SURGE: SINKHORN-SEEDED GRAPH EXPANSION FOR MULTIMODAL LONG-DOCUMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation over long documents is increasingly important in practice, yet most existing approaches are designed for purely textual corpora and fail to handle the heterogeneous mix of text, tables, and figures found in real-world technical documents. We present SURGE, a two-stage retrieval framework for multimodal long-context documents. In the first stage, Sinkhorn-regularized optimal transport is applied over all pairwise chunk similarities to produce a globally calibrated co-association graph that encodes inter-chunk relationships across modalities without privileging text. In the second stage, Personalized PageRank is seeded from a query-proportional personalization vector and propagates relevance through this graph, surfacing tables and figures that cosine-similarity retrieval systematically misses due to the embedding-space modality gap. We evaluate SURGE on long multimodal datasets spanning up to 620K tokens across vision-language models of different sizes and demonstrate advantages over baselines on document querying. Our analysis further identifies multimodal document segmentation as the critical open bottleneck: retrieval quality is fundamentally chunking-quality-dependent, and we call for research on robust multimodal chunking pipelines.

1 INTRODUCTION

Question answering over a single long multimodal document is one of the most practically significant and least studied problems in document understanding. Real-world documents that demand expert reasoning, such as technical manuals, financial filings, engineering specifications, scientific papers, are not purely textual: they interleave running prose with tables, figures, diagrams, and cross-referenced visual elements, often across hundreds of pages. Systems that can effectively answer queries over such long, multimodal documents would have transformative applications in fields ranging from scientific research to finance, engineering, and beyond.

A large body of work addresses long-context understanding through architectural scaling (Chen et al., 2023; 2024), retrieval-augmented generation (Zhao et al., 2024), graph-based summarization (Edge et al., 2024), and agentic context management (Singh et al., 2025; Li et al., 2026). Nearly all of these methods are designed and evaluated on purely textual corpora; the multimodal setting, where a query must be answered from one long document containing text, tables, and figures, remains largely underexplored. This gap poses three open challenges that text-only methods sidestep. *Document segmentation* is itself unsolved, as real-world multimodal documents resist clean chunking due to OCR degradation, layout ambiguity, and the absence of ground-truth chunk boundaries (Ma et al., 2024; Tripathi et al., 2025). *Multimodal retrieval* is limited by the modality gap and embedding granuarity, where the former makes chunks of different modality occupy geometrically distinct regions of the embedding space and distorts inter-chunk relationships, and the latter poses challenge for one single chunk vector to encode all information in a text piece, table or figure. Retrieval becomes difficult due to the lack of inter-chunk relationship modeling and query-time augmentations. Also, the capability of the *VLM backbone* is critical, given that existing vision-language models experience degradation when reasoning across extended contexts of mixed-modality retrievals.

We focus on the second limitation and propose SURGE (**S**inkhorn-seeded **U**nified **R**etrieval via **G**raph **E**xpansion), a two-stage retrieval framework that decouples a query-agnostic preprocessing step from a lightweight query-time retrieval step. During *preprocessing*, SURGE applies Sinkhorn-regularized optimal transport over the full pairwise chunk similarity matrix to construct a globally

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.

calibrated co-association graph that captures fine-grained inter-chunk relationships across modalities. This graph is computed once per document and cached. At *query time*, SURGE seeds Personalized PageRank from a query-proportional personalization vector and propagates relevance through the precomputed graph, enabling more accurate query-chunk matching by reaching non-textual evidence through its document-structural neighbors rather than relying on each chunk's compressed embedding alone. Our contributions are as follows:

1. We introduce SURGE, combining Sinkhorn optimal transport and Personalized PageRank for multimodal long-document retrieval, and demonstrate advanced accuracy performance across datasets and baselines.
2. We curate three multimodal long-document evaluation datasets, MMLongBench-MQ, FinanceBench, and a custom Chip Dataset with human-annotated chunk boundaries and ground-truth answers, for controling the segmentation challenge and benchmarking.
3. As a result of our analysis, we identify several open questions including the multimodal document segmentation bottleneck and call for dedicated research to address them.

