What the Reranker Sees: Multi-Aspect Page Annotation for Long-Document Multimodal Question Answering

Anonymous ACL submission

Abstract

Long-document visual question answering (VQA) over documents of tens to hundreds of pages mixing text, tables, charts, and figures typically follows retrieve-then-read pipelines. In our setting, the bottleneck shifts from retrieval recall to reranker-side evidence selection: on MMLongBench-Doc, BGE-M3 reaches Recall@20=0.86 but only F1@5=0.254, and even the visual retriever ColPali reaches only F1@5=0.332—a text-only rerank LLM seeing only raw snippets misses table, chart, and layout evidence even when the upstream retriever encoded images. We propose TRIDENT, with two complementary components: TRIDENT-R, a retriever-agnostic LLM reranker that converts each candidate into an LLM-readable semantic record—visual caption, section path, entity tags, multi-axis concept hits, and a text snippet—then performs a single adaptive-K rerank call; and TRIDENT-S, a generation-side module that prompts the VLM under topical, entity, and structural lenses before synthesis. On two long-document datasets, the annotation+rerank protocol substantially improves retrieval F1 across five heterogeneous pools, with every reranked pool exceeding the strongest adaptive-K baseline PageIndex. An LLM rerank without the annotation barely changes first-hit ranking, indicating the lift comes from the structured annotation. TRIDENT-S targets open-ended synthesis questions by design, adding up to 6.6 points in generation accuracy on these questions. The best TRIDENT configuration is the strongest downstream QA pipeline in our evaluation, with rankings consistent across two LLM judges (κ=0.913).

1 Introduction

Long-document visual question answering (VQA) requires answering natural-language questions over documents of tens to hundreds of pages that mix text, tables, charts, and figures. Although frontier Vision Language Models (VLMs) accept very long contexts in principle, feeding entire documents on every query is expensive and can suffer from position-induced degradation (Liu et al., 2024). Thus, retrieve-then-read pipelines (Lewis et al., 2020) remain practical: a retriever selects a small set of pages, and a VLM answers the question conditioned on them. On MMLongBench-Doc, oracle evidence pages alone lift Large Vision Language Model (LVLM) F1 by 10–30 points over full-document inputs (Ma et al., 2024b), showing evidence selection is central to long-document VQA.

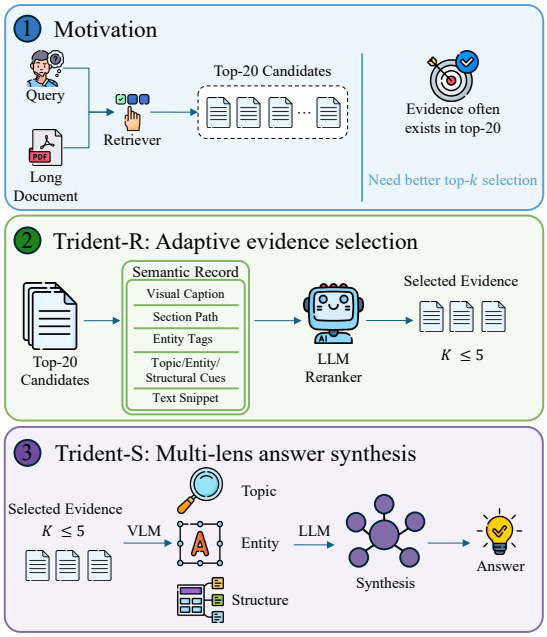

Figure 1: Overview of TRIDENT. Given a query and a long document, a retriever first returns a top-20 candidate pool where evidence pages often already appear. TRIDENT-R improves evidence selection by converting candidates into semantic records for adaptive LLM reranking, while TRIDENT-S synthesizes the final answer through topical, entity, and structural lenses.

Where does evidence selection fail? In our setting, BGE-M3 (Chen et al.) achieves Recall@20=0.86 on MMLongBench-Doc but only F1@5=0.254; the multimodal retriever ColPali (Faysse et al., 2025) improves this to

F1@5=0.332 but remains far from oracle. Evidence pages are typically present in the top-20 pool—the bottleneck is identifying the small evidence subset within 20 surface-similar candidates.

A natural next step is to insert an LLM reranker between the top-20 pool and the $K \leq 5$ evidence selection, an approach used by RankGPT (Sun et al., 2023) and increasingly common in long-document RAG. However, when this reranker only sees raw text snippets of each candidate page, it does not improve first-hit ranking on either pool: on BGE-M3, MRR is essentially unchanged ($0.524 \rightarrow 0.523$); on ColPali, MRR actually *drops* from 0.692 to 0.598, because the text-only LLM judgment overrides ColPali's stronger visual ranking signal. The reason is that long-document evidence is often locked in tables, charts, figures, and layout, all poorly represented in short text snippets. A rerank-stage LLM that only sees snippets is partially blind to exactly the evidence that matters most—ColPali's visual encoding occurs only at retrieval; the rerank LLM still consumes text-only candidate records and cannot access page images directly.

This rerank-stage blindness has been recognized: recent work addresses it by replacing the text LLM reranker with a multimodal LLM that directly sees page images (Chen et al., 2025; Xu et al., 2025). This is effective but expensive—every rerank query requires a full multimodal LLM call, and these rerankers typically need instruction-tuning or RL training on multimodal labels. We take an alternative architectural choice—a path used widely in industry RAG systems but under-explored in academic long-document QA: keep a cheap, training-free text-LLM reranker, and pre-compute a structured VLM-generated annotation once per document offline. Concretely, we convert each candidate into an LLM-readable semantic record—a VLM-generated visual caption surfaces table/chart/figure content, and structured fields (section path, entity tags, topic/entity/structure concept hits) expose pre-computed evidence signals. On the ColPali pool, attaching this annotation raises rerank F1 from 0.332 to 0.581. Consistent with the mechanism above, removing the caption hurts both pools but hurts BGE-M3 more than ColPali: a visual retriever partially compensates for the LLM's blindness at retrieval time, but does not eliminate the need for an explicit visual signal at rerank time.

Motivated by these observations, we propose TRIDENT, a long-document multimodal QA pipeline with two components addressing analogous bottlenecks at the rerank and generation stages. TRIDENT-R attaches the annotation above to each top-20 candidate and uses a single LLM call to select an adaptive set of $K \leq 5$ evidence pages. The same multi-aspect logic carries over to generation: a single VLM call must implicitly trade off attention across text, tables, and figures within each retrieved page, so we additionally propose TRIDENT-S, which prompts the VLM under topical, entity, and structural lenses before synthesizing a final answer. On MMLongBench-Doc, the annotation+rerank protocol applied to five heterogeneous candidate pools (BM25, BGE-M3, RRF, ColPali, and our multi-axis XKG) raises retrieval F1 on all five, with every reranked pool exceeding the strongest adaptive-K baseline PageIndex (Zhang et al., 2025a); the improvement also holds on LongDocURL. TRIDENT-S targets open-ended synthesis questions by design, providing consistent gains on that workload while leaving extractive workloads to a standard single-call VLM. End-to-end, the best TRIDENT configuration is the strongest downstream pipeline in our evaluation, with rankings consistent across two LLM judges (κ=0.913).

2 Related Work

Long-document VQA benchmarks and multimodal RAG systems. We evaluate on MMLongBench-Doc (Ma et al., 2024b) and LongDocURL (Deng et al., 2025), two benchmarks targeting VQA over hundred-plus-page documents with multi-modal evidence (text, tables, charts, figures, layout); MMLongBench-Doc reports that GPT-4o-vision with full-document access lags VLMs given oracle evidence pages by $\sim$30 points. Recent benchmarks complement these along different axes: retrieval-focused MMDocIR (Dong et al., 2025) (313 long documents, notably finding VLM image descriptions outperform OCR text—consistent with our caption-dominance result), fine-grained evidence-selection MMDocRAG (Dong et al., 2026) (metrics beyond page-level F1), and difficulty-graded REAL-MM-RAG (Wasserman et al., 2025b) (finance / technical). Multimodal RAG systems combine multi-modal retrievers with multimodal LLMs (M3DocRAG (Cho et al., 2024)), use hierarchical indexing with intra-page and cross-page chunks (MMRAG-DocQA (Gong et al., 2025)), build chunk–query graphs (MLDocRAG (Zhang

and Wu, 2026)), triage by document structure (PDFTriage (Saad-Falcon et al., 2024)), or use VLM-generated structured JSON descriptions for tables and figures (MultiFinRAG (Gondhalekar et al., 2025)). TRIDENT differs by factoring the pipeline into a swappable Stage 1 retriever and a Stage 2 annotation + rerank protocol we empirically isolate (§4.2).

