CROSSED-WITNESS GROUNDING FOR EVIDENCE-STATE-GUIDED VISUAL DOCUMENT QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Partial evidence makes document question answering a question-relative decision: the same incomplete page can retain a direct witness for one question while losing it for another. Marginal scores on separately altered packets do not establish whether the same incomplete packet elicits opposite question-conditioned decisions. We introduce *Crossed-Witness Grounding* (CWG), a source-bound OCR intervention, reader-adaptation, and reciprocal-audit protocol. CWG pairs two questions with distinct witnesses on one page and removes either witness in turn. Each altered packet requires one exact answer with a citation to a visible block and one explicit refusal; the two deletions reverse these decisions. At deployment the adapted reader sees one question and its visible packet. On 77 source-disjoint held-out pairs, a frozen Qwen3-VL-8B reader passes 3 (3.9%) four-decision tests, paired adaptation passes 26, and the full training sequence passes 33 (42.9%). On 308 original questions, paired adaptation improves joint intact-answer and own-witness-removal behavior by 32.1 percentage points. A 2B reader shows the paired-training gain on the same protocol. Answer-preserving deletion exposure further improves the 8B four-decision score over matched intact continuation. CWG provides a strict reader-level test of question-conditioned evidence use on incomplete known-page packets.

1 INTRODUCTION

In financial-document question answering, a plausible number is not enough: the number must be supported by the material actually delivered to the reader. A damaged scan, an omitted OCR row, or a missing page can change that material between indexing and answering. Suppose a report lists "Buildings: 20 years" and "Machinery: 5–10 years." If the buildings row is lost, the reader should withhold the buildings answer; if the machinery row survives, it should still answer the machinery question. Reliability therefore requires more than a tendency to abstain when a document looks incomplete. It requires an *answer-or-withhold decision conditioned on the question's surviving evidence*. We isolate this reader-level decision on known-page OCR packets before considering retrieval or pixel-level repair.

Document VQA benchmarks such as DocVQA and TAT-DQA test reading from visually rich pages (Mathew et al., 2021; Zhu et al., 2022); DUDE extends evaluation to diverse, multi-page documents and unanswerable questions (Van Landeghem et al., 2023). MMDocIR measures whether systems retrieve relevant pages and layouts from long documents (Dong et al., 2025). Retrieval and repair methods improve access to those pages through iterative search, visual inspection, or failure-aware revision (Jain et al., 2025; Zheng et al., 2026; Jiao et al., 2026). A system allocating time between rereading and retrieval needs a reliable evidence state for the page already in hand: which answer fields have a current witness, and which remain unresolved? Neither a correct answer on an intact page nor a relevant retrieved page establishes this property. In our MMDocIR development diagnosis, model-reviewed support improved under rereading on some questions but worsened on others; increasing inspection alone did not make this decision dependable.

Counterfactual evidence removal supplies useful supervision, but separate scores on intact and answer-deleted packets leave an important behavior untested. After a deletion, the remaining page may still answer a *different* question. A reader that withholds whenever a packet looks shortened can

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.

perform well on own-witness removal while discarding surviving fields. We ask whether document readers distinguish *question-conditioned witness availability* from generic sensitivity to damage.

We introduce *Crossed-Witness Grounding* (CWG) to make that question observable. Its *Witness Intervention* module constructs intact and evidence-removed views from source-bound OCR blocks. *Paired Answer-or-Withhold Adaptation* teaches a reader to produce a cited answer when its direct witness is visible and to withhold that direct span when the witness is removed. The *Crossed-Witness Audit* then pairs two questions about distinct blocks on the same physical page and asks both questions on each of the same two damaged packets. After one block is removed, its own question must withhold while the other question must still answer; after the other block is removed, the decisions reverse. A policy based only on whether a packet looks damaged cannot pass all four cells. Figure 1 illustrates this contrast.

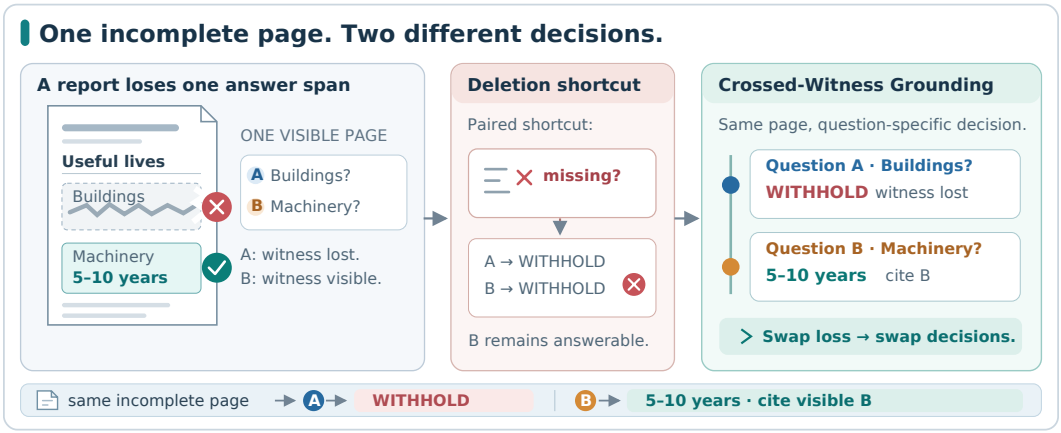

Figure 1: One damaged page, two different decisions. Losing an answer-bearing block should stop the question that needs it, while a question supported by the surviving block remains answerable. CWG crosses the two questions with both deletions, turning a generic "damage means refuse" shortcut into a visible error. The experiments implement damage as controlled OCR-block removal.

