DELTABIND: RELATION-BOUND RETRIEVAL FOR VISUALLY RENDERED TABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual table retrieval can fail when candidates contain the same words but associate values with different rows or columns. Page-wide late interaction can combine query-token matches from incompatible relations, a failure we call *relation collision*. We introduce DELTABIND, a training-free framework that makes a complete table relation the unit of query matching. Using predicted table structure, DELTABIND assembles source-image crops of a value, its row label, and its column header into a relation tile. A frozen retriever scores each tile with late interaction and ranks a table by its highest-scoring tile, keeping query-token matches within a shared relation. For aligned table versions, shared content can obscure the relations that distinguish candidates. A query-independent change index therefore focuses comparison on changed relations and enables sparse caching. DELTABIND improves ColQwen2's strict pair accuracy from 51.6% to 79.7% on TabFact-HumanBind. Relation binding alone also improves source-local retrieval on mined pools of unedited FinTabNet and PubTables-1M tables. OCR–BM25 controls and component ablations support the value of jointly preserving row, column, and value context.

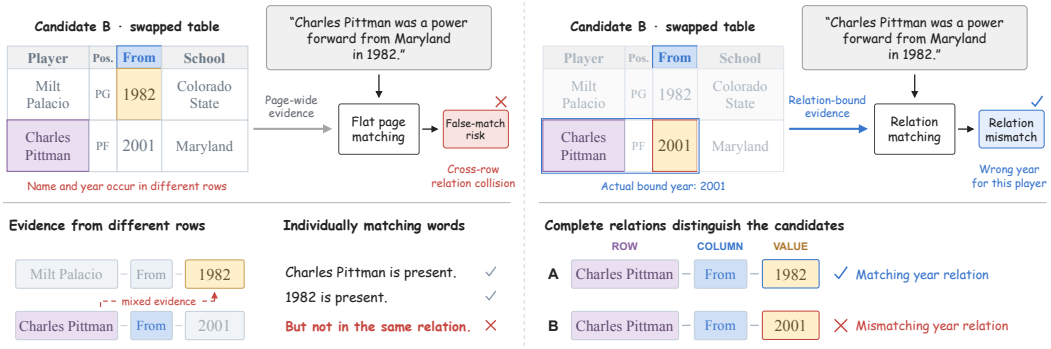


Figure 1: **Relation collision: matching words can disagree on their binding.** Left: page-wide matching can collect "Charles Pittman" and "1982" from different rows of Candidate B. Right: a shared row–column–value witness binds this player to 2001, while Candidate A contains the requested year. The schematic isolates the discriminating year relation in the displayed TabFact-HumanBind claim; the other requested attributes are shared by the two candidates.

1 INTRODUCTION

Reports repeat tables across periods, papers present closely related ablations, and document revisions change only a few cells. In these settings, a candidate can contain every query word and still be the wrong result: the requested value may belong to a different row or column. Retrieval must therefore preserve the relation between a value and its context, not just recognize their separate occurrences.

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.

Visual document retrieval represents rendered pages directly, retaining layout and typography that text extraction may discard (Faysse et al., 2025; Yu et al., 2025; Dong et al., 2025). In multi-vector retrievers such as ColPali, late interaction lets each query token select its best-matching image patch independently. Figure 1 illustrates the resulting *relation collision*: a player's name matches one row while the requested year matches another. The missing constraint is a shared relation: score all query tokens within one row–column–value witness before maximizing over witnesses. Section 3.1 formalizes this change in aggregation order.

Two design problems follow. First, retrieval needs an evidence unit that preserves row–column–value context: a value crop loses its labels, while a full table contains competing relations. Second, aligned versions need a comparison centered on their differences. Shared background content can dominate both candidates even when one changed relation determines their relative relevance.

DELTABIND addresses these problems with two components. **Relation binding** places a value and its row and column context on a single tile, performs late interaction within that tile, and maximizes over tiles to score the candidate. For aligned versions, **candidate-relative change filtering** restricts the comparison to shared relation identities whose pixels differ. Frozen structure prediction supplies the crops; either a visual encoder or OCR–BM25 scores the resulting evidence. Binding applies to unaligned tables as well as aligned versions, whereas filtering requires correspondence between candidate grids.

The central idea is to make the evidence unit match the retrieval decision. An atomic query specifies a value together with the labels that identify its meaning. A version comparison can hinge on the same local relation even when the full claim contains additional, shared attributes. In Figure 1, both tables agree on the player's position and school, while the bound year determines their relative relevance. DELTABIND uses this *relation-local discrimination* to rank candidates: the tile preserves the binding, and the change index identifies which bindings distinguish the versions.

We evaluate this principle in controlled and natural settings. TabFact-HumanBind uses within-column value swaps that preserve the page token multiset. DELTABIND raises ColQwen2 strict accuracy from 51.6% to 79.7%. On mined pools of unedited tables from held-out sources, relation binding raises source-local R@1 from 46.3% to 80.5% on FinTabNet and from 49.3% to 85.9% on PubTables-1M. ColSmol and OCR–BM25 provide additional scorer controls. For aligned pairs, caching changed tiles retains 5.8% of tiles while reproducing the decisions of the full-storage filtered implementation.

Our contributions are:

- **A diagnosis of relation collision.** We characterize how independent token matching can combine incompatible table evidence and make the shared-witness constraint explicit.
- **Relation binding and candidate-relative retrieval.** We introduce DELTABIND, a training-free framework that scores complete relation tiles and uses candidate differences to index aligned versions.
- **Controlled and natural evaluation.** We test content-preserving swaps, unedited tables, multiple scorers, and relation components, and characterize retrieval behavior across atomic and scientific claims.

2 RELATED WORK

Visual document and table retrieval. ColBERT introduced late interaction for text retrieval (Khattab & Zaharia, 2020); ColPali extended multi-vector matching to rendered pages (Faysse et al., 2025). VisRAG couples visual retrieval with generation (Yu et al., 2025), and MMDocIR evaluates page- and layout-level retrieval (Dong et al., 2025). Component graphs (Yun et al., 2025), query refinement (Uzan et al., 2026), and token-aware aggregation (Kang et al., 2025) address evidence granularity or matching. For tables, generic text representations can be competitive (Wang et al., 2022), while field-aware matching (Li et al., 2025) and HELIOS's table–text graphs (Park et al., 2025b) exploit finer structure. TaR-ViR studies open-world rendered-table retrieval (Li et al., 2026), and TabRAG combines multimodal reranking with answer generation (Xu et al., 2026). DELTABIND makes the relation the unit of query matching: all query tokens interact within one row–column–

value witness before candidate-level aggregation. This constraint specifies which local matches can jointly contribute to a candidate score.

Structured evidence and table verification. TabFact evaluates human-written claims against Wikipedia tables (Chen et al., 2020); SciTab emphasizes compositional reasoning (Lu et al., 2023). Structured evidence extraction (Wu et al., 2023) and table–text alignment (Ho et al., 2025) study the cells supporting a verdict. M^2-TabFact extends verification to multiple documents and modalities (Zhou et al., 2025), while FinCARDS uses schema-aligned textual cards for financial document reranking (Zhou et al., 2026). These approaches represent the roles of evidence explicitly. DELTABIND constructs its evidence from source-pixel crops of a value and its labels, using frozen structure predictions to define their relation. Visual and lexical scorers then operate on the same relation boxes. This design also connects our controls to complementary visual and textual retrieval strengths observed in chart retrieval (Nowak et al., 2024).