2 RELATED WORKS

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) grounds language-model outputs in evidence retrieved from an external index (Lewis et al., 2020; Gao et al., 2024). Beyond sparse lexical retrieval, dense and late-interaction methods such as ColBERT (Khattab & Zaharia, 2020) learn finer-grained query–passage matching, while HyDE (Gao et al., 2023) improves zero-shot retrieval through generated query representations. Later work introduces structure or adaptive control: RAPTOR (Sarthi et al., 2024) retrieves from a hierarchy of recursive summaries, GraphRAG (Edge et al., 2024) aggregates entity-level evidence over a knowledge graph, and Self-RAG and agentic RAG decide when or how to retrieve through learned or prompted reasoning loops (Asai et al., 2024; Singh et al., 2025).

Recent systems extend retrieval to document images. ColPali (Faysse et al., 2025) encodes each page as multi-vector visual representations and applies late interaction directly against a textual query. DMAP (Fu et al., 2026) combines image- and text-based retrieval with a hierarchical document map and a reflective agent that iteratively judges evidence sufficiency. These methods provide strong multimodal baselines, but ColPali does not explicitly model global inter-chunk structure, while DMAP introduces generative reasoning into the query-time retrieval loop. Our work instead constructs document structure offline and performs lightweight, non-generative graph propagation at retrieval time.

Long-Context Understanding. Long documents routinely exceed the effective processing capacity of current models. Larger context windows do not fully resolve this problem: models underuse evidence in the middle of long inputs (Liu et al., 2024), and attention sinks concentrate computation on a small set of initial tokens (Xiao et al., 2024). Positional interpolation (Chen et al., 2023) and LongLoRA (Chen et al., 2024) extend the sequences that pretrained models can process, but long-sequence attention remains computationally expensive and susceptible to quality degradation.

Retrieval offers a complementary solution by limiting the context presented to the model. LongRAG (Zhao et al., 2024) retrieves larger semantic units, while GraphRAG (Edge et al., 2024) uses corpus-level entity structure to aggregate distributed evidence. Agentic approaches instead treat context selection as sequential decision making: prompted agents refine retrieval iteratively (Singh et al., 2025), Sculptor learns context-management policies through reinforcement learning (Li et al., 2026), and DMAP applies reflective navigation to hierarchical multimodal document maps (Fu et al., 2026). Such systems can adapt their search to each query, but require additional generative inference; our method targets long-document structure without adding an agent to the retrieval path.

Multimodal Long-Context Retrieval. The challenge of long-context understanding becomes substantially harder when documents are multimodal, where the evidence relevant to a query may reside not only in running text but also in embedded tables, figures, charts, and diagrams. Real-world technical documents, financial filings, and scientific papers all exhibit this structure, yet most long-context methods remain designed and evaluated for plain text. Recent visual retrievers such as ColPali and map-based agents such as DMAP begin to address this setting, but robust segmenta-

tion, cross-modal retrieval, and downstream reasoning over heterogeneous evidence remain open problems.

A well-functioning multimodal long-context system must overcome at least three open problems:

1. *Document segmentation.* Partitioning a multimodal document into atomic retrieval units is itself an unsolved problem. Real-world documents impose additional difficulties through OCR degradation, layout parsing failures (Xu et al., 2020), and inherent ambiguity at the boundaries of running text, embedded tables, and inline figures (Blecher et al., 2023; Kim et al., 2022). Existing benchmarks provide no ground-truth chunk boundaries (Ma et al., 2024; Islam et al., 2023; Mathew et al., 2021), meaning that retrieval systems are evaluated end-to-end under a segmentation scheme they did not jointly optimize.
2. *Multimodal retrieval.* Even given clean segments, retrieving the right ones is harder when the query and the evidence occupy different modalities. Multimodal embedding models, including contrastive encoders descended from CLIP (Radford et al., 2021), embed text and non-textual content in a shared space but without an explicit objective that guarantees cross-modal semantic equivalence at the chunk level (See Appendix C.5). The resulting *modality gap* distorts inter-chunk relationships: text chunks cluster together and non-textual chunks cluster separately, so the co-association structure of the document is misrepresented in embedding space and the vanilla retrieval systematically underweights non-textual evidence regardless of its semantic relevance to the query. A compounding limitation is *embedding granularity*: a single chunk vector must compress all the information in a table or figure into one point, obscuring the fine-grained detail (e.g. specific cell values, timing annotations) that is decisive for accurate query-chunk matching—a query may be close in embedding space to a chunk's coarse topic yet far from the precise sub-element it actually asks about.
3. *VLM backbone capability.* Downstream accuracy also depends critically on whether the backbone vision-language model can reason faithfully over a retrieved multimodal context. Current VLMs, while strong on single-image benchmarks (OpenAI, 2024; Google DeepMind, 2025), degrade substantially when reasoning over extended mixed-modality contexts. Our experiments further reveal that this degradation is amplified on harder datasets: richer multimodal retrieval can exceed the reasoning capacity of weaker backbone models, creating a retrieval–reasoning gap that is largely absent in the text-only setting.