The adaptation is intentionally simple at use time: the adapted reader, frozen at deployment, receives one question and its currently visible blocks and generates a cited answer or explicit refusal. The crossed audit is offline; no companion question or deletion label enters the deployed prompt. To separate the effect of evidence-preserving exposure from the choice of which block to remove, we compare continuation on partner-witness deletions with matched repeated-intact and random eligible-deletion controls.

With the 8B reader on source-disjoint TAT-DQA-derived OCR packets, paired adaptation raises joint intact-answer/own-removal success by 32.1 percentage points on 308 questions. The crossed audit is harder: the frozen reader satisfies all four decisions for 3.9% of 77 held-out pairs, paired adaptation reaches 33.8%, and the full training sequence reaches 42.9%. A matched continuation identifies a four-pair improvement from adding an answer-preserving deletion view at 8B. Random eligible and partner-witness deletions tie on this strict endpoint. The same protocol raises a 2B reader from zero to 39.0% four-way success on these held-out pairs.

Our contribution is a reproducible *question-relative evidence test*: a source-bound procedure that makes two questions share each altered packet, an exact cited-answer/withhold contract, and a four-decision endpoint that defeats packet-only refusal. Paired adaptation establishes that readers can improve substantially on this test; the controlled continuation separates general deletion exposure from the choice of a partner witness. The empirical result concerns direct OCR-block witnesses on known pages; page recovery and image-pixel reading remain distinct steps in a complete document system.

2 Related Work

Finding and reading document evidence. DocVQA established document-image question answering at scale, and TAT-DQA targets financial pages containing tables, text, and numerical reasoning (Mathew et al., 2021; Zhu et al., 2022). DUDE and GRAM broaden document evaluation and modeling to multi-page settings (Van Landeghem et al., 2023; Blau et al., 2024); MMDocIR shifts attention to locating relevant pages and layouts in long, multimodal documents (Dong et al., 2025). Systems such as SimpleDoc and Doc-V* combine retrieval with iterative inspection, while S2G-RAG and Doctor-RAG revisit insufficient or failed evidence states (Jain et al., 2025; Zheng et al., 2026; Li et al., 2026; Jiao et al., 2026). This line of work improves the material made available to a reader. The next reliability question arises *within* the visible material: when part of a page is unavailable, which questions still have a direct witness? An intact-page answer score or page-relevance label does not by itself measure that question-relative decision.

Grounding, selective prediction, and counterfactual supervision. Selective VQA treats abstention as a risk–coverage decision (Whitehead et al., 2022); CLIP-UP studies answerable and unanswerable questions about the same image (Vardi et al., 2026). Visual attribution links predictions to source regions (Ma et al., 2025; Souibgui et al., 2025), while input interventions test semantic faithfulness in textual QA and synthesize counterfactual VQA examples (Chaturvedi et al., 2024; Chen et al., 2023). DeFacto also trains with relevant- and irrelevant-region masking, including for document QA, but does not study reciprocal answer-or-withhold decisions on the same incomplete document packet (Xu et al., 2026). CWG pairs two mapped witnesses on one page, alternately removes them, and asks both questions on each altered packet. The resulting four-decision endpoint requires retained answers with exact current-packet citations as well as corresponding refusals.

From evidence sensitivity to question dependence. Context-DPO aligns answers with supplied context, Pair-ID audits one failed RAG query under support addition and nonsupport deletion, and TTIQ uses image- and question-ablated likelihoods for test-time reinforcement learning (Bi et al., 2024; Du, 2026; He et al., 2026). These works establish that correctness alone can miss dependence on the actual inputs. CWG makes the dependence concrete for extractive document QA: the *same* partial page must support one question while withholding the other, and the roles reverse when the other witness is removed. Publisher mappings construct the interventions offline; the adapted reader subsequently answers from its visible packet in one generation. The resulting four-cell test is a direct measure of question-conditioned direct-witness behavior, rather than a reward inferred from answer confidence.

3 Crossed-Witness Grounding

Crossed-Witness Grounding (CWG) makes evidence availability a property of the *question–packet pair*. Its three modules construct controlled witness interventions, adapt an answer-or-withhold reader, and audit whether the reader distinguishes two questions on the same altered packet (Figure 2). Publisher mappings define witnesses and training labels offline. At deployment the frozen reader receives one question and its currently visible OCR packet; it needs neither a second question nor test-time labels.