Candidate-relative retrieval and change localization. Image-based document comparison can localize text changes without relying on OCR language models (Park et al., 2025a). SciClaimEval modifies evidence while retaining authentic scientific claims (Ho et al., 2026). Our controlled benchmark uses content-preserving value swaps to test whether retrieval distinguishes bindings when lexical content is fixed. For aligned candidates, visual change defines the admissible witness set; query relevance still determines the ranking. This separates detecting a difference from deciding whether that difference matters to the query. Unaligned candidates use relation binding without change filtering, so their gains test a different component of the framework.

3 DELTABIND

Given a query q and a supplied candidate group $\mathcal{C}_g$, DELTABIND ranks table images by their strongest relation-level evidence (Figure 2). It constructs relation tiles offline and, for aligned versions, defines a query-independent change index. Sparse storage materializes this index before query scoring. The structure parser and relevance encoders remain frozen throughout.

3.1 RELATION COLLISION AND THE SHARED-WITNESS CONSTRAINT

Let q have query vectors $\{\boldsymbol{q}_i\}_{i=1}^{L}$, and let a table image d have patch vectors $\mathcal{P}(d)$. Conventional late interaction assigns

$$s_{\mathrm{flat}}(q,d) = \sum_{i=1}^{L} \max_{\boldsymbol{p}\in\mathcal{P}(d)} \langle \boldsymbol{q}_i, \boldsymbol{p}\rangle. \tag{1}$$

Each token can select a patch from a different table relation. To formalize this freedom, consider a finite, nonempty witness set $\mathcal{W}(d)$ with fixed, finite, nonempty evidence sets $\mathcal{P}_w$. Define $\phi_i(w) = \max_{\boldsymbol{p}\in\mathcal{P}_w}\langle \boldsymbol{q}_i, \boldsymbol{p}\rangle$. Independent matching and shared-witness matching then give

$$s_{\mathrm{free}} = \sum_i \max_w \phi_i(w), \qquad s_{\mathrm{bind}} = \max_w \sum_i \phi_i(w) \;\le\; s_{\mathrm{free}}. \tag{2}$$

The difference is a nonnegative *binding gap*:

$$\Gamma(q,d) = s_{\mathrm{free}} - s_{\mathrm{bind}} = \min_{w\in\mathcal{W}(d)} \sum_i \left[\max_{v\in\mathcal{W}(d)} \phi_i(v) - \phi_i(w)\right]. \tag{3}$$

Each bracket measures a token's loss when all matches are assigned to the same witness. The gap is zero exactly when one witness attains every token's maximum. For two relations with match vectors (a,b) and (b,a), where $a>b$, independent matching scores $2a$, shared matching scores $a+b$, and the gap is $a-b$. When $\bigcup_w \mathcal{P}_w = \mathcal{P}(d)$, s_{free} equals Eq. 1. These expressions isolate the aggregation rule on fixed evidence vectors. The implementation realizes this constraint with re-rendered tiles, which present each relation in its own visual context.

The relation as the evidence unit. A relation links a value to its row label and column header. We call the corresponding unit of evidence a *witness*; its image representation is a *relation tile*. For

example, *the operating margin in 2023 was 17.2%* specifies a metric, a period, and a value within one relation. In a version pair, one changed relation can determine relative relevance while other attributes remain shared. DELTABIND scores the full query against each witness.

3.2 RELATION-BOUND EVIDENCE REPRESENTATION

A frozen Table Transformer (TATR) (Smock et al., 2022) predicts rows, columns, and a header region from the table image. We use the released v1.1-All checkpoint (Smock et al., 2023). With data row R_i, row-stub column K_0, data column K_j, and header row R_h, witness $w = (i, j)$ consists of three boxes:

$$b^v_{ij} = R_i \cap K_j, \qquad b^r_{ij} = R_i \cap K_0, \qquad b^c_{ij} = R_h \cap K_j. \tag{4}$$

The relation tile $T(d, w)$ places the value, row-label, and column-header crops in separate labeled bands on a 512×240 canvas. The bands preserve each component's role across tables with different cell dimensions and surrounding content. Witnesses in the same row share a row-label crop; those in the same column share a header crop. Construction is query-independent and uses source pixels and predicted structure. The first detected column supplies the row stub, and the selected header row supplies the column labels.

A frozen visual encoder scores each tile by within-witness late interaction:

$$r_{\text{vis}}(q, T) = \frac{1}{L} \sum_{i=1}^{L} \max_{\boldsymbol{t} \in \text{Enc}(T)} \langle \boldsymbol{q}_i, \boldsymbol{t} \rangle. \tag{5}$$

The query-length normalization leaves within-query rankings unchanged. Every query token interacts with the same relation tile, so its best match can use either the value or the associated context. The lexical variant assigns page OCR strings to the same component boxes, concatenates the three fields, and applies BM25 (Robertson & Zaragoza, 2009):

$$r_{\text{lex}}(q, T) = \text{BM25}\big(q, \text{OCR}(b^v) \parallel \text{OCR}(b^r) \parallel \text{OCR}(b^c)\big), \tag{6}$$

where $\parallel$ denotes text concatenation. This gives visual and lexical scorers the same relation structure. For unaligned candidates, all predicted witnesses compete:

$$S_{\text{rel}}(q, d) = \max_{w \in \mathcal{W}(d)} r(q, T(d, w)). \tag{7}$$

We call this relation-binding mode BINDTILE-Max. Its inner aggregation combines token evidence within a tile; its outer maximum selects the candidate's strongest relation. Each candidate can therefore select the witness most relevant to the query, including when candidate tables have different grids.

3.3 CANDIDATE-RELATIVE CHANGE FILTERING

Aligned versions provide an additional signal: relations that differ across candidates. Let $\mathcal{C}_g = \{d^a, d^b\}$ contain equal-size, structurally aligned versions with a nonempty set of shared row–column identities $\mathcal{W}_g$. For each component $u \in \{v, r, c\}$, we average the two normalized predicted boxes and apply the same box $\bar{b}^u_w$ to both images. This common comparison region reduces coordinate-jitter effects. Let $G(d, \bar{b})$ convert the crop to grayscale and resize it to 192×64 pixels. We measure change by

$$\delta^u_w = \frac{1}{192 \cdot 64} \sum_x \mathbf{1}\left[\left|G(d^a, \bar{b}^u_w)_x - G(d^b, \bar{b}^u_w)_x\right| \geq 20\right], \qquad \delta_w = \max_{u \in \{v,r,c\}} \delta^u_w. \tag{8}$$

Taking the maximum captures change in either the value or its identifying labels. A robust within-pair threshold defines the change index:

$$\tau = \max\big(\epsilon, \text{median}_{w \in \mathcal{W}_g}\, \delta_w + \lambda\, \text{MAD}_{w \in \mathcal{W}_g}\, \delta_w\big), \qquad \mathcal{D}_g = \{w \in \mathcal{W}_g : \delta_w \geq \tau\}. \tag{9}$$

Here $\epsilon = 0.002$, $\lambda = 3$, and MAD is the median absolute deviation. The median sets the pair-specific change level, MAD scales its dispersion, and ϵ supplies a minimum change threshold. An empty thresholded set is replaced by up to six witnesses with the largest δ_w, ensuring a nonempty index when $\mathcal{W}_g$ is nonempty. In sparse storage, structure and visual change determine which tiles are encoded and cached. Each retained identity indexes both candidates' own tiles, preserving their distinct content for relevance scoring.

