group-relative advantage estimation, construction of the two answer views, and the joint parameter update. A malformed trajectory is assigned zero reward before its answer or citation fields are scored, which makes the format gate explicit and keeps undefined fields out of either optimization term.

Table 1: Parsing quality on OmniDocBench, separated by language and content type. Baseline results follow prior OmniDocBench reports where available, and EviDocVLM is evaluated with the same metrics. Edit-based measures are minimized; TEDS is maximized. Best results are bold and second-best results are underlined.

Method	Overall ED↓		Text ED↓		Formula ED↓		Table TEDS↑		Table ED↓		Reading ED↓	
	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH
Modular Parsing Systems												
MinerU (Wang et al., 2024a)	0.150	0.357	0.061	0.215	<u>0.278</u>	0.577	78.6	62.1	0.180	0.344	<u>0.079</u>	0.292
Marker (Paruchuri, 2024)	0.336	0.556	0.080	0.315	0.530	0.883	67.6	49.2	0.619	0.685	0.114	0.340
Mathpix	0.191	0.365	0.105	0.384	0.306	0.454	77.0	67.1	0.243	0.320	0.108	0.304
Docling (Livathinos et al., 2025)	0.589	0.909	0.416	0.987	0.999	1.000	61.3	25.0	0.627	0.810	0.313	0.837
Pix2Text (Gurgurov and Morshnev, 2024)	0.320	0.528	0.138	0.356	0.276	0.611	73.6	66.2	0.584	0.645	0.281	0.499
Unstructured-0.17.2	0.586	0.716	0.198	0.481	0.999	1.000	-	-	1.000	0.998	0.145	0.387
OpenParse-0.7.0	0.646	0.814	0.681	0.974	0.996	1.000	64.8	27.5	0.284	0.639	0.595	0.641
OCR-Specialized VLMs												
Nougat (Blecher et al., 2024)	0.452	0.973	0.365	0.998	0.488	0.941	39.9	0.0	0.572	1.000	0.382	0.954
Mistral OCR	0.268	0.439	<u>0.072</u>	0.325	0.318	0.495	75.8	63.6	0.600	0.650	0.083	0.284
OLMOOCR-sglang	0.326	0.469	0.097	0.293	0.455	0.655	68.1	61.3	0.608	0.652	0.145	0.277
SmolDocling-256M	0.493	0.816	0.262	0.838	0.753	0.997	44.9	16.5	0.729	0.907	0.227	0.522
General-Purpose VLMs												
GPT-4o (Achiam et al., 2023)	0.233	0.399	0.144	0.409	0.425	0.606	72.0	62.9	0.234	0.329	0.128	0.251
Qwen2-VL-72B (Wang et al., 2024b)	0.252	0.327	0.096	0.218	0.404	0.487	76.8	76.4	0.387	0.408	0.119	0.193
InternVL2-76B (Chen et al., 2024)	0.440	0.443	0.353	0.290	0.543	0.701	63.0	60.2	0.547	0.555	0.317	0.228
Qwen2.5-VL-7B (Bai et al., 2025b)	0.220	0.265	0.142	<u>0.205</u>	0.393	0.530	<u>78.7</u>	<u>78.3</u>	0.155	<u>0.162</u>	0.191	0.169
Qwen3-VL-8B (Bai et al., 2025a)	0.187	<u>0.255</u>	0.085	0.289	0.359	<u>0.465</u>	73.4	81.2	0.351	0.155	0.165	<u>0.136</u>
InternVL3-8B (Zhu et al., 2025)	0.426	0.385	0.315	0.345	0.714	0.729	59.0	71.5	0.352	0.211	0.324	0.257
EviDocVLM	<u>0.180</u>	0.253	0.061	0.147	0.440	0.561	79.1	76.9	<u>0.162</u>	0.181	0.058	0.125