3.1 Module I: question-relative witness interventions

Let $C = ((i, t_i))_{i=1}^{K}$ be an ordered packet of OCR text blocks from a known page. The reader $f_\theta(q, C)$ returns either a span and its *current-packet* supporting block, (a, i), or $\perp$ when the packet does not supply an admissible direct witness. For a training question q, the publisher gives an answer a_q and mapped witness block W_q. Block IDs are reassigned after an intervention so the sequence does not reveal which block disappeared. The reader never receives W_q or the intervention label.

We restrict the direct-span task to unambiguous string witnesses. Writing $\chi(a, X) = 1$ when the normalized answer occurs in text X, an eligible question satisfies

$$\chi(a_q, W_q) = 1, \qquad \chi(a_q, q) = 0, \qquad \chi(a_q, C_q \setminus W_q) = 0. \tag{1}$$

Deleting W_q therefore removes the literal direct answer from the visible packet. The target is conservative abstention under this direct-witness contract; it does not assert that no alternative calculation from the remaining text exists.

CWG pairs questions q_A, q_B on one physical page when their distinct witnesses W_A, W_B survive each other's removal. Operationally, we build a graph for each page: vertices are eligible question–witness mappings, and an edge joins two vertices only if the answer strings differ, the witness blocks are distinct, and both questions remain qualified after forming a shared packet. A deterministic maximum-cardinality matching uses each question at most once. For every matched edge, we take the source-ordered union C_{AB} of the two local OCR windows, recheck $\chi(a_q, C_{AB} \setminus W_q) = 0$ for each question, and apply both witness deletions to this *same* packet. The question-conditioned label for deleting block U is

$$y(q, C_{AB} \setminus U) = \begin{cases} \bot, & U = W_q, \\ \big(a_q, \mathrm{idx}_{C_{AB}\setminus U}(W_q)\big), & U \neq W_q,\ W_q \in C_{AB} \setminus U. \end{cases} \tag{2}$$

The two deletions create three views for each question: intact, own-witness removed, and answer-preserving removal of another block. The same altered packet can therefore demand opposite decisions:

	$C_{AB} \setminus W_B$	$C_{AB} \setminus W_A$
q_A	(a_A, i_A)	$\bot$
q_B	$\bot$	(a_B, i_B)

. (3)

For either column, a question-blind answer/refusal policy $r(C)$ must make the same decision for both questions, while their targets differ. It can satisfy at most one of the two decisions in that column and cannot pass all four. With answer willingness encoded as one and refusal as zero, the binary targets satisfy $b(q_j, C_{AB} \setminus W_i) = \mathbb{I}[j \neq i]$ and

$$\sum_{i,j \in \{A,B\}} \mathbb{I}[r(C_{AB} \setminus W_i) = b(q_j, C_{AB} \setminus W_i)] = 2 \quad \text{for every packet-only } r. \tag{4}$$

The crossed construction turns a packet-level deletion shortcut into a falsifiable test.

3.2 Module II: answer-or-withhold adaptation

The first stage teaches both sides of the direct-witness decision using an original question packet C_q and its witness-deleted counterpart:

$$\mathcal{L}_{\mathrm{pair}}(\theta) = \mathbb{E}_q[\ell_\theta(y_q \mid q, C_q) + \ell_\theta(\bot \mid q, C_q \setminus W_q)], \quad \ell_\theta(y \mid x) = -\frac{1}{|y|}\sum_{t=1}^{|y|} \log p_\theta(y_t \mid x, y_{<t}). \tag{5}$$

Here $y_q = (a_q, i_q)$ includes the current block citation. Positive-only adaptation retains only the first term. Paired supervision supplies both answer and refusal targets, but alone can reward a packet-level deletion shortcut.

The second stage adds an *answer-preservation* view. For $q \in \{q_A, q_B\}$, sample an eligible $U \neq W_q$ from deletion distribution π_q and retain the answer target after removing U. Starting from the same paired checkpoint, the continued objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{view}}(\theta; \pi) = {} & \tfrac{1}{4}\mathbb{E}_q \ell_\theta(y_q \mid q, C_q) + \tfrac{1}{4}\mathbb{E}_{q,\bar{q}} \ell_\theta(y_q^{AB} \mid q, C_{AB}) \\ & + \tfrac{1}{4}\mathbb{E}_{q,\bar{q}} \ell_\theta(\bot \mid q, C_{AB} \setminus W_q) + \tfrac{1}{4}\mathbb{E}_{q,\bar{q}} \mathbb{E}_{U \sim \pi_q} \ell_\theta(y_q^U \mid q, C_{AB} \setminus U). \end{aligned} \tag{6}$$

The superscripts distinguish citations after block renumbering; all answerable targets contain the same span a_q. Original-packet anchors continue supervision on intact reading. The final two terms impose complementary responses to evidence loss: withhold when this question loses its witness, and keep answering when its witness remains.