Document retrievers and adaptive-K selection. Page-level retrievers span sparse (BM25), dense (BGE-M3 (Chen et al.); ColBERT-style late interaction (Khattab and Zaharia, 2020; Santhanam et al., 2022); RRF fusion (Cormack et al., 2009)), and vision-aware variants (ColPali (Faysse et al., 2025) with PaliGemma plus patch-level MaxSim; DSE (Ma et al., 2024a) with image-level dense embeddings); none condition on document-level structure. Adaptive-K selection has been addressed by section-tree navigation (PageIndex (Zhang et al., 2025a)), hierarchical summarization (RAPTOR (Sarthi et al., 2024)), reflection-controlled retrieval (Self-RAG (Asai et al., 2024)), summary-based re-ranking (SimpleDoc (Jain et al., 2025)), and relevance clustering (AVIR (Li et al., 2025)). Recent agentic approaches use iterative refinement (Doc-React (Wu et al., 2025)), unified MLLM retrieval-generation (URaG (Shi et al., 2026)), or multi-turn RL (MM-Doc-R1 (Lin et al., 2026)). These either rely on a single structural / semantic / visual signal or require end-to-end MLLM / agent training; TRIDENT is complementary—training-free, retriever-agnostic, and easily bolted onto any pipeline via multi-aspect annotation.

LLM rerankers with structured features. Retrieval-augmented generation (Lewis et al., 2020) traditionally couples retriever and generator via end-to-end training; TRIDENT sits at the opposite end of the design space—modular and prompting-based, with multi-aspect annotation serving as a frozen abstraction layer pairable with any pool and any VLM without re-training. RankGPT (Sun et al., 2023) exposes raw candidate text to an LLM for re-scoring; recent work instead conditions LLM rerankers on *structured per-candidate features*: CoRank (Tian et al., 2025) (categories / sections / keywords for scientific retrieval), FinCARDS (Zhou et al., 2026) (schema fields for financial QA), KeyB2 (Li et al., 2024) (block selection), and AcuRank (Yoon et al., 2026) (uncertainty-aware adaptive listwise)—all text-only. TRIDENT differs in three ways: (i) we target *long-document multimodal* QA, using VLM-generated visual captions as an LLM-readable interface for multimodal evidence; (ii) we test the interface via cross-pool isolation across five retriever paradigms, and show the gains transfer to the visual ColPali pool via asymmetric drop-caption and pool-invariant caption-only ablations (§4.2, Appendix I); (iii) our generation-side analog TRIDENT-S extends the same principle to multi-view VLM prompting. We additionally expose multi-aspect metadata—concept hits across topic / entity / structure axes plus section path, visual caption, and entity tags—enabling annotation-level reasoning when surface text is repetitive among top-20 candidates. Knowledge graphs have been used for retrieval over textual corpora (Gutiérrez et al., 2024) and document QA (Sanmartin, 2024); TRIDENT differs by treating annotation as a page-level reranker-friendly abstraction layer, not as a graph-walk substrate.

Caption-based vs MLLM-based multimodal reranking. Two architectural families incorporate visual content into multimodal rerankers. **Caption-based approaches** convert page images to text via VLM descriptions and feed a text-only LLM reranker; this pattern is common in industry RAG systems (Haystack, NVIDIA RAG Blueprint, IBM Multimodal RAG) but academically under-explored under a "loses visual details" assumption. **MLLM-based approaches** let a multimodal LLM see page images directly, spanning zero-shot prompt-based rerankers (Chen et al., 2025; Mortaheb et al., 2025), fine-tuned / RL-trained variants (Xu et al., 2025; Wasserman et al., 2025a; Lin et al., 2025), and listwise / production rerankers (Li et al., 2026); all are effective but costly per query and typically require multimodal instruction-tuning or RL training. Concurrent work also retrieves over VLM-encoded pages directly (Yu et al., 2025). TRIDENT systematically revisits the caption-based path for long-document multimodal QA and shows that with structured multi-aspect annotation, a text-only LLM reranker is competitive with or better than text-only / visual baselines at a fraction of per-query cost, while remaining drop-in compatible with any upstream retriever—tested via cross-pool isolation across five retriever paradigms (§4.2).

VLM-augmented LLM QA and multi-view prompting. SCRA-VQA (Zhang et al., 2025b) converts single-image VQA into a caption + LLM rerank pipeline, demonstrating that VLM captions

can substantially augment LLM-based QA at the single-image scale; MPCAR (Rahman et al., 2025) generates diverse complementary descriptions from multiple analytical perspectives and fuses them in a single-image LVLM reasoning prompt. TRIDENT extends this lineage to *long-document* multimodal retrieval and generation: visual caption is identified as the strongest universal annotation field on the retrieval side; the three orthogonal lenses on the generation side (TRIDENT-S) are structured around our T / E / S decomposition axes and are followed by a format-aware LLM synthesis call. Crucially, we *characterize* TRIDENT-S's effective regime (substantial gains on open-ended synthesis questions only) rather than claim universal lift.

3 Method

3.1 Problem Formulation

Let a document $D = \{p_1, \ldots, p_N\}$ contain N pages of mixed text and visual content (N may exceed 100). Given a natural-language question q with gold evidence set $E_q \subseteq D$ and gold answer a_q, the long-document VQA task is to produce an answer $\hat{a}$ approximating a_q. We decompose the task into retrieval and generation:

$$\hat{P}_q = \mathcal{R}(q, D),\ |\hat{P}_q| \leq K_{\max}; \quad \hat{a} = \mathcal{G}(q, \hat{P}_q), \tag{1}$$

where $\mathcal{R}$ returns up to $K_{\max}$ evidence pages and $\mathcal{G}$ is a VLM. Following the reranker-interface bottleneck identified in §1, TRIDENT factors $\mathcal{R}$ into a candidate-retrieval stage and an annotation + rerank stage:

$$\hat{P}_q = \mathcal{R}_2(q, \mathcal{R}_1(q, D),\ \Phi(D)), \tag{2}$$

where $\mathcal{R}_1$ returns a top-M=20 pool, $\Phi(D)$ is a multi-aspect annotation (§3.2), and $\mathcal{R}_2$ is an LLM reranker with $K_{\max}$=5 (§3.3). $\mathcal{G}$ is then realized by either a single VLM call or a multi-view generator TRIDENT-S (§3.5). $\mathcal{R}_1$ is treated as a swappable candidate-pool source (§3.4), while $\mathcal{R}_2$ and the page annotation Φ are the focus of this work.

3.2 TRIDENT-R-Index: Multi-Aspect Page Annotation

TRIDENT-R-Index is a function $\Phi(D) = (\mathcal{C}, \mathcal{M})$ that maps a document D to a set of concepts $\mathcal{C}$ and per-page metadata records $\{\mathcal{M}_p\}_{p \in D}$. The concept set decomposes into three disjoint axes:

$$\mathcal{C} \ = \ \mathcal{C}_T \cup \mathcal{C}_E \cup \mathcal{C}_S \tag{3}$$

where $\mathcal{C}_T$ contains topic concepts, $\mathcal{C}_E$ entity-bundle concepts, and $\mathcal{C}_S$ structural concepts. Each concept $c \in \mathcal{C}_X$ ($X \in \{T, E, S\}$) carries a name, an embedding $\mathbf{e}_c \in \mathbb{R}^d$ (where applicable), and a *weighted membership*—a set of (page, weight) pairs $\{(p, w_{c,p}) : p \in D,\ w_{c,p} > 0\}$ where $w_{c,p} \in [0, 1]$ reflects how strongly page p contributes to concept c.