Our work focuses on the retrieval challenge, targeting both difficulties directly. The Sinkhorn stage constructs a globally calibrated co-association graph that captures fine-grained inter-chunk relationships across modalities, compensating for the distortion of modality gap. Personalized PageRank then propagates query relevance through this graph at query time, enabling more accurate query-chunk matching by surfacing non-textual evidence through its document-structural neighbors rather than relying on each chunk's compressed embedding alone.

3 METHODOLOGY

3.1 OVERVIEW

Given a multimodal document decomposed into a set of atomic chunks $\mathcal{V} = \{v_1, \ldots, v_n\}$, where $\mathcal{V} = \mathcal{T} \cup \mathcal{A} \cup \mathcal{F}$ denotes text paragraphs, tables, and figures respectively, and a natural-language query q, the retrieval task is to select a ranked set of k chunks from $\mathcal{V}$ that best supports answering q. We assume this decomposition is given and do not address the chunking problem [1]. A pretrained multimodal encoder $\phi : \mathcal{V} \to \mathbb{R}^d$ maps each chunk into a shared embedding space. Pairwise chunk similarities $S_{ij} = \mathrm{sim}(\phi(v_i), \phi(v_j))$ and query-chunk similarities $s_i = \mathrm{sim}(\phi(q), \phi(v_i))$ are computed via cosine similarity.

Retrieval over multimodal long documents is particularly challenging due to the heterogeneous and often chaotic nature of the chunk set: text, table, and figure chunks occupy different regions of the embedding space, and the semantic relationships between chunks are not well-captured by pairwise cosine similarities alone. We therefore seek a soft, global encoding of inter-chunk associations

[1] We discuss the chunking problem in Section 4.4.

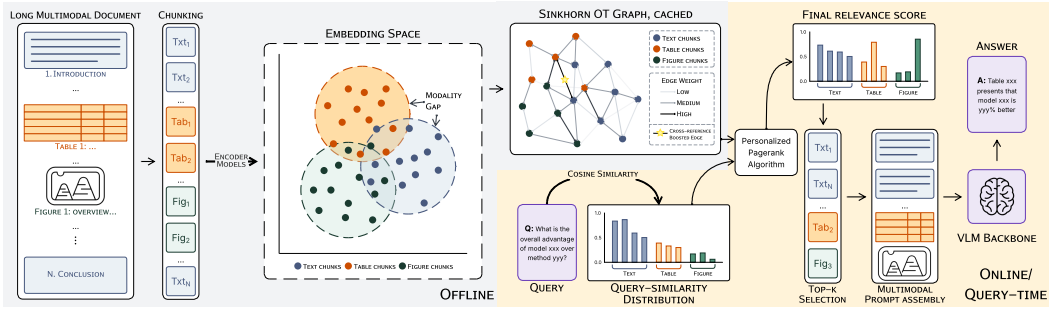


Figure 1: Overview of the SURGE pipeline. The Sinkhorn OT graph is constructed offline; at query time, Personalized PageRank propagates through the graph to produce the final ranked retrieval set.

that goes beyond local pairwise comparisons. Our approach proceeds in two stages: (i) Sinkhorn-regularized optimal transport to construct a globally calibrated co-association graph over all chunks, and (ii) Personalized PageRank to propagate initial query relevance through this graph and produce a final ranking. Figure 1 illustrates the full pipeline.

3.2 Sinkhorn Optimal Transport Graph

The Sinkhorn OT step is *query-agnostic*: given only the document chunks, it constructs a soft global graph encoding the co-association structure of the full chunk set. We convert the pairwise similarity matrix into a cost matrix $C_{ij} = 1 - \tilde{S}_{ij}$, so that semantically similar chunks incur lower transport cost. We define uniform marginal distributions $\boldsymbol{\mu} = \boldsymbol{\nu} = \frac{1}{n}\mathbf{1}_n \in \Delta^n$ over the n chunks, and solve the entropic optimal transport problem (Cuturi, 2013):

$$\min_{\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})} \langle \mathbf{C}, \boldsymbol{\pi} \rangle \; + \; \lambda \, \mathrm{KL}(\boldsymbol{\pi} \,\|\, \boldsymbol{\mu} \otimes \boldsymbol{\nu}), \tag{1}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ is the set of joint distributions with marginals $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, and $\lambda > 0$ is the entropic regularization coefficient.