We instantiate π_q in two ways. The *partner-witness* arm (crossed-witness augmentation, CWA) chooses $U = W_{\bar{q}}$; a *random-preserving* arm chooses an eligible non-own block without consulting the paired question and may select $W_{\bar{q}}$ by chance. A row-matched control repeats C_{AB} in place of the final deletion. All three share the paired initialization, questions, target balance, and

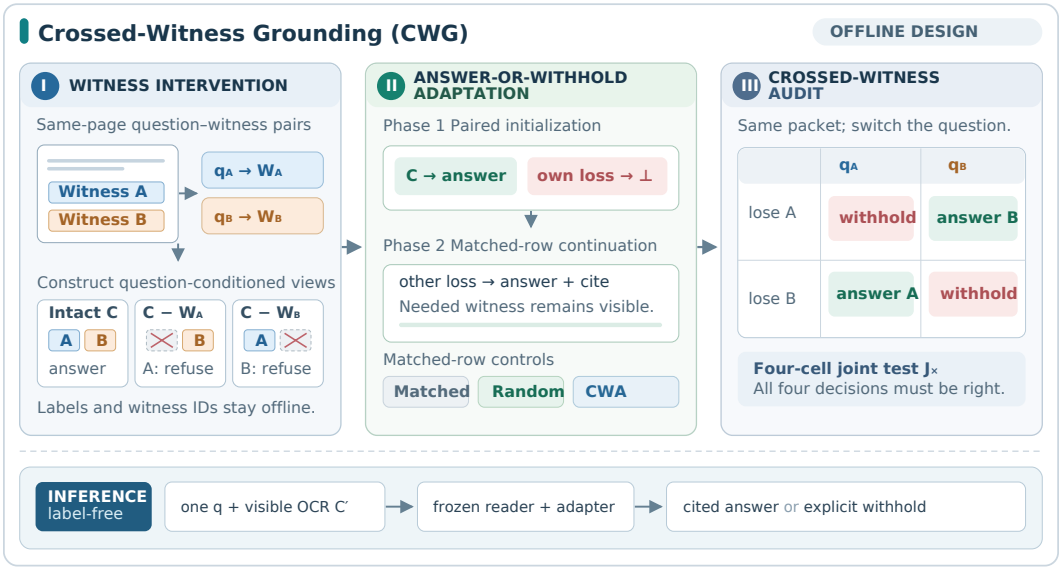

Figure 2: CWG flows from offline witness interventions through paired answer-or-withhold adaptation to a single-generation reader and an offline crossed audit. Matched, random-preserving, and partner-witness continuations compare the effect of the final answer-preserving view.

optimization schedule. Their comparison distinguishes answer-preserving deletion from deliberate partner-witness selection. Equation 6 is ordinary supervised sequence learning; the view distribution is the training intervention. CWG's distinctive construction crosses question identity with witness removal on a shared packet, while this adaptation tests whether the constructed views help a reader satisfy the reciprocal evidence contract.

3.3 Module III: Crossed-Witness Audit

Let $A_\theta(q, C)$ indicate an exact normalized answer with the correct current-packet citation, and $R_\theta(q, C)$ indicate a valid explicit refusal. For held-out pair k, CWG scores the four linked decisions as one outcome:

$$J_k^{\times} = A_\theta(q_A, C \setminus W_B)\, R_\theta(q_A, C \setminus W_A)\, R_\theta(q_B, C \setminus W_B)\, A_\theta(q_B, C \setminus W_A). \tag{7}$$

The pair succeeds only if both questions retain their answers after irrelevant deletion *and* withhold after own-witness deletion. We average over disjoint held-out pairs and resample uncertainty by document source. The matched arms expose two distinct contrasts:

$$\Delta_{\text{preserve}} = \mathbb{E}[J^{\times}_{\text{random}} - J^{\times}_{\text{matched}}], \qquad \Delta_{\text{partner}} = \mathbb{E}[J^{\times}_{\text{partner}} - J^{\times}_{\text{random}}]. \tag{8}$$

The first compares a generic answer-preserving deletion against repeated intact exposure; the second tests whether deliberately selecting the partner's witness improves on that generic deletion.

We also measure the question-by-deletion interaction independently of answer-string correctness. Let $s_\theta(q, C)$ be one for a valid non-refusal and zero otherwise. Then

$$\Gamma_k = \big[s(q_A, C \setminus W_B) - s(q_A, C \setminus W_A)\big] - \big[s(q_B, C \setminus W_B) - s(q_B, C \setminus W_A)\big]. \tag{9}$$

If answer willingness decomposes into a question propensity and a packet-only propensity, $s(q, C) = u(q) + v(C)$, both terms cancel and $\Gamma_k = 0$. A positive interaction isolates a question-relative response; $J_k^{\times}$ additionally requires correct spans and citations.