Topic concepts $\mathcal{C}_T$. We segment each page p into overlapping text windows of ≤ 512 tokens (stride 256), denoted $W_p = \{w_p^{(1)}, \ldots, w_p^{(n_p)}\}$. All windows $\bigcup_p W_p$ are embedded by a sentence embedding model $\mathcal{E}_W$ and L2-normalized. We cluster the embedding set via mini-batch K-means with $K_T = \min(64, \frac{|W|}{8})$. Each resulting cluster $c \in \mathcal{C}_T$ records its centroid $\mathbf{e}_c$, member windows W_c, and per-page weight:

$$w_{c,p} \ = \ \frac{|W_c \cap W_p|}{|W_c|}, \quad \text{so that} \sum_p w_{c,p} = 1. \tag{4}$$

Entity-bundle concepts $\mathcal{C}_E$. An extraction LLM $\mathcal{L}_{\text{ext}}$ extracts a flat set of typed entities $\mathcal{E}_p$ from each page—named entities, numbers, dates, monetary values, percentages, and table / figure references. We group co-occurring entities across pages using density-based clustering on a graph where entities are connected when they co-occur on at least one page, retaining bundles of size ≥ 2. Each bundle $c \in \mathcal{C}_E$ stores its member entity set $\mathcal{E}_c$ and the pages on which $\mathcal{E}_c$ surfaces, with weight $w_{c,p}$ equal to the fraction of $\mathcal{E}_c$ entities appearing on p.

Structural concepts $\mathcal{C}_S$. We obtain a hierarchical section tree of D from a public PDF structural parser $\mathcal{P}_{\text{sec}}$ (any document parser returning a section hierarchy can be substituted; we ground our choice in §4.1). Each leaf section s becomes a structural concept c with member pages $\{p : p \in \text{span}(s)\}$ (uniform weights), and embedding $\mathbf{e}_c = \mathcal{E}_W(\text{summary}(s))$, where $\text{summary}(s)$ is the section summary returned by $\mathcal{P}_{\text{sec}}$.

Per-page metadata $\mathcal{M}_p$. In addition to the concepts above, we attach four per-page fields to each p: (a) a *structural path* string (e.g. "Item 8 / Consolidated Balance Sheets"), read off the section tree; (b) a *visual caption* vc_p, generated by a captioning VLM $\mathcal{V}_{\text{cap}}$ prompted to describe the page's visible tables, figures, charts, and layout in ≤ 60 words; (c) a flat *entity-tag list* $\text{et}_p \subseteq \mathcal{E}_p$ retaining the top-15 most salient entities/numbers on the page; and

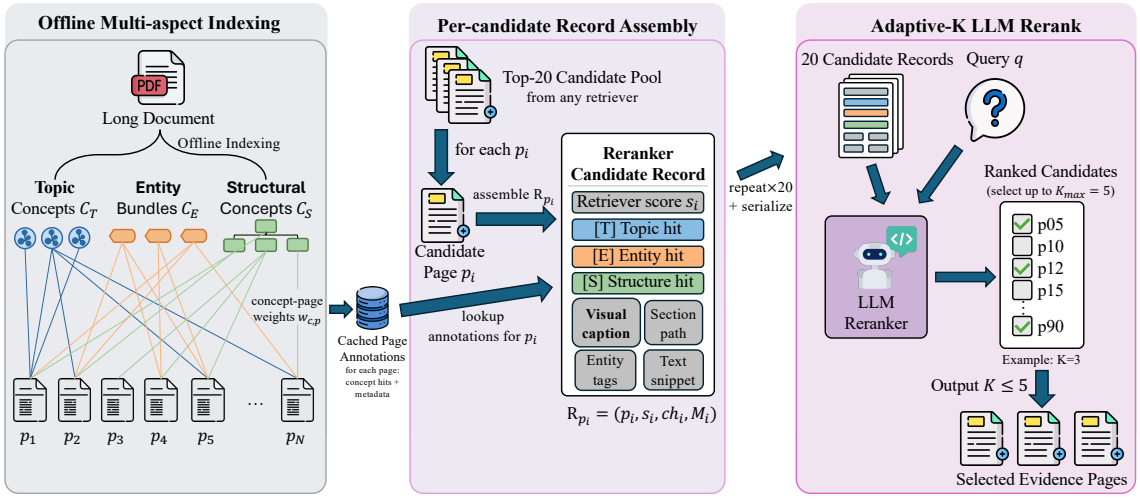

Figure 2: Overall framework of TRIDENT-R.

(d) the page's raw text snippet (truncated to 600 characters).

Offline cost. Index construction requires per-page LLM/VLM passes once per document and is cached on disk for reuse across all subsequent queries; full token-level accounting is in Appendix H.

3.3 TRIDENT-R: Adaptive-K LLM Rerank

Given a query q, we first obtain a candidate page pool $P_q \subseteq D$ with $|P_q| \leq M = 20$ from one of the retrievers in §3.4, then TRIDENT-R performs a single LLM call that returns an adaptive-K ordered subset $\hat{P}_q \subseteq P_q$ with $|\hat{P}_q| \leq K_{\max} = 5$.

Candidate payload. For each candidate page $p \in P_q$, we assemble a structured record $\mathcal{R}_p = (p,\ s_p,\ \mathrm{ch}_p,\ \mathcal{M}_p)$ where s_p is the underlying retriever score and ch_p is a *concept-hit list*

$$\mathrm{ch}_p = \{(X, c.\mathrm{name}) : c \in \mathcal{C}_X,\ w_{c,p} > 0,\ X \in \{T, E, S\}\} \quad (5)$$

collapsing all concept memberships involving p into a typed list of ≤ 6 tags (e.g. `T:Financial Statements`, `E:Nike`). Combined with $\mathcal{M}_p$, each candidate exposes *five* discriminating signals to the reranker: concept hits, section path, visual caption, entity tags, and a truncated text excerpt.

Rerank step. We serialize $\{\mathcal{R}_p\}_{p \in P_q}$ as a JSON list and prompt a rerank LLM $\mathcal{R}_\theta$ (at τ=0) with the question and candidate list, expecting a JSON response of the form:

$$\mathcal{R}_\theta(q, \{\mathcal{R}_p\}) = (\text{thinking},\ \hat{P}_q), \quad |\hat{P}_q| \leq K_{\max}. \quad (6)$$

The reranker may return fewer than $K_{\max}$ pages when the question's evidence is concentrated, yielding an adaptive $|\hat{P}_q|$. The verbatim rerank prompt is in Appendix B.

Why JSON-conditioned rerank? Compared to a fully-textual rerank prompt (e.g. RankGPT (Sun et al., 2023)), exposing structured per-candidate fields lets the LLM attend to specific evidence signals—visual caption, entity tags, concept memberships—rather than re-deriving them from the raw page text in its limited context. We empirically validate this design in §4.2 and isolate per-field contributions in §4.3.

3.4 Candidate Pool Sources

TRIDENT-R is agnostic to the source of P_q. We evaluate four pool choices, each returning the top-M=20 pages by score.

Single-signal text pools. **BM25** returns the top-20 pages by Okapi BM25 score over page text. **BGE-M3** (Chen et al.) returns the top-20 pages by cosine similarity between the query embedding and page embeddings (dim 1024). **RRF** (Cormack et al., 2009) fuses BM25 and BGE-M3 rankings by Reciprocal Rank Fusion (k=60).

Vision pool. **ColPali** (Faysse et al., 2025) encodes each page image with PaliGemma and returns top-20 pages by late-interaction MaxSim.

Multi-axis pool (xKG). We compose a multi-channel score from $\mathcal{C}_T, \mathcal{C}_E, \mathcal{C}_S$. A lightweight LLM router (Appendix D) first classifies q into a class $\kappa \in \{\text{topical, entity-level, structural, mixed}\}$ and

extracts query entities $\mathcal{E}_q$. The concept-level scores per axis are:

$$\text{score}_T(c, q) = \cos(\mathbf{e}_c, \mathbf{e}_q), \qquad c \in \mathcal{C}_T \tag{7}$$

$$\text{score}_E(c, q) = \frac{|\mathcal{E}_c \cap_{\text{sub}} \mathcal{E}_q|}{|\mathcal{E}_c|}, \qquad c \in \mathcal{C}_E \tag{8}$$

$$\text{score}_S(c, q) = \cos(\mathbf{e}_c, \mathbf{e}_q), \qquad c \in \mathcal{C}_S \tag{9}$$

where $\cap_{\text{sub}}$ denotes loose substring matching to handle entity surface variants. Page-level scores are obtained by aggregating concept contributions weighted by membership and a class-conditional fusion $(\alpha_T^\kappa, \alpha_E^\kappa, \alpha_S^\kappa)$:

$$s_p^{\text{XKG}}(q) \;=\; \sum_{X \in \{T,E,S\}} \alpha_X^\kappa \sum_{c \in \mathcal{C}_X} \text{score}_X(c, q) \cdot w_{c,p}. \tag{10}$$

The top-M pages by s_p^{XKG} form the XKG pool. Specific $(\alpha_T^\kappa, \alpha_E^\kappa, \alpha_S^\kappa)$ values for each κ, and a router-ablation comparison with uniform weights, are in Appendix D.