4 Experiments

4.1 Experimental Settings

Evaluation suite. We evaluate complementary capabilities rather than reducing document QA to a single score. OmniDocBench (Ouyang et al., 2024) and olmOCR-Bench (Poznanski et al., 2025) test transcription, structure recovery, and reading order. The downstream evaluation uses InfographicVQA (InfoVQA) (Mathew et al., 2022) as the in-domain benchmark and MP-DocVQA (Tito et al., 2023) as the out-of-domain benchmark, covering visually dense graphics and multi-page business documents. VisRAG-Ret (Yu et al., 2025) supplies the top $k=3$ page images for each query. We also record whether those pages contain the reference support, so missing retrieval evidence can be separated from failures on answerable inputs.

Comparison systems. The parsing study covers modular converters (MinerU (Wang et al., 2024a), Marker (Paruchuri, 2024), Docling (Livathinos et al., 2025), and Pix2Text (Gurgurov and Morshnev, 2024)), OCR-oriented generators such as Nougat (Blecher et al., 2024), and general VLMs including Qwen2-VL-72B (Wang et al., 2024b), Qwen3-VL-8B (Bai et al., 2025a), InternVL2-76B (Chen et al., 2024), Qwen2.5-VL-7B (Bai

et al., 2025b), and InternVL3-8B (Zhu et al., 2025). For parsing baselines, we use reported numbers from the corresponding benchmark or system papers when available. For document QA, we compare off-the-shelf VLMs with visual-reasoning RL models, including Vision-R1 (Zhan et al., 2025), MM-Eureka (Meng et al., 2025), Ocean-R1 (Lingfeng et al., 2025), ThinkLite-VL (Wang et al., 2025d), and OpenVLThinker (Deng et al., 2025). We also compare with retrieval-aware RL systems, namely R1-Router (Peng et al., 2025), MMSearch-R1 (Wu et al., 2025), and VRAG-RL (Wang et al., 2025c). A larger Qwen2.5-VL model and MiMo-VL-RL (Xiaomi, 2025) provide additional context for gains from scale and generic RL.

Scoring protocol. We adopt the sufficient-context formulation of Joren et al. (2025). When the retrieved set contains the information required by the annotation, Accuracy and token-level F1 compare the prediction with the reference answer. When that information is absent, credit is awarded only for the designated abstention response. Aggregating both cases evaluates answer quality and abstention behavior under one protocol. Parsing follows each benchmark’s official edit-distance, TEDS, and ordering measures; arrows in Table 1

Table 2: Representative in-domain and out-of-domain document QA results. Best entries are bold.

Method	ID: InfoVQA		OOD: MP-DocVQA	
	Acc	F1	Acc	F1
Off-the-Shelf VLMs				
Qwen2.5-VL-7B	60.86	54.61	63.28	56.03
MiMo-VL-7B-RL	68.11	45.93	74.11	47.67
Qwen2.5-VL-32B	78.13	66.06	83.93	73.78
Visual Reasoning with RL				
Vision-R1	29.53	27.73	32.49	30.04
Ocean-R1-7B	53.20	53.49	56.35	57.02
MM-Eureka	40.53	40.32	56.68	54.75
ThinkLite-VL-7B	61.70	61.62	62.61	62.37
OpenVLThinker	70.47	70.51	71.74	72.51
Retrieval-Aware Visual RL				
MMSearch-R1	57.94	57.71	61.59	60.82
VRAG-RL	64.86	12.21	73.22	22.39
R1-Router	60.58	15.17	75.97	25.25
EviDocVLM	70.55	69.90	71.70	72.15

indicate their direction.

Training data construction. We draw 30,000 questions from the training splits of ChartQA and InfographicVQA and vary the retained context from one to five pages. Qwen2.5-VL-72B and Qwen2.5-VL-7B propose candidate completions; after filtering with the task-specific answer checker and the required three-field output format, we retain 60,000 trajectories for SFT. For policy optimization, eight-rollout filtering removes prompts that are already solved consistently and prioritizes prompts with high within-group reward variation. Adding 400 cross-page MMLongBench-Doc (Ma et al., 2024) instances yields 4,000 RL prompts.