A shared, zero-call provenance gate checks whether a cited visible block literally contains the normalized answer. Let $I(C)$ be the visible block-ID set:

$$G(a, i, C) = \begin{cases} (a, i), & i \in I(C) \ \wedge \ \chi(a, t_i) = 1, \\ \bot, & \text{otherwise.} \end{cases} \tag{10}$$

For raw outcomes, a removed-witness success requires the model's explicit refusal. For gated outcomes, a rejected answer is a valid *system* abstention even if the model did not generate the refusal token. We report both for every arm. The gate verifies a literal citation link; the crossed test asks whether the decision follows the *current question's* remaining witness. At use time, the reader makes one generation from (q, C); all witness matching, interventions, and four-cell scoring are offline operations.

4 EXPERIMENTS

4.1 EVALUATION DESIGN

Datasets and evidence shifts. We study both a broad document-reading diagnostic and a source-disjoint intervention benchmark. On MMDocIR, 478 known-document questions from 246 sources and nine domains probe whether another visual reading repairs an initial answer. The main controlled benchmark uses TAT-DQA financial pages with publisher answer-to-OCR-block mappings (Zhu et al., 2022). A deterministic PDF-source split reserves 308 questions from 32 PDFs and trains on 1,407 questions from 135 others. Each question is evaluated with its direct witness intact and removed. Same-page matching yields 321 training pairs from 108 sources and 77 evaluation pairs from 29 sources, with distinct answers and nonoverlapping witnesses. The 77 pairs produce 154 questions and 462 intact/own-removed/other-removed views. Controlled block removal represents loss of a local evidence region while preserving surrounding page context; held-out packets contain a median of 11 visible OCR blocks (range 6–16), including neighboring rows and headers. The supplementary packet archive contains the frozen OCR views, pair IDs, witness mappings, and hashes with TAT-DQA attribution; it does not redistribute the source PDFs or page images.

Readers and controls. The main reader is Qwen3-VL-8B-Instruct (Bai et al., 2025) operating on supplied OCR text, with a fixed answer/citation schema and deterministic decoding. We compare the frozen reader, positive-only LoRA (Hu et al., 2022), and paired answer-or-withhold LoRA. LoRA keeps base weights fixed; the original-packet anchors in Eq. 6 explicitly preserve exposure to intact reading. From a single reproduced paired checkpoint, three row- and target-matched continuations repeat the intact packet, delete a random eligible answer-preserving block, or delete the partner question's witness. They share initialization, questions, optimization schedule, and all training views except the final deletion view. This experiment tests a general VLM's document-reading language path. Quantization, adapter rank, and exact training budgets appear in Appendix A.

Endpoints and uncertainty. For an answerable view we require both an exact normalized span and its correct current-packet block ID. A removed-witness view counts only a valid explicit `INSUFFICIENT_EVIDENCE` JSON response as a *raw* success; a parse error is not a refusal. In gated scores, a lexical-citation rejection counts as a system abstention. We report four-way pair success $J^{\times}$ (Eq. 7), its component success rates, the interaction Γ (Eq. 9), and performance on the original 308 intact/removed question pairs. The same zero-call lexical citation gate (Eq. 10) is applied to every arm. Paired confidence intervals resample PDF sources, keeping both questions and their views together. We also measure original-task retention, answer form, and computation.

4.2 PAIRED ADAPTATION RESTORES THE ANSWER/WITHHOLD DECISION

The TAT-DQA intervention supplies exact answer/citation and explicit refusal targets for a reader's direct-witness decision. Unlike the MMDocIR rereading diagnostic that motivates this setting, these source-disjoint packets test the adapted reader itself.

Table 2: Crossed-witness results on 77 disjoint pairs from 29 held-out PDFs. Four-way is the percentage of pairs satisfying both cited answers and both refusals; component counts use 154 questions and original joint uses 308 questions.

Reader	Four-way	Intact	Own $\perp$	Other answer	Three-view	Original joint
Frozen	3.9%	85	55	85	31	–
Paired start	33.8%	113	108	115	85	180
Repeated intact	37.7%	107	130	109	93	197
Random-preserving	42.9%	105	137	108	92	207
Partner-witness	42.9%	110	133	112	97	206

Table 1: Original TAT-DQA intervention, 308 questions from 32 held-out PDFs. Present requires an exact answer and citation; removed requires explicit model abstention for raw scores or model/gate abstention for gated scores. Joint requires both on the same question.

Reader	Present	Removed	Joint	Joint + gate
Frozen	172	126	78	106
Positive-only LoRA	224	0	0	24
Paired LoRA	218	243	177	179

Positive-only adaptation improves exact cited-span extraction to 224 of 308 but yields no valid raw refusals after direct-witness removal. Paired adaptation instead lifts the joint answer-and-withhold rate from 25.3% (78/308) to 57.5% (177/308): $+32.14$ percentage points with a PDF-source-cluster 95% interval of $[27.18, 37.04]$. It retains 218 intact-view exact-and-cited answers, close to the positive-only 224. Under the *same* zero-call citation gate, paired adaptation still achieves 179 joint successes versus 106 for the frozen reader. The paired training improves both the raw reader and the gated system.

The stronger crossed test reveals the remaining challenge. The frozen reader passes only 3.9% of held-out pairs (3/77); a reproduced paired reader reaches 33.8% (26/77), a $+29.87$-point source-cluster difference with interval $[21.52, 39.06]$. The new answer-preserving view is tested against this paired starting point.