3.5 TRIDENT-S: Multi-View Generation

Given retrieved pages $\hat{P}_q$, TRIDENT-S prompts a VLM $\mathcal{V}$ three times in parallel under three lenses applied to the *same* retrieved image set $\{\text{render}(p) : p \in \hat{P}_q\}$:

T: *Topical lens.* "What is each page about thematically? Which topics relate to the question?"

E: *Entity / value lens.* "Extract the named entities, numerical values, dates, table cells, and figure labels visible on each page. Use verbatim values."

S: *Structural lens.* "Where does each page sit in the document organization (section title, captions, headers)?"

Producing three candidate answers $\hat{a}^T, \hat{a}^E, \hat{a}^S$. A fourth *synthesis* call $\mathcal{S}$ merges them into a final answer:

$$\hat{a} \;=\; \mathcal{S}\big(q,\, \hat{a}^T,\, \hat{a}^E,\, \hat{a}^S\big), \tag{11}$$

with format-aware preferences encoded in the synthesis prompt (preferring the E-view answer for numeric questions, the S-view for location questions, and consensus otherwise). When all three lenses agree the question is unanswerable, $\mathcal{S}$ returns "Not answerable"—a calibration property we examine in §4.4. Full prompts are in Appendix C.

Why three lenses? Long-document evidence is often heterogeneous within a single retrieved page (e.g., a financial-report page contains both narrative text *and* a table). A single VLM call must implicitly trade off attention across modalities; by issuing three lens-conditioned prompts we explicitly induce the VLM to surface each evidence type, then defer the trade-off to a smaller textual synthesis call. This is conceptually analogous to multi-aspect annotation at generation time.

We treat TRIDENT-S as a question-format-conditioned extension; the empirical characterization of when it helps is in §4.4 and §4.5.

4 Experiments

4.1 Setup

Datasets: MMLongBench-Doc (Ma et al., 2024b) (1,082 QAs / 134 documents) and LongDocURL (Deng et al., 2025) (1,123 QAs / 200 documents). Retrieval metrics are reported on the N=830 answerable subset of MMLongBench-Doc. End-to-end accuracy is reported on a 300-QA stratified subset; the full sampling details are in Appendix E.

Baselines: BM25, BGE-M3 (Chen et al.), RRF (Cormack et al., 2009), ColPali (Faysse et al., 2025) at K=5 (fixed-K, no rerank), and PageIndex (Zhang et al., 2025a) adaptive-K cookbook with GPT-4.1. Our cross-pool isolation pairs each retriever pool with the same TRIDENT-R annotation + LLM rerank protocol.

Module assignments: We instantiate the abstract modules of §3 as follows—text embeddings ($\mathcal{E}_W$): `text-embedding-3-large`; extraction LLM $\mathcal{L}_{\text{ext}}$, rerank LLM $\mathcal{R}_\theta$, and synthesis LLM $\mathcal{S}$: GPT-4.1 at τ=0; captioning VLM $\mathcal{V}_{\text{cap}}$: GPT-4o-mini; answering VLM $\mathcal{V}_{\text{ans}}$: GPT-4.1 (vision); structural parser $\mathcal{P}_{\text{sec}}$: the PageIndex API (Zhang et al., 2025a) used purely as a PDF section-tree extractor, independent of the PageIndex *retrieval system* we benchmark against. The candidate JSON records produced by Φ are consumable by any instruction-tuned LLM; the annotation + rerank protocol is therefore not tied to a specific model family. Claude and Gemini additionally serve as LLM-as-judge evaluators (Appendix L).

Metrics: Retrieval is evaluated by page-level Recall, Precision, F1, MRR, and nDCG against gold evidence pages. End-to-end QA is evaluated by MMLongBench-Doc Generalized Accuracy and LLM-as-judge accuracy.

Method	avg. K	Recall	Prec.	F1	MRR	nDCG
Fixed-K baselines (no rerank, K=5)						
BM25	5.00	0.525	0.162	0.231	0.448	0.437
RRF	5.00	0.556	0.172	0.246	0.524	0.496
BGE-M3	5.00	0.572	0.178	0.254	0.524	0.502
ColPali	5.00	0.739	0.236	0.332	0.692	0.675
Adaptive-K, single-aspect signal						
PageIndex	2.99	0.644	0.442	0.480	0.569	0.583
Adaptive-K, pool + TRIDENT annotation + rerank (ours)						
BM25 (sparse) + ann. + rerank	2.89	0.665	0.471	0.503	0.707	0.681
RRF (hybrid) + ann. + rerank	3.78	0.748	0.454	0.511	0.751	0.728
XKG (graph) + ann. + rerank (TRIDENT-R)	**2.82**	0.692	0.485	0.527	0.714	0.690
BGE-M3 (dense) + ann. + rerank	2.89	0.726	0.499	0.546	0.744	0.723
ColPali (visual) + ann. + rerank	2.85	0.766	0.533	0.581	0.788	0.768

Table 1: **Cross-pool isolation on MMLongBench-Doc** (N=830 aligned QAs; all reranked rows use GPT-4.1 with $K_{\max}$=5). The same annotation + rerank protocol improves sparse, hybrid, graph, dense, and visual candidate pools.

Pool	Candidate view	F1	MRR
BGE-M3	raw top-5 retrieval	0.254	0.524
BGE-M3	text snippet + LLM rerank	0.374	0.523
BGE-M3	caption-only annotation + rerank	0.518	0.692
BGE-M3	full annotation + rerank	0.546	0.744
ColPali	raw top-5 retrieval	0.332	0.692
ColPali	text snippet + LLM rerank	0.423	0.598
ColPali	caption-only annotation + rerank	0.554	0.744
ColPali	full annotation + rerank	0.581	0.788

Table 2: Mechanism ablations. LLM reranking with only raw text snippets does not improve first-hit ranking on either pool (MRR flat on BGE-M3, and decreasing on ColPali because the LLM judgment overrides ColPali's visual ranking signal); the ranking lift requires the structured annotation, and the visual caption alone recovers most of it on both text and visual pools.

PageIndex output: PageIndex selects section nodes, which we expand to constituent pages and truncate at top-5 for apples-to-apples comparison. On MMLongBench-Doc this cap is rarely active (18.6% of queries); on LongDocURL it is more active (49.7%).

4.2 Cross-Pool Retrieval Isolation

Table 1 tests whether the proposed interface works only for a particular retriever or transfers across candidate-pool distributions. The answer is consistent: applying the same annotation + rerank protocol more than doubles raw retriever F1 for several pools (e.g., BGE-M3 $0.254 \rightarrow 0.546$) and raises all five pools above PageIndex. The visual ColPali pool is the most important case for our story: although ColPali already encodes page images, converting its candidates into LLM-readable annotations still raises F1 from 0.332 to 0.581. This indicates that the interface is not merely supplying missing visual information to text retrievers.

4.3 Mechanism: What the Reranker Sees

Table 2 isolates why the interface helps. The text-snippet-only rerank rows show that an LLM with raw page snippets improves F1 by shortening the output, but does not improve first-hit ranking on either pool: on BGE-M3 MRR is essentially unchanged ($0.524 \rightarrow 0.523$), and on ColPali MRR actually drops ($0.692 \rightarrow 0.598$) because the snippet-only LLM judgment overrides ColPali's stronger visual ranking signal. The ranking lift appears only once the reranker sees structured page annotation, and the visual caption alone is sufficient to recover most of it (BGE-M3 MRR $\rightarrow 0.692$, ColPali MRR $\rightarrow 0.744$ with caption-only; $\rightarrow 0.744/0.788$ with full annotation). Full per-field drop ablations (Appendix I) show the same pattern: removing visual caption hurts both pools, and the remaining topic/entity/structure axes add a small but nearly identical marginal gain (+0.028 / +0.027). Thus the caption is not just modality completion for text retrievers; it is a reranker-facing semantic representation that remains useful even when the upstream retriever is visual.