Problem equation 1 is solved efficiently via Sinkhorn iterations (Algorithm 1). The doubly stochastic constraint on $\boldsymbol{\pi}$, where each row and column summing to $1/n$, is critical, because it ensures every chunk contributes and receives equal total flow, directly avoiding the hub-dominance problem of raw cosine similarity graphs where a few high-similarity chunks monopolize connectivity. The entry π_{ij} encodes the soft global co-association between v_i and v_j, calibrated simultaneously over all n chunks. Before row-normalization, the raw Sinkhorn solution satisfies this doubly stochastic property; we row-normalize $\boldsymbol{\pi}$ to obtain the stochastic transition matrix $\mathbf{W}$ used in the subsequent PageRank step.

Algorithm 1 Sinkhorn OT Graph Construction

Require: Similarity matrix $\tilde{\mathbf{S}}$, regularization λ, iterations T
Ensure: Transport plan $\boldsymbol{\pi}$

$C_{ij} \leftarrow 1 - \tilde{S}_{ij}$; $K_{ij} \leftarrow \exp(-C_{ij}/\lambda)$
Initialize $\mathbf{u}$, $\mathbf{v} \leftarrow \mathbf{1}_n$
for $t = 1$ **to** T **do**
  $\mathbf{u} \leftarrow \boldsymbol{\mu}/(\mathbf{K}\mathbf{v})$; $\mathbf{v} \leftarrow \boldsymbol{\nu}/(\mathbf{K}^\top \mathbf{u})$ ▷ element-wise
end for
$\pi_{ij} \leftarrow u_i K_{ij} v_j$

Cross-Reference Boosting: When readers navigate technical documents, a paragraph may explicitly direct attention to a remote figure or table (e.g., *"Table 3 shows..."* or *"as illustrated in Fig. 2..."*). Inspired by this pattern, we introduce a lightweight preprocessing step that mirrors this structure. When a text chunk v_i explicitly references a non-textual chunk v_j via patterns such as

"Table X", "Fig. X", or "tab. X", we set $\tilde{S}_{ij} = 1$ before computing the cost matrix, ensuring that the OT solver treats these pairs as maximally similar. This step is optional; otherwise $\tilde{\mathbf{S}} = \mathbf{S}$.

3.3 Personalized PageRank Retrieval

The Personalized PageRank step is *query-dependent*. For each query q, chunks that are more similar to q should seed more PageRank mass, so we weight every chunk proportionally to its query-chunk cosine similarity and L_1-normalize to obtain the personalization vector $\mathbf{r} \in \Delta^n$:

$$r_i = \frac{s_i}{\sum_j s_j}. \tag{2}$$

We use $\mathbf{r}$ as the starting point of Personalized PageRank (Page et al., 1999) on $\mathbf{W}$, with $W_{ij} = \pi_{ij} / \sum_{j'} \pi_{ij'}$ denoting the row-normalized transition matrix. The score vector $\mathbf{p}$ is the stationary distribution of

$$\mathbf{p} = (1 - \alpha)\, \mathbf{W}^\top \mathbf{p} \; + \; \alpha\, \mathbf{r}, \tag{3}$$

where $\alpha \in (0, 1)$ is the teleportation probability and $\mathbf{W}^\top$ accumulates *in-flow* from neighbours (each column of $\mathbf{W}$ distributes mass to a node's in-links). The fixed point is found by power iteration (Algorithm 2). Chunks strongly connected in the OT graph to query-relevant chunks receive elevated scores even when their direct query similarity is low, naturally compensating for the modality gap affecting non-textual chunks.