4.3 CWG EXPOSES QUESTION-CONDITIONED EVIDENCE USE

The full CWG training sequence raises strict success from 3.9% for the frozen reader to 42.9% (33/77): $+38.96$ points with a PDF-source-cluster interval of $[27.27, 50.68]$ for the random-preserving implementation. The paired start already reaches 33.8%, and repeated intact continuation reaches 37.7%. This paired start is a separately reproduced checkpoint, with 180/308 original-task joint successes versus 177/308 for the run in Table 1. This strict gain requires *both* cited answers after irrelevant deletion and *both* own-witness refusals on the same pair; refusing every damaged packet cannot raise it. The common lexical gate changes frozen success from 3 to 6 pairs and leaves adapted four-way scores unchanged.

Compared with repeated-intact continuation, adding a random eligible answer-preserving deletion view raises the strict four-decision score by four pairs, or 5.19 points ($[1.23, 10.45]$ source-cluster interval). Deliberately deleting the partner witness reaches the same 33/77 as random deletion ($[-6.49, 5.81]$ for partner minus random). The 8B increment is consistent with a benefit from the answer-preserving view under this matched-row protocol, with no measured gain from choosing the partner witness specifically. The question-by-deletion interaction Γ rises from 0.71 (frozen) to 1.35 (paired) and 1.61/1.57 (random/partner), mirroring the stricter answer-and-citation endpoint.

The two preserving arms tie on the primary score but trade component errors (Table 2). Relative to repeated intact, random deletion changes retained answers from 109 to 108 of 154 and own-witness refusals from 130 to 137; the measured gain is in the *joint decision*, not a separately established increase in retained answers. On the original 308-question test, random and partner continuations reach 207 and 206 joint successes versus 180 for their shared paired start. Relative to that start, their intact-answer counts change from 226 to 216/217 while own-witness refusals rise from 232 to

Table 3: Qwen3-VL-2B crossed-witness replication with rank-eight LoRA. Four-way counts are out of 77 pairs; component counts use 154 questions; original joint counts use 308 questions. All entries are raw model scores.

Reader	Four-way	Intact	Own $\perp$	Other answer	Original joint
Frozen	0	50	19	49	–
Paired start	17	106	90	108	144
Repeated intact	29	101	135	101	184
Random-preserving	30	104	133	102	182
Partner-witness	28	99	134	101	179

289/286. Appendix Table 4 reports the full marginal breakdown, and Table 6 shows the four-cell error distribution.

An illustrative crossed pair. A held-out annual report asks where most revenue originates (witness B5) and for the firm-fixed-price total (witness B9: 3,574). After B9 is removed, the frozen and paired readers instead return an unrelated 10,194 from B3; the random-preserving reader withholds. After B5 is removed, that reader still answers the firm-fixed-price question and cites the renumbered B8. It also gives the reciprocal answer and refusal for the revenue-origin question. This single packet shows the distinction measured across all 77 pairs in Table 2.

4.4 Backbone scale and adapter rank

We repeat the frozen protocol with Qwen3-VL-2B-Instruct, using the same source split, 462 crossed and 616 original test views, answer schema, decoder, and scoring rule. The two Qwen3-VL tokenizers produce identical input token IDs for every test view. Table 3 reports the complete rank-eight training sequence.

The frozen 2B reader passes no four-way pair; paired adaptation reaches 17/77, a $+22.08$-point PDF-source-cluster gain with interval $[10.84, 34.18]$. Random-preserving continuation reaches 30/77 (39.0%) and retains 182/308 original-task joint decisions, versus 144/308 at the paired start ($+12.34$ points, $[5.75, 19.51]$), while intact-answer retention changes from 203 to 197/308 (-1.95 points, $[-6.98, 3.59]$). The row-matched intact continuation already reaches 29/77: random preservation adds one pair ($+1.30$ points, $[-2.86, 5.62]$) at this scale, compared with four pairs at 8B. The partner-witness arm scores 28/77, two pairs below random preservation ($[-7.61, 2.30]$ for their paired difference). The paired answer/withhold gain transfers to the smaller reader; the deletion-specific increment is more pronounced in the 8B experiment.

At rank sixteen, the 2B paired, random-preserving, and partner-witness readers pass 14, 25, and 28 pairs. These sensitivity runs do not resolve an adapter-rank advantage or a reliable partner-selection effect; Appendix Table 5 gives the component and original-task results.

4.5 Answer form and efficiency

Of the 154 crossed questions, 124 have an answer string containing a digit. In this stratum, the partner/random three-view joint counts are 82/77; on the other 30 questions both obtain 15. The full four-way pair outcome remains tied. The three 8B continuations are row-matched at 2,568 examples and 321 optimizer steps; measured target-inclusive token totals range from 1.356 to 1.392 million. Deployment makes one deterministic generation per question view, with no test-time parameter update. Detailed training and inference costs are in Appendix A.