We keep detailed secondary analyses in the appendix to keep the main text focused on the reranker-interface claim. Appendix I reports full field ablations on BGE-M3 and ColPali;

Method	GPT J.	Claude J.	Gen Acc	Gen F1
BGE-M3 (Plain)	44.7	44.7	43.5	41.3
ColPali (Plain)	48.3	46.3	46.4	45.9
PageIndex +TRIDENT-S	49.3	48.3	45.9	42.6
BGE-M3 +ann. +rk +TRIDENT-S	52.0	50.0	48.9	46.4
ColPali +ann. +rk, Plain	51.0	49.7	47.1	44.5
ColPali +ann. +rk +TRIDENT-S	**53.7**	**52.7**	**51.2**	**48.9**
TRIDENT-R (XKG pool, Plain)	50.7	48.0	47.2	42.9
TRIDENT-R (XKG pool) +TRIDENT-S	52.0	50.7	46.8	44.3

Table 3: End-to-end QA on the 300-QA MMLongBench-Doc subset (selected pipelines; all values in %). Full 14-pipeline results, including answer accuracy and unanswerable abstention, are in Appendix A.

Method	avg. K	Recall	Prec.	F1	MRR	nDCG
Fixed-K baselines (no rerank, K=5)						
BM25	5.00	0.623	0.191	0.281	0.616	0.568
BGE-M3	5.00	0.649	0.205	0.298	0.622	0.575
RRF	5.00	0.684	0.217	0.315	0.675	0.622
ColPali	5.00	**0.750**	0.235	0.343	**0.763**	**0.700**
Adaptive-K						
PageIndex	4.03	0.598	0.302	0.366	0.430	0.463
TRIDENT-R	3.19	0.546	**0.362**	**0.398**	0.603	0.552

Table 4: LongDocURL retrieval comparison (N=1,122 QAs with non-empty gold evidence). The retrieval-side trend transfers, although the full cross-pool isolation study is conducted on MMLongBench-Doc.

Appendix D reports XKG router details; Appendix J reports the adaptive-K distribution; and Appendix F reports bootstrap, Wilcoxon, and McNemar tests.

4.4 End-to-End QA as Supporting Evidence

Table 3 shows that the retrieval-side gain generally transfers to downstream QA. The ColPali pool with annotation and rerank is the best observed end-to-end configuration in this subset, and adding TRIDENT-S gives the highest Gen Acc and Gen F1 on MMLongBench-Doc. The full 14-pipeline table and inter-judge agreement are in Appendix A and Appendix L; the pipeline ranking is consistent across the two judges (κ=0.913).

4.5 LongDocURL Transfer Check

Table 4 provides an external transfer check on N=1,122 QAs. TRIDENT-R achieves the highest retrieval F1, showing that the retrieval-side trend transfers to LongDocURL. End-to-end results in Appendix A show the same pattern: ColPali plus annotation and rerank remains strongest.

When does TRIDENT-S help? TRIDENT-S delivers consistent positive gains on MMLongBench-Doc's open-ended synthesis questions (+1.7 to +6.6 Gen Acc across all eight evaluated pipelines, with the largest absolute gain on weaker baselines—PageIndex Plain → +TRIDENT-S adds +6.6). On LongDocURL's MCQ-style and short-extractive workloads, the three lenses converge on the same extracted answer, and TRIDENT-S is bypassed by design. We treat TRIDENT-S as a question-format-conditioned extension; an a-priori format router that gates TRIDENT-S invocation is a natural deployment refinement.

5 Limitations and Conclusion

Limitations. TRIDENT-R requires an offline per-document indexing pass ($0.256/100 pages, amortized across queries; Appendix H). The reranker uses GPT-4.1 here, but smaller or open-source LLMs are drop-in substitutes (Appendix M). The five-pool isolation is conducted on MMLongBench-Doc, with LongDocURL as an external transfer check (§4.5). We use PageIndex only as a section-path parser; public PDF parsers such as PyMuPDF or GROBID could replace it. Finally, since most lift comes from visual captions, a caption-first variant is preferable when other fields are costly.

Conclusion. We identify a reranker-interface bottleneck in long-document multimodal QA: evidence selection improves when candidate pages are converted into LLM-readable semantic annotations, even with multimodal retrievers. TRIDENT-R implements this as a retriever-agnostic annotation-based LLM reranker. Across five candidate-pool families on MMLongBench-Doc, the same interface substantially improves retrieval F1, and every reranked pool exceeds PageIndex. The gain on ColPali shows that the effect is not merely missing visual information for text retrievers. Mechanism ablations show that snippet-only LLM reranking barely changes first-hit ranking, while visual captions recover most of the full lift on both text and visual pools. TRIDENT-S is a secondary generation extension: useful for open-ended synthesis, but unnecessary for MCQ and short-extractive workloads. Overall, the results shift the focus from what the retriever scores to what the downstream reranker can see about each candidate page.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Zhanpeng Chen, Chengjin Xu, Yiyan Qi, Xuhui Jiang, and Jian Guo. 2025. Vlm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8140–8158.

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

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

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

Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. 2025. Mmdocir: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30959–30993.

Kuicai Dong, CHANG YUJING, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2026. Benchmarking retrieval-augmented multimomal generation for document question answering. *Advances in Neural Information Processing Systems*, 38.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pages 61424–61449.

Chinmay Gondhalekar, Urjitkumar Patel, and Fang-Chun Yeh. 2025. Multifinrag: An optimized multimodal retrieval-augmented generation (rag) framework for financial question answering. *arXiv preprint arXiv:2506.20821*.

Ziyu Gong, Yihua Huang, and Chengcheng Mai. 2025. Mmrag-docqa: A multi-modal retrieval-augmented generation method for document question-answering with hierarchical index and multi-granularity retrieval. *arXiv e-prints*, pages arXiv–2508.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569.

Chelsi Jain, Yiran Wu, Yifan Zeng, Jiale Liu, Shengyu Dai, Zhenwen Shao, Qingyun Wu, and Huazheng Wang. 2025. Simpledoc: Multi-modal document understanding with dual-cue page retrieval and iterative refinement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28398–28415.

Omar Khattab and Matei Zaharia. 2020. 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*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Minghan Li, Éric Gaussier, Juntao Li, and Guodong Zhou. 2024. Keyb2: Selecting key blocks is also important for long document ranking with large language models. *arXiv preprint arXiv:2411.06254*.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Zongmin Li, Yachuan Li, Lei Kang, Dimosthenis Karatzas, and Wenkang Ma. 2025. Avir: Adaptive visual in-document retrieval for efficient multi-page document question answering. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pages 1–7.

Jiahang Lin, Kai Hu, Binghai Wang, Yuhao Zhou, Zhiheng Xi, Honglin Guo, Shichun Liu, Junzhe Wang, Shihan Dou, Enyu Zhou, and 1 others. 2026. Mmdoc-r1: Training agents for long document visual question answering through multi-turn reinforcement learning. *arXiv preprint arXiv:2604.13579*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pages 44215–44234.

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

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024b. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Matin Mortaheb, Mohammad A Amir Khojastepour, Srimat T Chakradhar, and Sennur Ulukus. 2025. Re-ranking the context for multimodal retrieval augmented generation. *arXiv preprint arXiv:2501.04695*.

Amirul Rahman, Qiang Xu, and Xueying Huang. 2025. Mpcar: Multi-perspective contextual augmentation for enhanced visual reasoning in large vision-language models. *arXiv preprint arXiv:2508.12400*.

Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, Seunghyun Yoon, Ryan A Rossi, and Franck Dernoncourt. 2024. Pdftriage: Question answering over long, structured documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 153–169.

Diego Sanmartin. 2024. Kg-rag: Bridging the gap between knowledge and creativity. *arXiv preprint arXiv:2405.12035*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pages 32628–32649.

Yongxin Shi, Jiapeng Wang, Zeyu Shan, Dezhi Peng, Zening Lin, and Lianwen Jin. 2026. Urag: Unified retrieval and generation in multimodal llms for efficient long document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25357–25365.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14918–14937.