Algorithm 2 Personalized PageRank Retrieval

Require: Transition matrix $\mathbf{W}$, personalization vector $\mathbf{r}$, teleportation α, iterations T
Ensure: Ranked chunk scores $\mathbf{p}$
1: Initialize $\mathbf{p} \leftarrow \mathbf{r}$
2: **for** $t = 1$ **to** T **do**
3: $\quad \mathbf{p} \leftarrow (1 - \alpha)\, \mathbf{W}^\top \mathbf{p} + \alpha\, \mathbf{r}$
4: **end for**
5: **return** top-k indices of $\mathbf{p}$

3.4 Inference

Since the Sinkhorn OT graph depends only on the document and not on any specific query, it is computed once per document and cached. At query time, only the lightweight PPR step is executed, making the overall pipeline efficient. The top-k chunks ranked by p_i are concatenated in document order to form a multimodal context: text chunks are included as text and table or figure chunks are included as images. This context, together with the query q, is presented as a prompt to a backbone VLM to produce the final answer in a generative way. Additional details see Appendix B.3.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on three datasets of long, multimodal documents spanning different domains. **MMLongBench-MQ** is derived from MMLongBench (Ma et al., 2024): starting from the full benchmark, we exclude documents below a minimum length threshold and queries whose answers require no visual or tabular evidence, retaining only instances that demand genuine reasoning over multimodal content in long documents. **FinanceBench** (Islam et al., 2023) consists of corporate financial reports requiring precise numerical reasoning over tables and figures. **Chip Dataset** is a custom collection of microprocessor and chip architecture documentation that is dense in technical diagrams, specification tables, and cross-referenced text. Together, the three datasets span document lengths from under 2K to over 620K tokens and contain rich mixtures of text, tables, and figures throughout, making them among the most demanding long-context multimodal evaluation settings available. We leave tables as images instead of structured data to prevent additional noise. For all

Dataset	#Documents (Queries)	Min #Tokens	Max #Tokens	Avg. #Tokens
MMLongBench-MQ	60 (60)	10,598	332,377	42,683
FinanceBench	84 (150)	1,933	526,931	116,460
Chip Dataset	6 (72)	415,969	620,250	522,472

Table 1: Overview of the datasets used in our evaluation. We record the number of documents in each dataset, the token counts of the largest and the smallest documents, and the average.

	Representation	Document preprocessing		Query-time retrieval	
Method	backend	Representation	Additional index	Query encoding	Retrieval/navigation
SURGE	Voyage MM (API)	12.24 s	84.5 ms	162.6 ms	0.99 ms
ColPali	ColQwen2-v1.0 (GPU)	18.88 s	N/A	27.7 ms	1.64 ms
DMAP	Voyage MM + ColQwen2	31.12 s	30.1 ms	190.3 ms	5.84 ms

Table 2: Stage-separated mean computational cost on FinanceBench. Representation and query-encoding times are reported separately because Voyage is a managed API whereas ColQwen2 runs locally on a GPU. *Additional index* is the work performed after representations are cached; *retrieval/navigation* measures the retrieval kernel by either the graph expansion or the DMAP-style decisions. The final answer generations are excluded from all latency columns.

three datasets, documents are human-annotated to assign each chunk to the appropriate semantic unit, eliminating the confounding effect of chunking quality on retrieval performance. An overview of dataset statistics is given in Table 1; the full preprocessing pipeline for each dataset is described in Appendix B.1. We will release all three datasets together with our main method implementation.

Baselines. We compare SURGE against four baselines. **Top-k RAG** is the standard pipeline that ranks all chunks by cosine similarity to the query embedding and returns the top k. **GraphRAG** (Edge et al., 2024) constructs a knowledge graph from document text and uses entity-aware ranking for retrieval. Since GraphRAG operates exclusively on text, we provide it with enriched text chunks that redundantly incorporate tabular and visual information via textual descriptions, placing it in a deliberately favourable environment for multimodal documents. We also compare our method with **ColPali**(Faysse et al., 2025) and **DMAP**(Fu et al., 2026), two recent strong document-retrieval methods. For the fairness of the experimentations, we use, by default, `voyage-multimodal-3` to produce multimodal embeddings and query embeddings and pose no restrictions on additional index construction. For methods like Colpali that uses a pre-trained VLM for retrieval, we use their open-sourced weights. Appendix B.2 reports the implementation and stage-separated timing details.