5 Conclusion

Crossed-Witness Grounding makes question-relative evidence use measurable: on the same incomplete document packet, one question must receive an answer cited to a surviving block while another must be withheld, and switching the removed witness reverses those decisions. On the source-held-out OCR benchmark, the full training sequence raises strict four-decision success from 3.9% to 42.9% for Qwen3-VL-8B and from zero to 39.0% for Qwen3-VL-2B. The crossed intervention and

audit supply the distinctive test; the reader uses supervised adaptation, with an 8B gain from an answer-preserving deletion view and no measured gain from choosing the partner witness over an eligible random block.

For document assistants, this question-relative behavior matters whenever scans, OCR pipelines, or indexes deliver only part of a report: a surviving field can still support a traceable answer, while an absent direct witness should not be replaced by a plausible guess. The four-cell audit offers a regression test for that behavior as document pipelines change. A next step is to connect the reader's evidence state to budgeted rereading and page retrieval, then test it on pixel-level damage and naturally incomplete multi-page collections with independently assessed answerability.

Reproducibility Statement

The supplementary archive provides the frozen source-disjoint OCR packets, pair IDs, witness mappings, and input hashes. Appendix A records the construction, training controls, scoring, and uncertainty procedure; the underlying TAT-DQA source is needed to rebuild the OCR packets from documents.

AI Use Statement

OpenAI ChatGPT and Codex assisted with literature discovery and manuscript drafting; the design of the controlled OCR study and its four-cell formulation; and data and code preparation, experiment execution, and analysis. We verified cited sources, experimental records, and reported claims. The authors take responsibility for the paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, et al. Context-DPO: Aligning language models for context-faithfulness. arXiv:2412.15280, 2024. URL https://arxiv.org/abs/2412.15280.

Tsachi Blau, Sharon Fogel, Roi Ronen, Alona Golts, Roy Ganz, Elad Ben Avraham, Aviad Aberdam, Shahar Tsiper, and Ron Litman. GRAM: Global reasoning for multi-page VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15598–15607, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Blau_GRAM_Global_Reasoning_for_Multi-Page_VQA_CVPR_2024_paper.html.

Akshay Chaturvedi, Swarnadeep Bhar, Soumadeep Saha, Utpal Garain, and Nicholas Asher. Analyzing semantic faithfulness of language models via input intervention on question answering. *Computational Linguistics*, 50(1):119–155, 2024. doi: 10.1162/coli_a_00493. URL https://aclanthology.org/2024.cl-1.5/.

Long Chen, Yuhang Zheng, Yulei Niu, Hanwang Zhang, and Jun Xiao. Counterfactual samples synthesizing and training for robust visual question answering. IEEE TPAMI; arXiv:2110.01013v2, 2023. URL https://arxiv.org/abs/2110.01013.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract.html.

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of EMNLP*, pp. 30971–31005, 2025. doi: 10.18653/v1/2025.emnlp-main.1576. URL https://aclanthology.org/2025.emnlp-main.1576/.

Wenzhang Du. What would fix this RAG failure? auditing counterfactual response with paired evidence interventions. arXiv:2608.08944, 2026. URL https://arxiv.org/abs/2608.08944.

Xinrui He, Ting-Wei Li, Junting Wang, Mengting Ai, Xinyu He, Hanghang Tong, and Jingrui He. Harnessing image question dependence for better VLM test-time reinforcement learning. arXiv:2609.13296, 2026. URL https://arxiv.org/abs/2609.13296.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, and Huazheng Wang. SimpleDoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. In *Proceedings of EMNLP*, pp. 28410–28427, 2025. doi: 10.18653/v1/2025.emnlp-main.1443. URL https://aclanthology.org/2025.emnlp-main.1443/.

Shuguang Jiao, Chengkai Huang, Shuhan Qi, Xuan Wang, Yifan Li, Quanchi Weng, Lingchuan Liu, Xunliang Cai, and Lina Yao. Doctor-RAG: A failure-aware repair framework for agentic retrieval-augmented generation. arXiv:2604.00865, 2026. URL https://arxiv.org/abs/2604.00865.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. arXiv:2604.23783, 2026. URL https://arxiv.org/abs/2604.23783.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. VISA: Retrieval augmented generation with visual source attribution. In *Proceedings of ACL*, 2025. URL https://aclanthology.org/2025.acl-long.1456/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Mohamed Ali Souibgui, Changkyu Choi, Andrey Barsky, Kangsoo Jung, Ernest Valveny, and Dimosthenis Karatzas. DocVXQA: Context-aware visual explanations for document question answering. arXiv:2505.07496, 2025. URL https://arxiv.org/abs/2505.07496.

Jordy Van Landeghem, Ruben Tito, Lukasz Borchmann, Michal Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Sien Moens, and Tomasz Stanislawek. Document understanding dataset and evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19528–19540, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Van_Landeghem_Document_Understanding_Dataset_and_Evaluation_DUDE_ICCV_2023_paper.html.