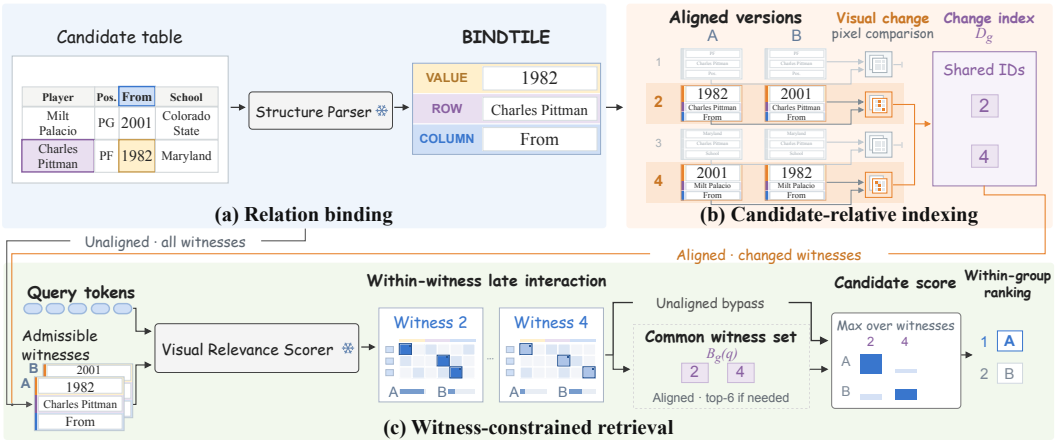


Figure 2: **DeltaBind: relation binding with candidate-relative indexing.** (a) Frozen structure prediction groups value, row, and column crops into a relation tile. (b) Pixel comparison of aligned versions selects changed identities $\mathcal{D}_g$, preserving each candidate's own content. (c) Query tokens interact within each tile. Aligned candidates maximize over the common set $\mathcal{B}_g(q)$, while unaligned candidates use all their own relations. The schematic shows two changed IDs; top-six selection is applied when more than six remain.

Focusing the candidate comparison. Suppose the highest unchanged-witness score is the same value m in both versions and exceeds both changed-witness maxima. All-relation max scoring then ties the candidates at m, even if their changed-witness maxima differ. Restricting the comparison to $\mathcal{D}_g$ makes those maxima determine the ranking. The aligned mode therefore focuses on the relations that can distinguish versions for the query.

At query time, witnesses in the change index are prioritized by their best relevance score across candidates:

$$\rho_w(q) = \max_{d \in \mathcal{C}_g} r(q, T(d, w)), \qquad \mathcal{B}_g(q) = \mathrm{TopK}_{w \in \mathcal{D}_g}\, \rho_w(q). \tag{10}$$

We use $K = 6$ and retain all witnesses when $|\mathcal{D}_g| \le K$. Both candidates are scored on these common identities:

$$S_\Delta(q, d \mid \mathcal{C}_g) = \max_{w \in \mathcal{B}_g(q)} r(q, T(d, w)). \tag{11}$$

The common set aligns the evidence available to the comparison. Each candidate retains its own pixels and selects its own maximum. The change index supplies the admissible relations; relevance selects the strongest query match among them.

3.4 Retrieval and Computational Cost

Let $\mathcal{A}_d(q) = \mathcal{W}(d)$ in relation-binding mode and $\mathcal{A}_d(q) = \mathcal{B}_g(q)$ in aligned mode. Use $S = S_{\mathrm{rel}}$ or $S = S_\Delta$, respectively. The top-ranked candidate and its selected witness are

$$\widehat{d} \in \arg\max_{d \in \mathcal{C}_g} S(q, d), \qquad \widehat{w} \in \arg\max_{w \in \mathcal{A}_{\widehat{d}}(q)} r(q, T(\widehat{d}, w)), \tag{12}$$

with the group conditioning of S_Δ implicit in this notation. The same scores rank the remaining candidates. The selected witness retains its source row, column, and crop coordinates, allowing the retrieved relation to be located in the original table. For integration into a larger ranking, the rank-slot procedure reorders the positions occupied by group members while preserving all other positions.

Offline storage and online scoring. For $M_d = |\mathcal{W}(d)|$ witnesses and P embeddings per tile, unaligned storage is $O(P \sum_d M_d)$ and query scoring is $O(LP \sum_d M_d)$, treating the embedding dimension as fixed. Sparse storage for an aligned pair keeps both candidates' tiles for each identity in $\mathcal{D}_g$, giving $O(P|\mathcal{D}_g|)$ storage and $O(LP|\mathcal{D}_g|)$ scoring. Encoding is performed once per retained

tile and reused across queries. This separates the cost of building an index from the cost of consulting it.

The role of top-K. Computing ρ_w scores the changed witnesses before top-K selection. Thus the change threshold controls storage and scoring, while K controls final aggregation. For fixed scores and $K \geq 1$, the selected set preserves a unique changed-set winner. The globally highest tile score also maximizes ρ_w, so top-K retains a witness attaining that score. A unique winning candidate therefore keeps its maximum, while every other candidate remains below it. Tied maxima can change under truncation; Appendix E gives the full derivation. Sparse caching and full storage implement this same filtered rule, compared in Sec. 4.5.

4 EXPERIMENTS

Our experiments examine discrimination under content-preserving swaps, retrieval over mined unedited tables, relation components and candidate aggregation, and sparse-index efficiency. A scientific-claim evaluation examines the method beyond atomic queries.

4.1 SETUP

Datasets. Table 1 summarizes the four evaluation regimes. **TabFact-HumanBind** uses human-written entailed claims from TabFact (Chen et al., 2020), with one table and query per source article. Each negative swaps the matched value with a different value in the same column. The candidates therefore contain the same page token multiset but differ in row–value binding. Both are standardized table renderings. **FinTabNet** (Zheng et al., 2021) and **PubTables-1M** (Smock et al., 2022) use unedited crops mined from the `test` portion of local public-data copies. Mining selects same-source tables with overlapping query words but different bindings; their positive and hard-negative tables form the candidate manifest. We split these sources into our own train/calibration/sealed sets and rank every same-source table in the manifest. Atomic queries use public annotations during construction; retrieval uses predicted structure. **SciClaimEval** (Ho et al., 2026) supplies authentic scientific claims and modified evidence for testing the relation-local assumption.

Table 1: **Evaluation regimes.** Natural test rows are our sealed source splits; candidate counts refer to the mined manifests. SciClaimEval uses development data.