Training configuration. EviDocVLM starts from Qwen2.5-VL-7B (Bai et al., 2025b). LLaMA-Factory (Zheng et al., 2024) runs one SFT epoch in bf16 with global batch size 32, cosine decay, 0.1 warmup, and a peak learning rate of 5×10^{-7} . Easy-R1 (Yaowei et al., 2025) then runs EA-GRPO for four epochs with eight responses per prompt, temperature 1.2, global and rollout batch sizes of 32, learning rate 10^{-6} , weight decay 10^{-2} , and clipping thresholds 0.2/0.28. All QA systems receive the same retrieval depth. The evidence view adds one teacher-forced pass during training but is not used for rollout generation or decoding.

4.2 Overall Performance

Document parsing. Table 1 shows a mixed profile rather than uniform dominance. Specialized pipelines remain strong on some formula and table criteria, where task-specific recognizers provide

Table 3: Component analysis on the out-of-domain MP-DocVQA benchmark. Best results are bold.

Variant	OOD: MP-DocVQA	
	Acc	F1
EviDocVLM	71.70	72.15
<i>Optimization</i>		
Standard GRPO	64.41	64.70
SFT Only	62.78	63.05
<i>Evidence-Answer Coupling</i>		
w/o Format Gate	66.01	66.28
w/o Reliability Conditioning	70.56	70.80
w/o Support Alignment	70.26	70.45
<i>Citation Integrity</i>		
w/o Identifier Validity	69.54	69.92
w/o Quotation Fidelity	70.81	71.12

reliable local transcription. EviDocVLM is most competitive on text recovery and reading order, the two aspects most directly involved in constructing stable handles for later citation. Its performance also varies less sharply across the English and Chinese partitions than several OCR-oriented systems. The olmOCR breakdown in Table 4 gives a complementary stress test: the model is robust on scientific pages, dense tables, multi-column layouts, and small text, but header and footer extraction remains weaker. Together, the results support the use of a lightweight source map for citation, while showing that evidence-aware training does not replace every specialized recognition component.

Multi-page question answering. Table 2 includes one in-domain collection with visually dense layouts and one out-of-domain collection of multi-page documents. EviDocVLM improves over its initialization in both settings and maintains similar behavior after the domain shift. Its token-overlap quality is competitive with reasoning-oriented models using the same backbone, although the larger general model remains stronger on strict answer accuracy. These results suggest that evidence-aware optimization improves the link between cited content and answers, without eliminating retrieval errors or replacing the benefit of larger models.

4.3 Component Analysis

Table 3 separates optimization, evidence-answer coupling, and citation integrity on MP-DocVQA, the out-of-domain multi-page benchmark from Table 2. Supervised training alone is the weakest variant, while standard GRPO recovers only part

Table 4: olmOCR (Poznanski et al., 2025) results across document domains and layout stress conditions. Baseline results follow the benchmark report where available; higher is better. Best results are bold and second-best results are underlined.