Metrics. We evaluate the retrieval quality using mean reciprocal rank (MRR), recall at k (R@k), and normalized discounted cumulative gain at k (nDCG@k). Then, we report methods' answer accuracy on each query. The criterion for examining the accuracy is strict: a query is considered correctly answered only if the backbone model answers *every* sub-question it contains correctly given the retrieved chunks. Many queries in our evaluation require jointly identifying a numerical value, its unit, and a contextual qualifier from a table or figure and necessary reasoning; a partial answer is marked incorrect. Answers are verified by an LLM judge with human checking and validation. We test each retrieval method across six backbone vision-language models spanning a range of scales and providers: *GPT-5.2* and *Gemini 2.5 Pro* as large frontier models; *GPT-4o*, *Gemini 2.5 Flash*, and *Qwen3-VL-32B* as mid-range models; and *Qwen3-VL-8B* as a compact model. This breadth allows us to assess whether retrieval improvements are consistent across backbone capability levels.

4.2 OVERALL RESULTS

We begin our analysis with computational costs. Table 2 decomposes the computational cost by stage. Including representation encoding and index construction, SURGE achieves the lowest observed preprocessing cost, at 12.32 s per document, compared with ColPali and DMAP. With representations cached, SURGE also achieves the fastest query-time retrieval kernel, at 0.99 ms, compared with 1.64 ms for ColPali and 5.84 ms for DMAP's multi-stage navigation. Although a strictly fair comparison is difficult because the baselines rely on different embedding models and preprocessing

Strategy	MRR	R@5	R@10	nDCG@5	nDCG@10
Top-k RAG	0.325	0.337	0.493	0.280	0.319
GraphRAG	0.318	0.379	<u>0.502</u>	0.279	0.324
ColPali	0.294	0.337	0.446	0.263	0.302
DMAP	**0.340**	**0.389**	0.500	**0.309**	**0.352**
SURGE	<u>0.327</u>	<u>0.383</u>	**0.503**	<u>0.282</u>	<u>0.327</u>

Table 3: Retrieval quality on FinanceBench. Best results are shown in **bold** and second-best results are <u>underlined</u>.

Method	MMLongBench-MQ	FinanceBench	Chip Dataset	Average
Top-k RAG	<u>54.1</u>	58.0	<u>52.8</u>	<u>55.0</u>
GraphRAG	34.6	56.7	50.0	47.1
ColPali	39.0	44.7	45.6	43.1
DMAP	39.2	<u>63.3</u>	45.6	49.4
SURGE	**56.2**	**64.0**	**56.9**	**59.0**

Table 4: Best answer accuracy (%) achieved by each retrieval method over the six backbone VLMs. The final column averages the three dataset-best scores. The best result in each column is shown in **bold**. Second-best results are underlined.

pipelines, our results demonstrate the query-time efficiency of SURGE and its independence from locally deployed, specialized visual models. Appendix C.1 reports the complete runtime breakdown and total costs.

We next examine retrieval performance. We evaluate on FinanceBench because it provides a relatively well-structured benchmark with a sufficient number of contextual passages. Table 3 shows that SURGE remains competitive with state-of-the-art multimodal retrievers. It outperforms ColPali across all five retrieval metrics and trails DMAP by only a small margin. DMAP achieves the highest MRR and nDCG scores. However, these overlap-based metrics implicitly assume that the annotated chunks constitute the unique valid evidence path. In long documents, relevant information may appear in redundant tables, adjacent explanatory text, cross-references, or repeated specifications. As a result, a retrieval method may identify an alternative evidence set that fully supports the answer while receiving little or no credit relative to the single annotated evidence path. Appendix C.3 provides a qualitative example of this phenomenon.

The downstream answer-quality results provide further evidence for our hypothesis. Table 4 reports the best accuracy achieved by each retrieval method across the six backbone VLMs; complete backbone-wise results are provided in Appendix C.2 due to space constraints. SURGE achieves the best result on every dataset: 56.2% on MMLongBench-MQ, 64.0% on FinanceBench, and 56.9% on the Chip Dataset. Its average across the three dataset-level best scores is approximately 10% higher than those of ColPali and DMAP, with basic Top-k RAG being the closest competitor. These results suggest that SURGE is broadly comparable to the strongest baselines under strict annotated-overlap metrics while providing more useful evidence to the downstream VLM. We attribute part of this gap to the alternative-evidence-path phenomenon discussed above and use these observations to support the argument that conventional retrieval metrics can understate practical answer quality in long, redundant documents.