Dataset	Queries	Images	Sources	Cand./q	Query / image
TabFact-HumanBind	256	512	256	2	human / controlled
FinTabNet test	41	53	14	2–8	template / unedited
PubTables test	71	84	38	2–4	template / unedited
SciClaimEval dev	215	430	81	2	authentic / modified

Baselines and scorers. We compare the released ColQwen2-v1.0 checkpoint with its matched full-image MaxSim baseline, and repeat the comparison with ColSmol-500M. OCR–BM25 uses predicted structure and PP-OCR (Du et al., 2020) version 4: flat OCR scores page text, relation OCR scores every relation, and filtered relation OCR uses the change index. On natural sets, DELTABIND uses relation binding alone; aligned controlled pairs also use change filtering. All relation-based systems use frozen models and predicted structure. Full-image and tile-based visual retrieval use the same checkpoint; ColSmol and OCR–BM25 test the pattern across backbones and lexical matching.

Qwen2.5-VL-7B-Instruct (Bai et al., 2025) compares full images in both A/B orders. It sees the controlled pair directly and reranks flat ColQwen2's top two natural candidates. Top-2 positive coverage is 75.6% on FinTabNet and 100% on PubTables. The reranker combines next-token A/B logits across both orders and leaves candidates below the top two in place. Appendices A–B provide dataset construction and implementation details.

Metrics. Controlled evaluation reports strict pair accuracy, counting zero margins as errors. Natural retrieval reports R@1 over all same-source manifest candidates and pair accuracy against a designated hard negative. OCR R@1 also counts ties as errors.

4.2 CONTROLLED RELATION COLLISIONS

Table 2 tests retrieval when word content is held constant. Flat ColQwen2 reaches 51.6% strict accuracy, while DELTABIND reaches 79.7%, a gain of 28.1 percentage points. The ColSmol replication improves from 44.9% to 85.2%. Thus the combined binding-and-filtering pipeline helps both visual encoders distinguish candidates with the same visible tokens.

Table 2: **TabFact-HumanBind strict pair accuracy (%).** †marks post-hoc replications or diagnostics. Qwen is a bidirectional full-page A/B verifier; all ties count as wrong. Gains use unrounded outcomes.

Scorer	Evidence unit	Accuracy	Gain vs. flat
ColQwen2	full page	51.6	–
ColQwen2	filtered relation tiles	79.7	+28.1
ColSmol-500M†	full page	44.9	–
ColSmol-500M†	filtered relation tiles	**85.2**	+40.2
Qwen2.5-VL-7B	full-page verifier	61.3	–
OCR–BM25†	full page	4.3	–
OCR–BM25†	relation tiles (text)	68.4	+64.1
OCR–BM25†	filtered tiles (text)	84.8	+80.5

Relation-bound ColQwen2 also outperforms the full-page Qwen verifier, which obtains 61.3% strict accuracy. The resulting 18.4-point advantage shows the effectiveness of compact relation evidence in this comparison. The lexical control holds the scorer family fixed: moving from page OCR to relation OCR raises accuracy from 4.3% to 68.4%, and change filtering further raises it to 84.8%. Flat OCR ties on 89.1% of pairs: page-level lexical matching largely loses the row assignment changed by the swap. Relation OCR restores the association between each value and its labels, while filtering directs the comparison to the altered relations. The progression from 4.3% to 68.4% to 84.8% separates the contribution of relation context from the additional benefit of candidate differences within the same lexical scorer.

A source-annotation audit identifies 12 of 256 pairs with duplicate support for the normalized triple: another row still matches it after the swap. The full-set results above include these cases. On the 244 unique-relation cases, ColQwen2 improves from 51.2% to 82.0% and ColSmol from 45.5% to 86.5%. The gains persist when the requested triple distinguishes the pair unambiguously.

4.3 NATURAL SOURCE-LOCAL RETRIEVAL

On unedited FinTabNet tables, relation binding raises ColQwen2 R@1 from 46.3% to 80.5%; on PubTables-1M it raises R@1 from 49.3% to 85.9% (Table 3). The improvements are 34.1 and 36.6 percentage points, respectively. The method scores all predicted relations in each table. Candidates can have different grids, and their strongest query matches can occur at different row and column positions. This setting tests relation binding on the tables' original content and layout.

The two natural metrics distinguish full-pool ranking from the designated hard-negative comparison. ColQwen2 reaches 82.9% pair accuracy and 80.5% R@1 on FinTabNet, and 85.9% on both metrics on PubTables. The gains therefore extend from the mined pair to competition with every table in the same-source candidate pool.

ColSmol gains 12.2 points on FinTabNet and 31.0 on PubTables. Relation OCR–BM25 reaches 65.9% and 85.9% R@1, respectively, versus 46.3% and 54.9% for flat OCR. Relation context thus benefits both visual and lexical retrieval, with ColQwen2 giving the strongest visual results.

On PubTables, visual and lexical relation scoring have the same aggregate R@1 but each answers nine queries missed by the other. On FinTabNet, visual scoring alone answers nine and OCR alone three. The scorers therefore succeed on partly different queries despite using the same relation representation. Equal aggregate accuracy on PubTables conceals this difference in which queries the two scorers resolve correctly.

Table 3: **Natural same-source retrieval.** Values are R@1 / designated hard-negative pair accuracy (%). †denotes a post-hoc replication or diagnostic. OCR R@1 requires a strict win. Qwen reranks the flat visual top two; other rows score every same-source manifest candidate.

Scorer	Evidence unit	FinTabNet	PubTables-1M
ColQwen2	full image	46.3 / 53.7	49.3 / 52.1
ColQwen2	relation tiles	**80.5 / 82.9**	**85.9 / 85.9**
ColSmol-500M†	full image	46.3 / 56.1	52.1 / 56.3
ColSmol-500M†	relation tiles	58.5 / 65.9	83.1 / 84.5
Qwen2.5-VL-7B†	top-2 full images	51.2 / 58.5	69.0 / 69.0
OCR–BM25†	full image text	46.3 / 56.1	54.9 / 56.3
OCR–BM25†	relation tiles (text)	65.9 / 70.7	**85.9 / 85.9**

4.4 What makes relation binding work?

Complete relation context. Table 4 compares the three components on FinTabNet. A value crop alone obtains 29.3% R@1, below the full-image baseline. Adding either row or column context raises R@1 to 53.7%, and including both reaches 80.5%. The complete tile gains 26.8 points over either partial-context variant, showing the value of presenting row and column context together.

Table 4: **Relation component ablation on FinTabNet.** All rows use ColQwen2, predicted structure, and the full natural-language query.

Evidence in each tile	R@1	Pair acc.
Value only	29.3	41.5
Value + row	53.7	68.3
Value + column	53.7	68.3
Value + row + column	**80.5**	**82.9**

The two partial tiles leave different ambiguities. Value–row evidence identifies the entity but may omit the metric or period; value–column evidence identifies the field but may omit the entity. A repeated value can therefore match the wrong relation in either case. The complete tile provides both identifiers, improving full-pool ranking and hard-negative discrimination under the same scoring rule.