System	All	ArXiv	Scanned Math	Tables	Scans	Headers/Footers	Multi-Column	Tiny Text	Base
olmOCR v0.1.68 (Anchored)	<u>77.4</u>	75.6	75.1	70.2	<u>44.5</u>	93.4	<u>79.4</u>	81.7	99.0
olmOCR v0.1.68 (No Anchor)	76.3	72.1	74.7	71.5	43.7	91.6	78.5	80.5	98.1
Mistral OCR API	72.0	<u>77.2</u>	67.5	60.6	29.3	93.6	71.3	77.1	99.4
GPT-4o (Anchored)	69.9	53.5	74.5	70.0	40.7	93.8	69.3	60.6	96.8
GPT-4o (No Anchor)	68.9	51.5	75.5	69.1	40.9	<u>94.2</u>	68.9	54.1	96.7
Qwen 2.5 VL (No Anchor)	65.5	63.1	65.7	67.3	38.6	73.6	68.3	49.1	98.3
Gemini Flash 2 (Anchored)	63.8	54.5	56.1	<u>72.1</u>	34.2	64.7	61.5	71.5	95.6
MinerU v1.3.10	61.5	75.4	47.4	60.9	17.3	96.6	59.0	39.1	96.6
Marker v1.6.2	59.4	24.3	22.1	69.8	24.3	87.1	71.0	76.9	<u>99.5</u>
Gemini Flash 2 (No Anchor)	57.8	32.1	56.3	61.4	27.8	48.0	58.7	<u>84.4</u>	94.0
Qwen 2 VL (No Anchor)	31.5	19.7	31.7	24.2	17.1	88.9	8.3	6.8	55.5
EviDocVLM	79.8	86.0	<u>75.3</u>	78.0	46.8	78.3	85.0	89.6	99.6

of the gap to the complete objective. Outcome optimization helps, but answer correctness by itself does not account for the full improvement. Reliability conditioning and support alignment affect different failure modes: the former changes which sampled trajectories are preferred through citation integrity, while the latter adds a token-level constraint on the answer. Their effects are consistent with the intended division between traceability and support sufficiency.

The format gate also matters because malformed outputs prevent the remaining checks from being evaluated reliably. Within the citation score, identifier validity contributes more than quotation fidelity, especially on collections that require selecting among several pages. This suggests that resolving the correct source unit is the harder prerequisite, while faithful copying becomes useful once that unit has been located. These ablations support the objective components, but they do not imply that the generated source map is error-free.

5 Conclusion

We studied the evidence utilization gap in multi-page document question answering: a citation may be structurally valid without containing the information needed for the answer. EviDocVLM addresses this gap while preserving direct access to the retrieved page images. Its reliability-conditioned return uses citation integrity to modulate the value of a complete answer trajectory. Evidence-Aligned GRPO then reuses frozen rollout log-probabilities to supervise one evidence-view pass on the sampled answer tokens. These two objectives encourage support that is both traceable and useful for answering without a learned grader,

additional rollout, or inference-time module. Results across document parsing and multi-page QA tasks show that coupling trajectory reward with support-conditioned answer behavior is a practical step toward more accountable document reasoning. Future work may extend the same principle to longer document collections and richer evidence types, where compact citations must support cross-page, multi-step answers.

Limitations

Our study uses a focused set of public document understanding and visual RAG benchmarks, and broader evaluation across domains, languages, and document genres is left for future work. The reported results use one backbone size and a fixed retrieval depth; experiments with other model scales and retrieval budgets would better show how evidence-conditioned optimization interacts with model capacity and visual context length. Citation checks are resolved against a model-generated source map, so an internally consistent transcription error can still pass them; direct page-grounded verification would provide a stronger guarantee. Finally, our reward design relies on rule-based, automatically computed signals. This avoids reward modeling overhead and improves reproducibility, but lightweight verifiable signals for richer document modalities, such as handwriting, mixed-language pages, and low-resolution scans, remain an open direction.

Ethics Statement

This work does not involve human subjects and does not release a new dataset; all experiments use existing public document understanding and

visual question answering benchmarks under their intended research licenses. EviDocVLM operates on document page images that may contain personally identifiable information, financial details, or other sensitive content. In real deployments, retrieved pages, source maps, and selected support should be sanitized and access-controlled before being passed into the model or stored as training data. Citation traceability reduces, but does not eliminate, the risk of ungrounded or biased outputs: support can pass identifier-validity and quotation-fidelity checks while still being incomplete, and a source map can misread the page. Citation-based evidence should therefore be treated as a support signal rather than a guarantee of factual correctness, and human oversight is recommended for high-stakes document settings such as legal, medical, or financial decision-making.