Ben Vardi, Oron Nir, and Ariel Shamir. CLIP-UP: CLIP-based unanswerable problem detection for visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5898–5908, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Vardi_CLIP-UP_CLIP-Based_Unanswerable_Problem_Detection_for_Visual_Question_Answering_WACV_2026_paper.html.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. arXiv:2204.13631, 2022. URL https://arxiv.org/abs/2204.13631.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, Jing Liu, and Feng Chen. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. In *Proceedings of the International Conference on Machine Learning*, 2026. URL https://arxiv.org/pdf/2509.20912.

Yuanlei Zheng, Pei Fu, Hang Li, Ziyang Wang, Yuyi Zhang, Wenyu Ruan, Xiaojin Zhang, Zhongyu Wei, Zhenbo Luo, Jian Luan, Wei Chen, and Xiang Bai. Doc-V*: Coarse-to-fine interactive visual reasoning for multi-page document VQA. arXiv:2604.13731, 2026. URL https://arxiv.org/abs/2604.13731.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of ACM Multimedia*, 2022. doi: 10.1145/3503161.3548422. URL https://arxiv.org/abs/2207.11871.

A Construction and Reproducibility Details

Source split and qualification. We split the TAT-DQA publisher training set by a deterministic hash of PDF source identity, keeping every question from a source on one side. Eligible direct-span questions have a mapped OCR block containing the normalized answer; the answer appears neither in the question nor elsewhere in the constructed packet after that block is removed. Within a page, we match questions with different answers and distinct witness blocks using a deterministic maximum-cardinality matching, then recheck eligibility on their shared packet.

Shared packets. Each packet is the source-ordered union of both questions' local OCR windows: three blocks on either side of each witness and two fixed distractor positions, restricted to their common page. Block text follows the parent experiment's 1,800-character cap. Removing a complete witness renumbers visible blocks from B1. Reader inputs contain the current question and visible blocks, but no witness map, partner question, deletion label, or reference answer. The frozen split contains 321 training and 77 held-out question pairs.

Matched adaptation. All three 8B continuations start from the same reproduced paired LoRA checkpoint. This checkpoint scores 180/308 original-task joint decisions; the earlier paired run in Table 1 scores 177/308 and is not used as a continuation baseline. Each continuation uses rank-eight LoRA with 4-bit NF4 base weights (Dettmers et al., 2023), one epoch of AdamW, learning rate 5×10^{-5}, gradient accumulation eight, and a 3,072-token input-plus-target limit. Its 2,568 rows comprise the same 642 original-packet anchors, 642 intact shared packets, and 642 own-witness removals, plus one 642-row fourth component. That component repeats the intact packet, removes a random eligible answer-preserving block, or removes the partner question's witness. Question order and target balance are matched across arms.

The random control excludes the current witness and blocks uniquely carrying visible question words or multi-digit numbers. It favors blocks within a factor of two of a hash-sampled witness length and samples independently of the partner question; its deletion differs from the partner witness in 546 of 642 rows. The arms match row count and optimizer steps, while processed token counts can differ.

Table 4: Original-packet retention for the shared 8B paired checkpoint and its three continuations. Each column counts questions out of 308; joint success requires both an intact exact-and-cited answer and an own-witness-removal refusal.

Reader	Intact answer	Own-removal refusal	Joint
Shared paired start	226	232	180
Repeated intact	219	274	197
Random-preserving	216	289	207
Partner-witness	217	286	206

Scoring and uncertainty. An answerable view requires an exact normalized answer and its current block citation; raw own-witness-removal success requires explicit `INSUFFICIENT_EVIDENCE` with null citation. Invalid JSON fails. The common lexical gate rejects an answer absent from its cited block. Intervals resample PDF sources, keeping both questions and all views of a pair together. Every continuation is also evaluated on the original 308 intact/removed questions.

Scale and cost. The 8B continuations process 1.356–1.392 million target-inclusive tokens each. Evaluation makes 462 crossed and 616 original-packet deterministic generations per adapter. Qwen3-VL-2B follows the same data, prompts, decoding, and rank-eight protocol. The rank-sixteen sensitivity doubles trainable adapter parameters from 8.7 to 17.4 million and compares paired, random-preserving, and partner-witness arms; it has no matched continuation. Table 5 reports its raw outcomes.

Table 5: Qwen3-VL-2B rank-sixteen sensitivity on the identical held-out questions. Four-way is out of 77 pairs; the next three columns are out of 154 questions, and original-task columns are out of 308.

Reader	Four-way	Intact	Own $\perp$	Other	Orig. pres.	Orig. rem.	Orig. joint
Paired	14	107	75	109	201	184	129
Random-preserving	25	100	126	97	194	274	175
Partner-witness	28	97	134	96	189	276	171

Table 6: Held-out pairs with 0–4 correct crossed cells, computed from cached raw outputs. The four-cell column equals the primary four-way score.

Reader	0	1	2	3	4
Frozen	9	21	25	19	3
Paired start	3	7	11	30	26
Repeated intact	0	5	11	32	29
Random-preserving	0	3	13	28	33
Partner-witness	0	4	11	29	33