Candidate aggregation. With the same changed witnesses and ColQwen2 scores on TabFact-HumanBind, max-per-candidate accuracy is 79.7%. Mean signed score differences, the signed score difference with the largest absolute magnitude, and the mean of the two largest-magnitude signed differences each obtain 45.7%. The distinction can be expressed directly in the candidate margin. Writing $r_w^a = r(q, T(d^a, w))$ and $r_w^b = r(q, T(d^b, w))$ on the common selected set $\mathcal{B} = \mathcal{B}_g(q)$ gives

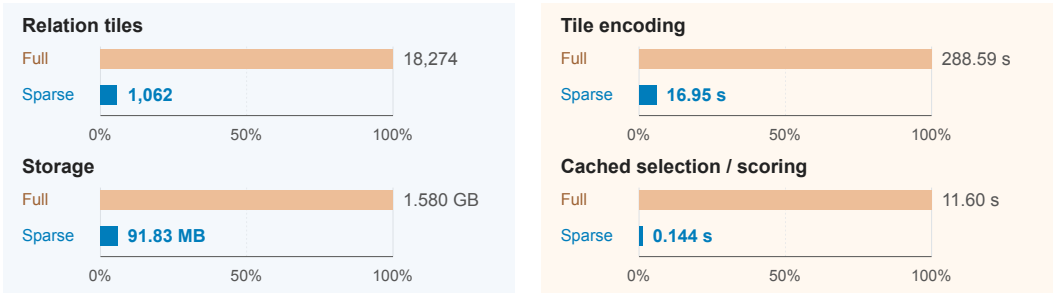
$$\Delta_{\max} = \max_{w \in \mathcal{B}} r_w^a - \max_{w \in \mathcal{B}} r_w^b, \qquad \Delta_{\text{mean}} = \frac{1}{|\mathcal{B}|} \sum_{w \in \mathcal{B}} (r_w^a - r_w^b). \tag{13}$$

The first compares each candidate's strongest relation; the second averages signed evidence across changed relations. A query may concern only one of those changes, making its strongest match the relevant comparison. Mean aggregation lets scores on other changed relations offset the relation sought by the query. Selecting the largest absolute difference instead prioritizes contrast between candidates, which can arise on a tile with little query relevance. Max-per-candidate keeps each candidate's strongest query match as the decision criterion, matching its advantage in this controlled comparison.

4.5 Change filtering and efficiency

Across the 512 controlled images, the full index stores 18,274 tiles (1.580 GB). Caching only the change index retains 1,062 tiles (5.81%, 91.83 MB) and reproduces every binary decision of the

full-storage implementation of the same filtered rule. Figure 3 reports encoding times of 288.59 and 16.95 seconds and cached CPU-loop times of 11.60 and 0.144 seconds. The full loop scores all tiles, the page baseline, and aggregation controls, and computes changes online. The sparse loop scores retained tiles using a prebuilt change index. The timing difference therefore combines fewer tiles, offline change detection, and a narrower scoring workload. Appendix C.3 specifies the timing boundaries.



(a) Index size **(b) Measured cost**

Figure 3: **Index size and recorded implementation costs.** Both indexes preserve the same filtered decisions on TabFact-HumanBind. CPU-loop workloads differ as described in the text; sparse timing excludes offline change-index construction. Bars are normalized within each metric.

4.6 Scope beyond relation-local retrieval

The controlled and natural results concern candidate comparisons for which one explicit relation can provide decisive evidence. SciClaimEval tests a broader claim class: its scientific statements can require aggregation, discourse context, or implicit reasoning. On 215 task-2 development pairs, flat ColQwen2 obtains 56.3% strict accuracy and change-filtered top-six DeltaBind 54.0% (116/215), a change of -2.3 percentage points. Here, a value may need to be interpreted with another quantity, an aggregate over cells, or surrounding prose. Such comparisons require combining evidence across relations; maximizing over individual tiles provides only the strongest local match.

5 Limitations

Evaluation scope. Natural evaluation uses mined same-source candidate pools and templated queries; controlled evaluation uses standardized TabFact renderings. FinTabNet has 41 queries from 14 reports, with possible dependence among queries sharing a source. Open-world search, automatic group discovery, and multi-relation claims extend beyond this evaluation setting.

Structural assumptions. The parser uses a row stub and one header row, making complex spanning headers, transposed layouts, and detection errors challenging. The frozen TATR-v1.1-All checkpoint was pretrained on PubTables-1M and FinTabNet.c; our source splits describe the retrieval evaluation. Aligned filtering assumes comparable geometry and a query-relevant change; scan noise and layout shifts can affect its pixel statistic. The component analysis measures the combined tile design, including its crop layout, role labels, and resolution. A selected witness is a retrieval result whose semantic support depends on the parser and scorer.

6 Conclusion

We presented DeltaBind, a training-free framework that binds values to row and column context before candidate-level aggregation. Relation tiles preserve the context needed to distinguish bindings, while the candidate-relative index focuses aligned comparisons on changed relations. The resulting framework improves retrieval across controlled swaps and mined unedited tables with both visual and lexical scorers. Sparse caching retains 5.8% of controlled relation tiles while preserving the full-storage filtered decisions. These findings establish complete table relations as an effective unit for fine-grained candidate comparison with frozen retrieval models.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript, including correcting grammatical errors and improving clarity and readability. We have reviewed AI-assisted work for accuracy and consistency. We take responsibility for the final content of this work, including the text, claims, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

The study uses existing public research datasets and does not collect personal or human-subject data. Counterfactual tables are used only to measure retrieval behavior and are clearly separated from original records. A practical risk is that a retrieval system may surface a visually plausible but incorrect table; the method should therefore be used as evidence selection, not as an autonomous factual authority. Dataset licenses and redistribution conditions must be respected when releasing processed images.

REPRODUCIBILITY STATEMENT

Appendices A and B describe dataset construction, evaluation protocols, model identifiers, thresholds, prompting, and statistical tests. Appendix C gives extended numerical results. These appendices are included in this PDF; no task-specific model training is required.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. TabFact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkeJRhNYDH.

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 the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30971–31005, 2025. doi: 10.18653/v1/2025.emnlp-main.1576. URL https://aclanthology.org/2025.emnlp-main.1576/.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020. URL https://arxiv.org/abs/2009.09941.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, pp. 61424–61449, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/99e9e141aafc314f76b0ca3dd66898b3-Paper-Conference.pdf.

Xanh Ho, Sunisth Kumar, Yun-Ang Wu, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Table-text alignment: Explaining claim verification against tables in scientific papers. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2509–2517, 2025. doi: 10.18653/v1/2025.findings-emnlp.135. URL https://aclanthology.org/2025.findings-emnlp.135/.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Tian Cheng Xia, Florian Boudin, André Greiner-Petter, and Akiko Aizawa. SciClaimEval: Cross-modal claim verification in scientific papers. In *Proceedings*

of the Fifteenth Language Resources and Evaluation Conference, pp. 11060–11071, 2026. doi: 10.63317/4ap9rg2gnwmf. URL https://aclanthology.org/2026.lrec-1.864/.

Hyukkyu Kang, Injung Kim, and Wook-Shin Han. TRIAL: Token relations and importance aware late-interaction for accurate text retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16864–16877, 2025. doi: 10.18653/v1/2025.emnlp-main.854. URL https://aclanthology.org/2025.emnlp-main.854/.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Da Li, Keping Bi, Jiafeng Guo, and Xueqi Cheng. Tailoring table retrieval from a field-aware hybrid matching perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27693–27704, 2025. doi: 10.18653/v1/2025.emnlp-main.1409. URL https://aclanthology.org/2025.emnlp-main.1409/.