4.3 ABLATION STUDIES

Core Algorithmic Components. We first ablate the two core components of SURGE to examine where the performance gain comes from. We introduce two variants: **w/o Sinkhorn** and **w/o PageRank**. For the former, we replace the Sinkhorn OT graph with a cosine-similarity graph. Since raw cosine similarities do not directly provide a calibrated graph structure, we dynamically select an edge threshold by fitting a Gaussian Mixture Model (GMM) to the pairwise similarity distribution. For the latter, we retain the Sinkhorn transport plan and start with the query-similarity. For each

Methods	MMLongBench-MQ	FinanceBench	Chip Dataset
Ours	50.3	64.0	56.9
Ours w/o Sinkhorn	11.9	48.7	31.9
Ours w/o PageRank	34.6	54.7	55.6

Table 5: Ablation study on the core algorithmic components of our retrieval system. Accuracy is reported across all three datasets using the GPT-5.2 backbone.

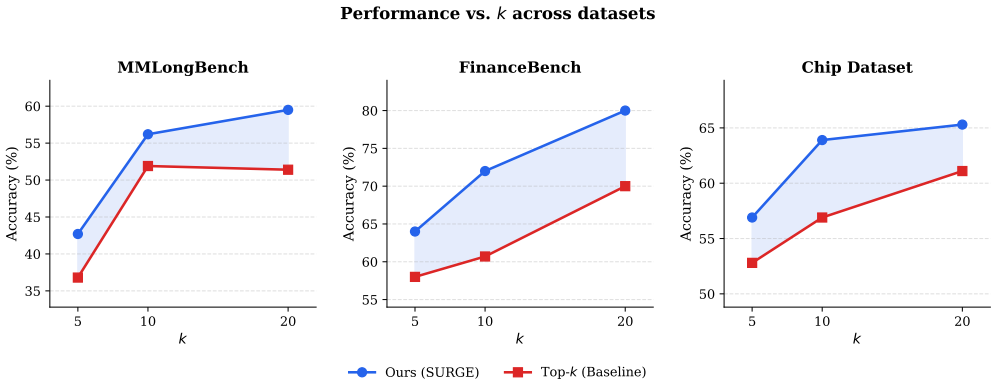


Figure 2: Accuracy as a function of the number of retrieved chunks k for SURGE and Top-k RAG across all three datasets. SURGE outperforms the baseline at every k and reaches a plateau earlier, indicating that PPR concentrates the most informative chunks near the top of the ranking.

selected seed chunk, its Sinkhorn co-associated neighbours that exceeds a threshold t are greedily added until the retrieval budget k is filled. t is set to 0.2 after testing.

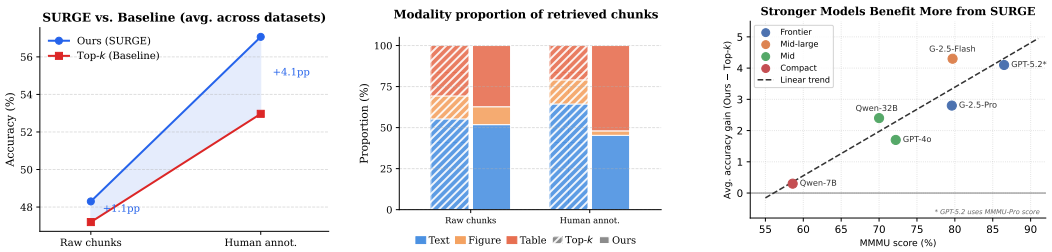
All variants use the GPT-5.2 backbone and are evaluated in Table 5. Removing Sinkhorn causes the largest degradation. The severe collapse suggests that cosine-based graph construction fails to provide a reliable global co-association structure: without the doubly stochastic OT constraint, the graph is more vulnerable to high-similarity hubs and noisy same-modality connections, leaving PageRank with weak structural signal to propagate. Removing PageRank also substantially hurts performance. This confirms that the Sinkhorn graph alone is insufficient: while it captures inter-chunk relationships, PPR is needed to transform these relationships into a query-conditioned global ranking. Overall, the ablation supports the design of SURGE as a coupled system, where Sinkhorn provides globally calibrated document structure and PPR performs principled relevance propagation over that structure.

Sensitivity to k. We next study the sensitivity of SURGE to the retrieval budget k, which controls how many chunks are passed to the backbone model. We compare SURGE against Top-k RAG under GPT5.2 with varying k choices on all three datasets.