Acknowledgements

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2026M791639).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv:2303.08774*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 44 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.
- Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. *arXiv:2308.13418*.
- Zongsheng Cao, Anran Liu, Yangfan He, Jing Li, Bo Zhang, and Zigan Wang. 2026a. Vig-rag: Video-aware graph retrieval-augmented generation via temporal and semantic hybrid reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 48–56.
- Zongsheng Cao, Anran Liu, Jun Xie, Feng Chen, Lang Chen, Jing Li, Zhepeng Wang, and Zigan Wang. 2026b. Videotrace-r1: Long video-based retrieval-augmented generation via reinforcement learning. In *International Conference on Machine Learning*.
- Zongsheng Cao, Anran Liu, Jun Xie, Lang Chen, Feng Chen, and Zigan Wang. 2026c. Unidocvllm: Enhancing visual reasoning and document understanding for vlm via reinforcement learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 275–284.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. 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*, pages 24185–24198.
- Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *CoRR*, abs/2503.17352.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. *CoRR*, abs/2407.01449.
- Daniil Gurgurov and Aleksey Morshnev. 2024. Image-to-latex converter for mathematical formulas and text. *arXiv preprint arXiv:2408.04015*.
- Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*.
- Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Ming Lingfeng, Li Yadong, Chen Song, Xu Jianhua, Zhou Zenan, and Chen Weipeng. 2025.

- Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. <https://github.com/VLM-RL/Ocean-R1>. Accessed: 2025-04-03.
- Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, and 1 others. 2025. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*.
- Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Wei Yao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. 2024. *Kosmos-2.5: A multimodal literate model*. *Preprint*, arXiv:2309.11419.
- Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.
- Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, and 1 others. 2025. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *CoRR*, abs/2503.07365.
- Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, and 1 others. 2024. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *arXiv preprint arXiv:2412.07626*.
- Vik Paruchuri. 2024. *Marker*.
- Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Yishan Li, Yukun Yan, Shuo Wang, Zhiyuan Liu, Yu Gu, Minghe Yu, Ge Yu, and 1 others. 2025. Learning to route queries across knowledge bases for step-wise retrieval-augmented reasoning. *CoRR*, abs/2505.22095.
- Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. 2025. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. VLM-R1: A stable and generalizable r1-style large vision-language model. *CoRR*, abs/2504.07615.
- Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognit.*, 144:109834.
- Baode Wang, Biao Wu, Weizhen Li, Meng Fang, Zuming Huang, Jun Huang, Haozhe Wang, Yanjie Liang, Ling Chen, Wei Chu, and Yuan Qi. 2025a. Infinity-parser: Layout-aware reinforcement learning for scanned document parsing. *arXiv preprint arXiv:2506.03197*.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv:2409.18839*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025b. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *CoRR*, abs/2502.18017.
- Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025c. VRAG-RL: empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. *CoRR*, abs/2505.22019.
- Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025d. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *CoRR*, abs/2504.07934.
- Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. 2025. Mmsearch-r1: Incentivizing llms to search. *CoRR*, abs/2506.20670.

- LLM-Core-Team Xiaomi. 2025. [Mimo-vl technical report](#). *Preprint*, arXiv:2506.03569.
- Shilin Xu, Yanwei Li, Rui Yang, Tao Zhang, Yueyi Sun, Wei Chow, Linfeng Li, Hang Song, Qi Xu, Yunhai Tong, and 1 others. 2025. Mixed-r1: Unified reward perspective for reasoning capability in multimodal large language models. *CoRR*, abs/2505.24164.
- Zheng Yaowei, Lu Juntong, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. 2025. Easyrl: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyR1>.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. 2025. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. *CoRR*, abs/2503.18013.
- Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, and 1 others. 2024. Mapneo: Highly capable and transparent bilingual large language model series. *CoRR*, abs/2405.19327.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.