Da Li, Keping Bi, Jiafeng Guo, Wei Yuan, Fang Yang, Tingting Gao, and Xueqi Cheng. Beyond text-only: Towards multimodal table retrieval in open-world. In *International Conference on Learning Representations*, pp. 71279–71293, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7422317f84e8c83e4c1ad2ff87e1e88e-Paper-Conference.pdf.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7787–7813, 2023. doi: 10.18653/v1/2023.emnlp-main.483. URL https://aclanthology.org/2023.emnlp-main.483/.

Averi Nowak, Francesco Piccinno, and Yasemin Altun. Multimodal chart retrieval: A comparison of text, table and image based approaches. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5488–5505, 2024. doi: 10.18653/v1/2024.naacl-long.307. URL https://aclanthology.org/2024.naacl-long.307/.

Doyoung Park, Naresh Reddy Yarram, Sunjin Kim, Minkyu Kim, Seongho Cho, and Taehee Lee. Text change detection in multilingual documents using image comparison. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 5218–5227, 2025a. doi: 10.1109/WACV61041.2025.00510. URL https://doi.org/10.1109/WACV61041.2025.00510.

Sungho Park, Joohyung Yun, Jongwuk Lee, and Wook-Shin Han. HELIOS: Harmonizing early fusion, late fusion, and LLM reasoning for multi-granular table-text retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32424–32444, 2025b. doi: 10.18653/v1/2025.acl-long.1559. URL https://aclanthology.org/2025.acl-long.1559/.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019. URL https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf.

Brandon Smock, Rohith Pesala, and Robin Abraham. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4634–4642, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Smock_PubTables-1M_Towards_Comprehensive_Table_Extraction_From_Unstructured_Documents_CVPR_2022_paper.html.

Brandon Smock, Rohith Pesala, and Robin Abraham. Aligning benchmark datasets for table structure recognition. In *Document Analysis and Recognition - ICDAR 2023*, pp. 371–386. Springer, 2023. doi: 10.1007/978-3-031-41734-4_23. URL https://link.springer.com/chapter/10.1007/978-3-031-41734-4_23.

Omri Uzan, Asaf Yehudai, Roi Pony, Eyal Shnarch, and Ariel Gera. Guided query refinement: Multimodal hybrid retrieval with test-time optimization. In *International Conference on Learning Representations*, pp. 495–522, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/015d6a156390d23c19d17651d60b3db6-Paper-Conference.pdf.

Zhiruo Wang, Zhengbao Jiang, Eric Nyberg, and Graham Neubig. Table retrieval may not necessitate table-specific model design. In *Proceedings of the Workshop on Structured and Unstructured Knowledge Integration (SUKI)*, pp. 36–46, 2022. doi: 10.18653/v1/2022.suki-1.5. URL https://aclanthology.org/2022.suki-1.5/.

Zirui Wu, Nan Hu, and Yansong Feng. Enhancing structured evidence extraction for fact verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6631–6641, 2023. doi: 10.18653/v1/2023.emnlp-main.409. URL https://aclanthology.org/2023.emnlp-main.409/.

Zhuoyan Xu, Haoyang Fang, Boran Han, Bonan Min, Bernie Wang, Cuixiong Hu, and Shuai Zhang. Efficient table retrieval and understanding with multimodal large language models. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4327–4340, 2026. doi: 10.18653/v1/2026.findings-eacl.226. URL https://aclanthology.org/2026.findings-eacl.226/.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, pp. 21074–21098, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3640a1997a4c9571cea9db2c82e1fc35-Paper-Conference.pdf.

Joohyung Yun, Doyup Lee, and Wook-Shin Han. LILaC: Late interacting in layered component graph for open-domain multimodal multihop retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20540–20559, 2025. doi: 10.18653/v1/2025.emnlp-main.1037. URL https://aclanthology.org/2025.emnlp-main.1037/.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 697–706, 2021. doi: 10.1109/WACV48630.2021.00074. URL https://doi.org/10.1109/WACV48630.2021.00074.

Mingyang Zhou, Lingyu Zhang, Sophia Horng, Maximillian Chen, Kung-Hsiang Huang, and Shih-Fu Chang. M^2-TabFact: Multi-document multi-modal fact verification with visual and textual representations of tabular data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26239–26256, 2025. doi: 10.18653/v1/2025.findings-acl.1345. URL https://aclanthology.org/2025.findings-acl.1345/.

Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, and Zhuohan Xie. FinCARDS: Card-based analyst reranking for financial document question answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24836–24852, 2026. doi: 10.18653/v1/2026.findings-acl.1244. URL https://aclanthology.org/2026.findings-acl.1244/.

A Frozen protocols and data construction

A.1 TabFact-HumanBind

The final controlled protocol was frozen on 4 September 2026 before the final dataset was built or scored. Its source is the official TabFact train split. All Wikipedia article identifiers used by earlier development, validation, and confirmation sets were excluded; selection was lexicographic after exclusions, with at most one table per article. Tables with more than 12 data rows or eight columns

were rejected. For each retained table, the construction selected an official human-written entailed statement with a conservatively matched row, column, and numeric value. The counterfactual exchanged the value with a different row in the same column, keeping the whole-page token multiset invariant. The final set contains 256 articles, 256 queries, and 512 images. The metadata records 126 duplicate-article, 902 excluded-article, 3,541 relation-match, and 4,298 shape rejections.

Both candidates had to yield valid predicted structure. This eligibility test occurred before retrieval and did not depend on the outcome. The detector threshold was 0.3, target-anchor metadata was disabled, and the retrieval settings were fixed to minimum pixel difference 0.002, MAD multiplier 3, per-pixel grayscale threshold 20, comparison size 192×64, changed-witness budget six, and max aggregation. The primary endpoints were strict pair accuracy relative to flat ColQwen2 and a full-page Qwen2.5-VL-7B verifier. We report effect sizes and two-sided exact paired binary tests for both pre-specified comparisons.

A.2 Natural table sets

Construction starts from the `test` portion of the local public-data copies. A table must provide an extractable row label, column label, and numeric value; PubTables additionally requires a regular rectangular grid without merged cells. Mining identifies same-source positive and hard-negative tables that contain the query words with different bindings, retaining at most one question per positive table. The candidate manifest contains the union of positive and hard-negative tables involved in these questions. Retrieval compares all manifest tables from the query's source. Thus the pool is defined by this construction, rather than an inventory of every table in the original report or paper.

For FinTabNet, a deterministic source-report assignment with seed `binding-collision-v1` partitions 90 reports into 63 train, 13 calibration, and 14 sealed reports. The sealed split contains 41 queries and 53 distinct table crops. The exact BindTile-Max protocol, including all same-report manifest candidates, predicted TATR structure at threshold 0.3, no target anchor, and full natural-language queries, was frozen before inspecting sealed results. Candidate sets contain between two and eight tables.

For PubTables-1M, seed `pubtables-controlled-v1` partitions 254 source papers into 178 train, 38 calibration, and 38 sealed papers. The sealed split contains 71 queries and 84 embedded table images, with two to four candidate tables per source. The confirmatory result uses predicted TATR structure; an earlier protocol with oracle OTSL boxes is not used as main evidence. For both natural datasets, labels and query templates are built from public annotations, but the inference pipeline sees only the query, source pixels, and source-group identifier. Our train/calibration/sealed partitions are source splits of these mined manifests, distinct from the public datasets' original partitions. Source separation applies within this evaluation and makes no claim about model pretraining overlap. All tables in both natural sealed manifests have valid predicted structure and nonempty tile sets; tile availability removes no additional candidates at this stage.