Runchu Tian, Xueqiang Xu, Bowen Jin, SeongKu Kang, and Jiawei Han. 2025. Corank: Llm-based compact reranking with document features for scientific retrieval. *arXiv preprint arXiv:2505.13757*.

Navve Wasserman, Oliver Heinimann, Yuval Golbari, Tal Zimbalist, Eli Schwartz, and Michal Irani. 2025a. Docrerank: Single-page hard negative query generation for training multi-modal rag rerankers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8651–8669.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025b. Real-mm-rag: A real-world multimodal retrieval benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31660–31683.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, and 1 others. 2025. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. 2025. Mm-r5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*.

Soyoung Yoon, Gyuwan Kim, Gyu-Hwung Cho, and 1 others. 2026. Acurank: Uncertainty-aware adaptive computation for listwise reranking. *Advances in Neural Information Processing Systems*, 38:26388–26417.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pages 21074–21098.

Mingtian Zhang, Yu Tang, and PageIndex Team. 2025a. Pageindex: Next-generation vectorless, reasoning-based rag. *PageIndex Blog*. Https://pageindex.ai/blog/pageindex-intro.

Yan Zhang, Jiaqing Lin, Miao Zhang, Kui Xiao, Xiaoju Hou, Yue Zhao, and Zhifei Li. 2025b. Scravqa: Summarized caption-rerank for augmented large language models in visual question answering. In *International Conference on Database Systems for Advanced Applications*, pages 37–52. Springer.

Yongyue Zhang and Yaxiong Wu. 2026. Mldocrag: Multimodal long-context document retrieval augmented generation. *arXiv preprint arXiv:2602.10271*.

Yixi Zhou, Fan Zhang, Yu Chen, Haipeng Zhang, Preslav Nakov, and Zhuohan Xie. 2026. Fincards: Card-based analyst reranking for financial document question answering. *arXiv preprint arXiv:2601.06992*.

A Additional End-to-End Results

This appendix complements §4.4 and §4.5 with the full end-to-end QA tables on both benchmarks. Table 5 reports all 14 evaluated pipelines on the 300-QA MMLongBench-Doc subset, including the answer-accuracy and unanswerable-abstention columns omitted from the main paper. Table 6 reports the corresponding end-to-end results on the 300-QA LongDocURL stratified subset (all questions answerable), comparing TRIDENT-R against the strongest baseline ColPali + annotation + rerank, each with and without TRIDENT-S.

B TRIDENT-R Rerank Prompt

The full rerank prompt sent to GPT-4.1 is shown below. Variables in angle brackets are filled at runtime. The candidate JSON list (<candidates_json>) is the array of records described in §3.3; <query> is the natural-language question; <k> is the maximum number of pages (K=5 throughout this work).

```
You are picking the top-K pages most
likely to answer the question, from a
candidate list. Each candidate page
comes with the document signals on it
AND a list of "concept hits"—high-level
concepts that pulled it into the
candidate pool. Pages hit by multiple
distinct concepts are stronger evidence.

Question: <query>

Candidate pages (<n> total):
<candidates_json>

Reply as JSON: {"thinking": ". . . ",
"page_list": [<page_id>, . . . ]}
Pick the strongest evidence pages first.
Return at most <k> pages. Return only
JSON.
```

Each candidate record in <candidates_json> has the form:

```
{page_id, retriever_score,
concept_hits: ["T:<name>", "E:<name>",
"S:<name>"], section, visual, entities,
text}
```

The retriever_score field holds the underlying Stage-1 pool score (Okapi BM25 for BM25, cosine for BGE-M3, RRF fused rank for RRF, late-interaction MaxSim for ColPali, or our multi-channel XKG fusion of Eq. 10); the field name is uniform across all pools so the rerank prompt template is pool-agnostic. (In our released code the field is currently named graph_score for historical reasons; the value semantics are as described.)

C TRIDENT-S Lens Prompts

TRIDENT-S prompts the VLM three times in parallel with the three lenses below, then synthesizes a final answer with the synthesis prompt. All four calls share the same retrieved page images.

(T) Topical lens.

```
Question: <query>

You are given <n> page images from a
long document. Read them through the
TOPICAL lens: what is each page about
thematically? Which topics on which
pages relate to the question? - Identify
the topical content of each page. -
Focus on how the topics connect to the
question.

Provide a clear, concise answer based
on the topical analysis. If none of
the provided pages contain information
sufficient to answer the question
through this topical lens, reply exactly:
"Not answerable".
```

(E) Entity / value lens.

```
Question: <query>

. . . Read them through the ENTITY /
VALUE lens: focus on the specific
named entities, numerical values, dates,
currencies, percentages, table cells,
and figure labels that appear on each
page. - Extract every named entity,
number, date, or specific value. -
Be precise—use verbatim values from the
images.

Provide a clear, concise answer focused
on the specific entities and values
surfaced. If none of the provided pages
contain the entities or values needed to
answer the question, reply exactly: "Not
answerable".
```

(S) Structural / section lens.

```
Question: <query>

. . . Read them through the STRUCTURAL
/ SECTION lens: focus on where
each page sits in the document's
organization—section titles visible on
the page, headers, chapter / part
markers, document structure, captions
of figures / tables. - Identify each
page's structural role. - Use captions
of tables/figures and section paths.

Provide a clear, concise answer that
leverages structural context. If
the structural context does not
contain information needed to answer
the question, reply exactly: "Not
answerable".
```

Synthesis prompt.

```
You are answering a long-document QA
question. We elicited three candidate
```

Method	GPT-4.1 J.	Claude J.	Gen Acc	Gen F1	Ans. Acc	Unans. Abst.
BM25 (Plain)	42.3	42.7	41.3	38.2	33.7	65.3
BM25 +TRIDENT-S	45.0	44.3	44.8	43.8	39.7	61.1
RRF (Plain)	44.3	44.3	42.6	39.6	36.3	62.5
RRF +TRIDENT-S	44.7	43.3	43.9	41.7	39.3	58.3
BGE-M3 (Plain)	44.7	44.7	43.5	41.3	37.9	61.1
BGE-M3 +TRIDENT-S	46.0	45.7	43.9	43.4	41.1	52.8
ColPali (Plain)	48.3	46.3	46.4	45.9	44.4	52.8
ColPali +TRIDENT-S	50.0	49.0	49.2	48.1	47.2	55.6
PageIndex (Plain)	42.7	43.0	41.3	37.2	29.7	**77.8**
PageIndex +TRIDENT-S	49.3	48.3	45.9	42.6	40.6	62.5
BGE-M3 +ann. +rk, Plain	48.3	46.7	44.6	42.9	43.4	48.6
BGE-M3 +ann. +rk, +TRIDENT-S	52.0	50.0	48.9	46.4	48.1	51.4
ColPali +ann. +rk, Plain	51.0	49.7	47.1	44.5	45.3	52.8
ColPali +ann. +rk, +TRIDENT-S	**53.7**	**52.7**	**51.2**	**48.9**	**50.7**	52.8
TRIDENT-R (Plain)	50.7	48.0	47.2	42.9	41.9	63.9
TRIDENT-R + TRIDENT-S	52.0	50.7	46.8	44.3	45.4	51.4

Table 5: Full end-to-end QA results on the 300-QA MMLongBench-Doc subset (all values in %). *J.*=LLM-as-judge accuracy; *+ann.*=multi-aspect annotation; *+rk*=adaptive-K rerank; *Plain*=single multimodal VLM call; +TRIDENT-S=three-lens generation with synthesis.