A.3 SciClaimEval

SciClaimEval task 2 development contains 215 evidence pairs from 81 source papers in the evaluated table subset. The recorded protocol is `source_pair_changed_aligned_bindtile_late_interaction`, with $K = 6$ and max-per-candidate aggregation. This run computes relevance for all tiles, restricts the admissible identities by pixel change, and retains at most six according to query relevance. If the thresholded set is empty, the largest pixel changes supply the fallback set. The final decision uses each candidate's maximum on the selected identities, not its unrestricted all-tile maximum. The 215 query records retain between one and six relations; 116 strict wins give $116/215 = 53.95348837\%$, reported as 54.0%. This development evaluation examines performance beyond relation-local claims.

B Implementation details

Checkpoints. The primary retriever is the local snapshot of `vidore/colqwen2-v1.0`; the post-hoc replication uses `vidore/colSmol-500M`. The structure checkpoint is `microsoft/table-transformer-structure-recognition-v1.1-all`, pretrained on PubTables-

1M and FinTabNet.c (Smock et al., 2023). We use its released weights without further fine-tuning. The generative baseline is Qwen2.5-VL-7B-Instruct (Bai et al., 2025). OCR uses `rapidocr-onnxruntime==1.4.4` with PP-OCRv4 detector and recognizer ONNX files and the mobile-v2 direction classifier. The version-specific architecture is documented in the official PP-OCRv4 technical report.

Predicted structure and tile implementation. The frozen recognizer uses confidence threshold 0.3. All coordinates are normalized to $[0, 1]^2$. Axis detections are processed in descending confidence order and retained when their one-dimensional IoU with every retained detection is below 0.7; retained boxes are sorted spatially. The header row is the last row whose center lies in the predicted header region, or the first detected row if no such region is found. The first detected column is treated as the row stub. Data rows below the header and non-stub columns define witnesses by Eq. 4. Each component crop receives four source-pixel padding units and is enlarged isotropically using Lanczos interpolation to fit a 396×68 content area. The tile is a 512×240 white canvas with three 80-pixel bands and a 104-pixel role-label column, labeled VALUE, ROW, and COLUMN. No title band is used.

Change statistic and fixed settings. Equation 8 uses grayscale crops resized by bilinear interpolation. For each component, the two normalized boxes are averaged and the common box is applied to both source images before comparison. Equation 9 uses $\epsilon = 0.002$ and $\lambda = 3$. An empty thresholded set is replaced by the up-to-six witnesses with largest δ_w; query-time prioritization uses $K = 6$. The robust deviation is $\mathrm{MAD}_w\, \delta_w = \mathrm{median}_w\, |\delta_w - \mathrm{median}_{w'}\, \delta_{w'}|$. These settings were fixed before the final controlled evaluation.

Scoring implementation. Visual embeddings are used as produced by each released checkpoint. The flat baseline uses official sum-MaxSim. Equation 5 expresses mean-MaxSim across query tokens. The natural visual retrieval script divides this tile score by query length once more, using r_{vis}/L. This positive query-wise factor preserves tile and candidate rankings. It changes the relative scale in any additive fusion with full-image scores, so rankings from score fusion depend on the implementation's normalization. PP-OCRv4 strings are assigned to predicted component boxes by geometry and concatenated into value, row, and column fields. Text is tokenized by `casefold()` followed by the regular expression `[a-z0-9]+`, without stemming or stop-word removal. BM25 uses $k_1 = 1.2$ and $b = 0.75$. Page-level IDF is computed over every page in the input manifest; relation-level IDF uses every relation text from those pages, across sources. DeltaBind-Lex shares this relation corpus and its IDF with unfiltered relation BM25, changing only the admissible relation set. Both visual and lexical paths use predicted structure, not gold cells or text.

Rank-slot integration. In relation-binding mode, the group defines the ranking pool and each candidate's score depends on its own tiles. In aligned mode, the group also defines which witnesses are admissible: the same table can be compared through different relations when paired with a different version.

Given a base ranking π, a supplied group $\mathcal{C}_g$, and an eligibility flag, the implementation permutes only the positions already occupied by group members. If the flag is false or fewer than two members are present, it returns π unchanged. This integration routine uses stable sorting on ties; the natural retrieval scripts' tie behavior is specified below. Positions outside the group are preserved by construction. This contract does not evaluate automatic group discovery, query eligibility, or end-to-end open-world retrieval.

Qwen verifier. The controlled and FinTabNet prompt was:

> You are comparing two candidate table images. Determine which table directly supports the claim. Check every comparison, row label, column header, metric, unit, and value; do not choose a table just because it shares words with the claim. Claim: {question}. Which table supports the claim? Answer only A or B.

The recorded PubTables prompt was:

> You are selecting which of two table images contains all the specific evidence needed to answer a question. Image A is shown first and Image B second. Compare row labels,

> column headers, dates, units, and values. Do not choose a table merely because it shares broad keywords.
>
> Question: {question}
>
> Which image contains the more complete and directly relevant evidence needed to answer the question? Answer only A or B.

For each image order, the verifier reads the next-token logit difference $\ell = \mathrm{logit}(A) - \mathrm{logit}(B)$. Its preference for the first candidate in the original order is $(\ell_{\mathrm{forward}} - \ell_{\mathrm{reverse}})/2$. The controlled metric counts zero preference as wrong. Natural retrieval reranks only the original ColQwen2 top two and preserves the positions of all remaining candidates. It uses these logits directly rather than voting over generated answers.

Strict ties and statistics. For the controlled pair, correctness is $\mathbf{1}[S(q, d^+) > S(q, d^-)]$. Natural visual R@1 uses PyTorch descending score order with one labeled relevant candidate and no explicitly requested stable tie rule. The primary ColQwen2 records have no top-score ties on either natural dataset. Natural OCR R@1 requires the positive to score strictly above every same-source manifest competitor; a tie is incorrect. Qwen has two zero preferences on FinTabNet and one on PubTables. Its sorting code does not explicitly guarantee preservation of the original top-two order on ties. For diagnostics only, tie-aware accuracy divides credit among all candidates whose scores match the maximum within absolute tolerance 10^{-8}. The exact test is the two-sided sign/McNemar test over discordant binary outcomes. For natural sets, these query-level tests do not account for dependence among queries from the same source. Gains use unrounded outcomes, so differences between displayed percentages can differ by 0.1 point.

Empty evidence sets. The retrieval scripts raise an error for invalid structure or missing scoreable tiles, with no automatic full-image fallback. Both natural sealed sets contain valid, nonempty tile sets throughout.

C Extended results

C.1 Effect sizes and exact tests

Table 5: **Controlled comparisons.** Gains are percentage points. Tests compare paired query outcomes; each source article has one query.