Method	GPT-4.1 J.	Claude J.	Gen Acc	Gen F1
TRIDENT-R (Plain)	51.3	47.7	42.6	47.3
TRIDENT-R +TRIDENT-S	50.0	50.0	41.1	45.5
ColPali+ann. +rk, Plain	**68.7**	65.7	56.0	58.0
ColPali+ann. +rk, +TRIDENT-S	66.3	**66.0**	**56.7**	**58.5**

Table 6: End-to-end QA on the 300-QA LongDocURL stratified subset (all values in %; all questions are answerable). LongDocURL is dominated by MCQ-style and short-extractive answers, so TRIDENT-S gives little average benefit.

```
answers from a vision-language model
under three orthogonal analytical lenses
applied to the same K=<n> retrieved page
images:

(T) Topical lens   (E) Entity / value
lens   (S) Structural / section lens

Question: <query>

Candidate answers:
[T-view] <ans_T>
[E-view] <ans_E>
[S-view] <ans_S>

Task: produce a single, clear, concise
final answer.

Synthesis rules: - If two or three lenses
converge, prefer that answer. - If only
one lens has a precise answer while
others are vague, prefer the specific
one—unless contradicted. - If question
is numeric, prefer E-view's value. -
If question is "where / which section",
prefer S-view. - If question is "what is
the topic", prefer T-view. - Match the
question's expected answer format. Do
not introduce information beyond what
the three candidate answers contain. Do
not invent values. If all three lenses
agree the question is not answerable,
reply exactly: "Not answerable".
```

D Query-Type Router

The router prompts an LLM (GPT-4.1, τ=0, JSON mode) to classify the question into one of four classes and to extract salient entities. The full prompt is:

```
You are routing a question against a
multi-view document index that has three
concept types: topical (semantic content
clusters), entity-bundle (groups of
co-occurring named entities/numbers),
and section-derived (structural
sections).

Question: <query>

Output JSON with three fields:
"entities": list of salient entities /
key terms / numbers from the question.
Include named entities, technical
terms, monetary amounts, percentages,
figure / table references. Drop generic
stopwords.
"class": one of "entity-level" (asks
about a specific named entity / number
/ figure), "topical" (asks for thematic
content / overview / explanation),
"structural" (asks about location /
section / where in the document), or
"mixed" if it spans multiple.
"rationale": one short sentence
explaining the class choice.

Reply with JSON only.
```

The class label determines the channel-fusion weights $(\alpha_T, \alpha_E, \alpha_S)$ applied when aggregating topical / entity / structural concept scores into a candidate-pool score:

Axis (category)	Full %	Subset %	Δ pp
Answer format			
Float	14.8	20.7	+5.9
Int	26.4	24.0	−2.4
List	13.3	14.3	+1.0
Str	23.1	17.0	−6.1
None (unans.)	22.4	24.0	+1.6
Evidence source			
Chart	16.1	14.2	−1.9
Figure	26.3	28.9	+2.6
Layout	10.9	8.0	−2.9
Plain-text	26.9	28.4	+1.5
Table	19.8	20.4	+0.6
Evidence pages			
0 (unanswerable)	22.6	25.7	+3.1
1	44.5	44.3	−0.1
2-3	26.3	24.3	−2.0
4+	6.7	5.7	−1.0
Document length			
0-30 pages	52.9	22.0	−30.9
31-60 pages	22.6	15.3	−7.2
61-120 pages	19.4	49.0	+29.6
121-200 pages	4.7	12.7	+8.0
200+ pages	0.5	1.0	+0.5

Table 7: Distribution match between the 300-QA subset and the full MMLongBench-Doc dataset. Most axes (Evidence Source, Evidence Pages, Unanswerable rate) match within ±3pp. Answer format is moderately biased toward Float (+5.9pp) and away from Str (−6.1pp); document length is intentionally biased toward longer documents (median 72 vs. 28 pages), aligning with our framework's long-document focus.

Class	α_T	α_E	α_S
`topical`	0.60	0.20	0.20
`entity-level`	0.20	0.60	0.20
`structural`	0.20	0.20	0.60
`mixed`	0.40	0.30	0.30

Replacing this class-conditional router with uniform weights $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ causes a small but consistent -0.003 F1 drop on the cross-pool isolation setup (§4.2, measured on the same N=830 aligned subset as Table 1), validating the router's contribution.

E 300-QA Subset Construction

The 300-QA subset is drawn from MMLongBench-Doc with seed 42. Sampling emphasizes the long-document setting that motivates this work: the subset has median document length 72 pages, compared to 28 in the full set, aligning with our framework's focus on documents that exceed any current VLM's input window. Table 7 reports the per-axis distribution match.

The LongDocURL subset of 200 QAs (§4.5) is sampled analogously, stratified by document length, answer format, and evidence source. Released QA-id lists for both subsets accompany the code release.

F Bootstrap CIs, Wilcoxon and McNemar tests

Retrieval (N=830). We report (i) non-parametric 95% bootstrap CIs on per-QA F1 (B=10,000 resamples drawn with replacement from the 830 aligned QAs); and (ii) pairwise Wilcoxon signed-rank tests on per-QA F1 against TRIDENT-R (xKG pool). We prefer Wilcoxon on F1 over McNemar on STRICT for retrieval because methods have different effective K (5 for fixed-K baselines; mean $K \in [2.7, 3.0]$ for adaptive-K methods), which would structurally favor the larger-K methods. CIs cluster tightly within ± 0.024. All five raw baselines are significantly worse than TRIDENT-R ($p < 0.001$). Among the +ann. + rerank variants, RRF and BGE-M3 swaps are statistically indistinguishable from TRIDENT-R (p=0.077 and 0.055), the BM25 swap is borderline (p=0.026), and the ColPali swap is significantly higher ($p < 0.001$). Taken together, all five +ann. + rerank rows in Table 1 fall within the $[0.503, 0.581]$ F1 band without pairwise dominance (except ColPali), validating the pool-agnostic claim.

End-to-end (N=300). For end-to-end Gen Acc on the 300-QA MMLongBench-Doc subset we use McNemar on a binary correct-score (≥ 1) outcome per QA, with continuity correction. None of the pairwise comparisons between top-tier pipelines (ColPali+ann. +TRIDENT-S, TRIDENT-R+TRIDENT-S, BGE-M3+ann. +TRIDENT-S, PageIndex+TRIDENT-S) reach $p < 0.05$ at N=300, reflecting the small sample size; however the ranking is consistent across all three LLM judges (Appendix L, inter-judge κ=0.913). Full pairwise results are released with our code.

G Case Study Records

We list the full records for the three case-study QAs in §G. All extracted answers are the GPT-4o-extracted short forms after the MMLongBench-Doc official evaluation pipeline. "Score" is the rule-based Generalized Accuracy score (1.0 = correct, 0.0 = wrong).

Case 1: Chart axes. Doc `2401.18059v1.pdf`; Q: *What are the horizontal and vertical axis of*

Figure 3 respectively? Gold: `[Context Length, F1]`.

- TRIDENT-R: retrieved pages $\{6, 5\}$, ext. "`[Context Length, F1]`", score 1.0.
- PageIndex: retrieved $\{3\}$, ext. "`Not answerable`", score 0.0.
- BGE-M3 raw: retrieved $\{5, 2, 16, 19, 22\}$, ext. "`Not answerable`", score 0.0.

Case 2: Activision R&D ratio. Doc `ACTIVISIONBLIZZARD_2019_10K.pdf`; Q: *What is R&D to asset ratio for Activision Blizzard in FY 2019?* Gold: `5.03%`.

- TRIDENT-R: retrieved $\{70, 69\}$, ext. "`0.0503`", score 1.0.
- PageIndex: retrieved $\{37, 62, 67, 68, 69\}$, ext. "`Not answerable`", score 0.0.
- BGE-M3 raw: retrieved $\{47, 29, 92, 174, 175\}$, ext. "`Not answerable`", score 0.0.

Case 3: Calibrated refusal. Doc `2311.16502v3.pdf`; Q: *In which image type does GPT-4o demonstrate least proficiency . . . ?* Gold: `Not answerable`.

- TRIDENT-R: retrieved $\{21, 111, 26\}$, ext. "`Not answerable`", score 1.0.
- PageIndex: retrieved $\{110, 111, 112\}$, ext. "`Chemical`", score 0.0.
- BGE-M3 raw: retrieved $\{112, 7, 26, 110, 2\}$, ext. "`Music`", score 0.0.

H Detailed Cost Breakdown

Token-level cost accounting for TRIDENT-R-Index, measured empirically on the 134-document MMLongBench-Doc corpus (6,492 pages, mean 103 text tokens per page, 40 caption tokens per page) at public OpenAI 2025 rates ($2.00/$8.00 per 1M input/output tokens for gpt-4.1; $0.15/$0.60 for gpt-4o-mini; $0.13 per 1M tokens for text-embedding-3-large; image tokens ≈ 1500 per full page at ZOOM=2.0).