Comparison	Gain	Exact p
DeltaBind-ColQwen2 vs flat	+28.1	2.16×10^{-11}
DeltaBind-ColSmol† vs flat	+40.2	2.81×10^{-20}
DeltaBind-ColQwen2 vs Qwen-7B	+18.4	9.45×10^{-6}
DeltaBind-ColSmol† vs Qwen-7B	+23.8	4.39×10^{-10}
DeltaBind-Lex† vs Relation OCR	+16.4	5.12×10^{-12}

Table 6: **Natural comparisons against matched flat scorers.** Gains are query-level percentage-point differences. Exact tests operate on paired queries, without accounting for shared-source dependence.

Dataset	Relation scorer	Gain	Exact p
FinTabNet	ColQwen2	+34.1	5.19×10^{-4}
FinTabNet	ColSmol†	+12.2	.383
FinTabNet	OCR–BM25†	+19.5	.115
PubTables	ColQwen2	+36.6	1.29×10^{-5}
PubTables	ColSmol†	+31.0	1.95×10^{-4}
PubTables	OCR–BM25†	+31.0	2.98×10^{-6}

Comparisons to Qwen2.5-VL-7B. On FinTabNet, BindTile-ColQwen2 is 29.3 points above the top-2 Qwen reranker, with exact $p = .00183$. The comparison is bounded by Qwen's 75.6% top-2 positive coverage. On PubTables, the corresponding ColQwen2 difference is +16.9 points (exact $p = .0169$). ColSmol is +14.1 points, with exact $p = .0639$.

C.2 Controlled diagnostics

Table 7: **Aggregation ablation on TabFact-HumanBind.** Accuracy is strict. All non-flat rows use the same changed witnesses and ColQwen2 scores.

Candidate comparison	Accuracy (%)
Flat page MaxSim	51.6
Mean signed witness difference	45.7
Largest-absolute signed difference	45.7
Mean of top-two absolute differences	45.7
Max relevance per candidate	**79.7**

Let $\Delta_w = r(q, T(d^a, w)) - r(q, T(d^b, w))$ on the fixed admissible witness set. The signed-difference controls use $|\mathcal{B}_g(q)|^{-1} \sum_w \Delta_w$, Δ_{w_1}, and $(\Delta_{w_1} + \Delta_{w_2})/2$, respectively, where w_1, w_2 have the largest absolute score differences. Their sign selects a candidate and zero counts as a strict tie. Max-per-candidate instead compares the two candidates' maximum relevance scores. "Absolute difference" here refers to relevance-score differences, not the pixel-change statistic used to build the index.

Table 8: **Post-hoc controls using source HTML annotations.** These diagnostic controls have access to annotated table structure and text.

Control	Strict acc. (%)	Tie rate (%)
Whole-table HTML BM25	0.0	100.0
Gold-incidence relation BM25	78.1	21.5
Exact normalized triple oracle	95.3	4.7

The exact-triple audit identifies the same 12 examples with duplicate triple support discussed in Sec. 4. On the unique subset of 244, Qwen's strict score is 61.9%. DeltaBind-ColQwen2 improves over its flat counterpart by 30.7 points; DeltaBind-ColSmol improves by 41.0 points. This post-hoc subset analysis isolates cases with unique normalized-triple support; the primary endpoint uses all 256 pairs.

C.3 Efficiency measurements

Table 9: **Aligned controlled index measurements.** Historical encoding times and cached CPU-loop times have the boundaries described below. The two CPU loops execute different workloads.

Index	Tiles	Storage	Encode / CPU loop (s)
All Relation Tiles	18,274	1.580 GB	288.59 / 11.60
Changed Relation Tiles	1,062	91.83 MB	16.95 / 0.144

The sparse index retains 5.8115% of tiles and reproduces every binary decision of the full-storage implementation of the same change-filtered rule. The recorded encoding times (288.59 and 16.95 seconds) include image reading, preprocessing, model forward passes, and transfer of output vectors to CPU. They exclude model loading, tile generation, and cache-file writes.

Both scoring timers start after cache loading. Cached vectors remain on CPU during scoring. The 11.60-second full loop includes page-baseline scoring, all-tile scoring, candidate-image reads, online pixel-change computation, selection, and multiple aggregation rules. The 0.144-second sparse loop scores only retained tiles and reads the prebuilt change index. Its timing excludes offline index construction. These recorded loop costs therefore combine tile reduction, moving change detection

offline, and removal of the additional scoring work. The components were not timed separately, and the original hardware model and actual batch size are not recoverable from the available records. These values characterize the recorded implementations; they are not GPU end-to-end inference latencies.

OCR of all 512 controlled pages took 168.22 seconds with eight workers; lexical indexing and scoring took 0.982 seconds. The Qwen verifier required 411.99 seconds for 512 newly scored ordered pairs; this workload differs from cached retrieval scoring.

D DIAGNOSTIC INTERPRETATION

Fixed operating point. The aligned protocol fixes the threshold, MAD multiplier, fallback, and witness budget specified in Appendix B before confirmation. The threshold and fallback determine the indexed witness set; top-K controls its final aggregation. Appendix E establishes the unique-winner property of top-K for fixed relevance scores.

Visual and lexical complementarity. On PubTables, OCR-only and visual-only correctness counts are both nine; on FinTabNet they are three and nine. Label-based oracle selection yields union scores of 98.6% and 87.8%, respectively. These oracle upper bounds quantify the complementary successes available to a potential visual–lexical selector.

Witness interpretation. The selected witness identifies the tile that determines the candidate score, and its source coordinates enable inspection of the value and context. The reported metrics evaluate candidate ranking. Establishing the witness's semantic support requires checking its content against the query.

E PROPERTIES OF THE SCORING RULE

Shared-witness score gap. Let $m_i = \max_w \phi_i(w)$ under the finite, nonempty evidence-set assumptions in Sec. 3.1. Subtracting $\max_w \sum_i \phi_i(w)$ from $\sum_i m_i$ and moving the constant inside the minimum gives Eq. 3. Each summand is nonnegative, so the gap is nonnegative. It is zero exactly when some witness satisfies $\phi_i(w) = m_i$ for every token i. This proves both the inequality and its equality condition. The derivation compares aggregation over the same fixed evidence vectors; separate page and tile encodings also change the representations.

Top-K and the winning candidate. Denote the tile scores by $r_{dw} = r(q, T(d, w))$. Their maximum across candidates and changed witnesses is $M = \max_d \max_{w \in \mathcal{D}_g} r_{dw}$. Because $\rho_w = \max_d r_{dw}$, every nonempty top-K set under ρ contains at least one witness $w^\star$ for which $\rho_{w^\star} = M$. If the full changed-set comparison has a unique winning candidate $d^\star$, only that candidate can attain M. It therefore still attains M on the selected set, while every other candidate remains below M. The winner is preserved, although other scores and the winning margin may change. With ties, different candidates may attain M at different witnesses; truncation can remove one of these witnesses and break the tie. This argument holds for a fixed change set and fixed relevance scores.

Coherence, relevance, and support. The within-tile constraint determines which evidence can be combined, while the scorer evaluates its relevance to the query. Numerical equality, negation, units, and clause-level support depend on that semantic evaluation. Aligned candidates share an admissible set of relation identities, and each selects its own maximizing witness. Their ranking therefore compares each candidate's strongest admissible evidence, as expressed by $\Delta_{\max}$ in Eq. 13. This separates relation coherence, candidate ranking, and semantic support.