The reported $0.256 per 100-page document is within the $\sim$ $0.30 range stated in the main text and is conservative for dense documents: doubling per-page text tokens (to ~ 200) raises total cost to approximately $0.36. Per-query cost is dominated by the multimodal VLM call: Plain generation is $0.012, TRIDENT-S multi-view is $0.039 (3 lens calls + 1 synthesis).

Component (per 100-page doc)	Model	Calls	USD
Visual caption	gpt-4o-mini	100	0.028
Entity extraction	gpt-4.1	100	0.201
Page embedding	emb-3-large	100	0.001
Concept naming	gpt-4.1	20	0.026
Offline total			**0.256**
Online per-query (avg K=2.85):			
Stage 2 LLM rerank	gpt-4.1	1	0.010
Plain VLM generation	gpt-4.1 (vision)	1	0.012
TRIDENT-S (3 lens + synth)	gpt-4.1 (vision)	4	0.039

Table 8: Empirical token-level cost breakdown. Entity extraction dominates offline indexing (78%); concept naming and visual captioning contribute $\sim 10\%$ each; page embedding is negligible (1%).

I Detailed Field Ablation on Two Pools

We complement §4.2 with the full per-field ablation on both the BGE-M3 (text-only) and ColPali (visual) pools. Each row removes exactly one annotation field from the rerank candidate JSON.

Configuration	F1	ΔF1	Recall	Prec.
BGE-M3 pool (text retriever, $F1_{full}$=0.546)				
+ full annotation	0.546	—	0.726	0.499
drop visual_caption	0.416	−0.130	0.657	0.354
drop concept_hits	0.515	−0.032	0.688	0.477
drop entities	0.507	−0.039	0.701	0.455
caption-only (only visual_caption)	**0.518**	−0.028	0.669	0.484
ColPali pool (visual retriever, $F1_{full}$=0.581)				
+ full annotation	0.581	—	0.766	0.533
drop visual_caption	0.498	−0.083	0.715	0.444
drop concept_hits	0.593	+0.013	0.751	0.556
drop entities	0.588	+0.007	0.773	0.542
caption-only (only visual_caption)	**0.554**	−0.027	0.723	0.516

Table 9: Per-field annotation ablation on two retrievers (N=830 aligned, $K_{\max}$=5). Standard rows: remove one field at a time. Caption-only rows : keep only the visual caption (drop concept_hits, section path, entity tags, text snippet). Caption alone achieves $\sim$90% of the full annotation lift on both pools (−0.028 / −0.027 from full), with the gap remarkably pool-invariant. Combined with the asymmetric drop-caption result (−0.130 on BGE-M3 vs. −0.083 on ColPali; 36% smaller on visual pool), this rules out pure modality completion as the sole mechanism: the lift stems from structuring page evidence as LLM-readable input, with visual caption acting as a *semantic interface* for the text-only LLM reranker (effective regardless of whether the upstream retriever is visual or textual).

J Adaptive-K Distribution by Question Type

We bin the rerank-selected K by MMLongBench-Doc question type (using the `evidence_sources`

field) across four annotated-rerank pools.

Question type	BM25 +ann	BGE-M3 +ann	ColPali +ann	TRIDENT-R (xKG)+ann
Pure-text	2.95	3.08	3.11	2.94
Figure	2.64	2.61	2.60	2.71
Table	2.90	2.90	2.98	2.75
Chart	2.81	2.77	2.55	2.59
Layout	2.84	2.78	2.97	2.75
Multi-source	3.01	2.98	2.95	2.97
Unanswerable	2.69	2.72	2.67	2.62

Table 10: Mean adaptive K per question type, by candidate pool ($K_{\max}$=5). Two patterns: (1) K is highly stable *across pools* (range 2.55–3.11), suggesting that the reranker's K choice is driven by the question's evidence requirement, not by which retriever fed the pool. (2) K varies *by question type*: Pure-text questions request the most pages ($\bar{K} \approx 3.0$, K=5 cap rate $\sim 30\%$); Figure / Chart questions request fewer ($\bar{K} \approx 2.6$, K=1 rate $> 30\%$ on Figure), consistent with visual evidence being concentrated on individual pages.

K Visual Caption QC: Substring and N-gram Overlap

To address the concern that VLM-generated captions might leak the gold answer text and thereby trivialize Stage 2 reranking, we audit 200 randomly sampled answerable QA / candidate-page pairs (N=238 caption-page records after dropping "Text-only page." captions).

- **Exact substring** (gold answer string contained verbatim in caption): $22/238 = 9.2\%$.
- **Bigram Jaccard** ≥ 0.3 (significant sentence-level overlap): $0/238 = 0.0\%$.
- **Unigram coverage** ≥ 0.7 of gold tokens in caption: $32/238 = 13.4\%$.

The 0% bigram match indicates no caption near-verbatim paraphrases the answer sentence; the 9.2% substring rate reflects captions correctly encoding visual content (e.g. chart legend colours such as "Blue", or diagram labels such as "Elastic Compute Service") that happen to coincide with the gold answer — the intended behaviour of caption-as-visual-evidence rather than answer leakage.

L Inter-Judge Agreement

We use two independent LLM judges (GPT-4.1 and Claude Sonnet 4.5) to score the same 300-QA end-to-end gen outputs from each of the 14 pipelines in Table 5. Both judges use the identical 3-level rubric (correct / partial / wrong) and the same MMLongBench-Doc gold answers. Table 11 reports per-pipeline pairwise agreement and Cohen's κ.

Pipeline	Agree.	Cohen κ
BM25 (Plain / MAVS)	94.0 / 96.7	0.89 / 0.94
BGE-M3 (Plain / MAVS)	94.3 / 97.3	0.90 / 0.95
ColPali (Plain / MAVS)	93.0 / 95.7	0.87 / 0.92
PageIndex (Plain / MAVS)	97.3 / 97.7	0.95 / 0.96
BGE-M3+ann. (Plain / MAVS)	93.7 / 96.3	0.89 / 0.93
TRIDENT-R (Plain / MAVS)	92.7 / 94.7	0.87 / 0.91
ColPali+ann. (Plain / MAVS)	93.0 / 95.7	0.88 / 0.92
Mean (14 pipelines)	**95.1%**	**0.913**
Range	92.7–97.7	0.87–0.96

Table 11: Inter-judge agreement and Cohen's κ between GPT-4.1 and Claude on the same 300-QA outputs. Mean κ=0.913 corresponds to "almost perfect agreement" on the Landis-Koch scale. The pipeline-level ranking under the two judges is identical (ColPali + ann. + rk + TRIDENT-S is best under both; TRIDENT-R + TRIDENT-S second under Claude, tied second under GPT). This indicates that the observed end-to-end ranking is not dependent on a single judge's quirks.

We do not perform human evaluation due to scale ($14 \times 300 = 4{,}200$ prediction-gold pairs); we view the high inter-judge agreement and the agreement of pipeline rankings between two independently developed LLMs as a reasonable proxy.

M Extended Limitations Discussion

We elaborate the limitations summarized in §5.

Reranker LLM dependence. The rerank stage in this work uses GPT-4.1 at τ=0. We expect the annotation-conditioned rerank protocol to be reranker-agnostic in principle, because the candidate JSON exposes per-page evidence in a form that any instruction-tuned LLM can consume. A pilot replacement with a smaller reranker would falsify or strengthen this expectation; we leave a full small-reranker / open-source-reranker sweep for follow-up work due to budget constraints.

LongDocURL coverage. LongDocURL serves as an external transfer check: the retrieval-side comparison (Table 4) reports adaptive-K methods on the full N=1,122 QAs with non-empty gold evidence, but the full five-pool cross-pool isolation conducted on MMLongBench-Doc is not separately replicated on LongDocURL. End-to-end LongDocURL results (Appendix A, Table 6) confirm the broad pattern.

TRIDENT-S on extractive workloads. On LongDocURL's MCQ-style and short-extractive questions, the topical, entity, and structural lenses converge on the same extracted answer; the synthesis call therefore contributes mean $\Delta \approx 0$ across four LongDocURL pipelines (Appendix A, Table 6). This motivates the on-demand deployment regime described in §3.5: TRIDENT-S is gated by question format.

Parser substitution. We use the PageIndex API as a structural parser for the section-path field while also benchmarking against the PageIndex retrieval system; the two roles are independent. Any public PDF parser that yields a section tree (PyMuPDF, GROBID) could fill this role. A parser-substitution robustness study is left to future work.