The results are shown in Figure 2. Across all tested values of k, SURGE consistently outperforms Top-k RAG. Moreover, SURGE reaches its performance plateau earlier: increasing k beyond 10 yields only limited additional gains, whereas Top-k RAG benefits more gradually from larger retrieval budgets. This indicates that SURGE concentrates relevant evidence near the top of the ranking by propagating query relevance through the Sinkhorn co-association graph.

4.4 ADDITIONAL DISCOVERIES

Chunking quality. The first challenge we proposed in Section 2 calls for better chunking quality. Our main experiments use human-annotated chunks to isolate parsing errors. Figure 3a compares this setting with the initial stage of the preprocessing, where an automatic pipeline using standard layout heuristics. Automated chunking lowers accuracy for all methods, but SURGE still consistently outperforms Top-k RAG. This shows that SURGE is robust to imperfect chunking, but still depends on chunk quality. OCR errors, wrong boundaries, and poorly extracted tables or figures can limit



(a) Effect of chunking pipeline. (b) Modality proportion of chunks. (c) Acc. gain vs. VLM capability.

Figure 3: (a) Accuracy of SURGE and Top-k RAG under an automatic and human-based chunking pipeline. (b) The modality proportion of the top-5 retrieved chunks on the Chip dataset, with comparison between retrieval methods and chunking pipelines. (c) Average accuracy gain of SURGE over Top-k RAG as a function of VLM multimodal capability; the dashed line suggests a linear fit.

retrieval effectiveness. We list some examples in Appendix C.4. Thus, multimodal chunking remains an important open problem rather than a solved preprocessing step.

Retrieved modality composition. Since queries are textual, Top-k RAG is strongly biased toward text chunks and retrieves fewer tables or figures. This reflects the modality gap in the embedding space as mentioned. In the Chip dataset, we inspect the modality composition of chunks retrieved by our method and the baseline Top-k to explore if our method changes the modality distribution. The results are shown in Figure 3b. SURGE retrieves more tables and figures by propagating relevance from query-aligned text chunks to associated non-text chunks through the Sinkhorn OT graph. Moreover, such trend, despite weaker, is also observed in the non-human-annotated chunks, suggesting a more global enhancement of modality balance.

Backbone capability. The benefit of retrieval also depends on if the backbone model can use the retrieved evidence. Figure 3c shows the performance gain of backbones from SURGE against its reported MMMU score, a metric for multimodal reasoning capability. The results suggest that the gains of our method increase with backbone strength. For mid-to-large VLMs, SURGE gives consistent gains over Top-k RAG. For the compact models, however, the improvement is limited, suggesting that richer multimodal contexts do not automatically translate into better answers. This result indicates that retrieval and reasoning capacity must improve together, which agrees with the third challenge we proposed in Section 2.

5 LIMITATIONS

Despite the strong performance of SURGE over existing baselines, our method still depends on the quality of document chunking. In addition, although our method are evaluated on long documents with 300–500K context length, the lack of off-the-shelf benchmarks with reliable multimodal chunk annotations prevents us from evaluating this setting more comprehensively. We therefore view high-quality long-document multimodal chunking and benchmark construction as important future directions. We also provide a failure case study in Appendix C.6.

6 CONCLUSION

We presented SURGE, a graph-based retrieval framework for multimodal long-context documents that combines Sinkhorn-regularized optimal transport with Personalized PageRank. The Sinkhorn step constructs a globally calibrated co-association graph over all document chunks, and PPR propagates query relevance through it to surface non-textual evidence that cosine-similarity retrieval systematically misses. SURGE demonstrates performance advantages across all three evaluation datasets and all backbone models tested, with the largest gains on documents rich in cross-modal structure. We also demonstrated the how improved chunking quality and backbone capability benefits our method, and therefore encourage researches on high-quality multimodal chunking on long document cases.

AI USE STATEMENT

In this work, we used generative AI tools only to revise author-written text for clarity, grammar, and style. All AI-assisted revisions were reviewed and approved by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

MMLongBench is released under the Apache License 2.0, and FinanceBench is released under the CC BY-NC license. Our Chip Dataset is constructed from publicly available online chip manuals, which we will identify the files and describe the modification to be made to construct our dataset upon release. We use these resources only for research and evaluation purposes, and follow the corresponding license terms and attribution requirements.

All human annotation and verification work in this project was conducted by recruited volunteers, who were properly compensated for their time and labor. No third-party annotators, crowdworkers, or external labor platforms were involved. The annotation procedure did not involve sensitive personal information, harmful content, or tasks that could reasonably cause harm to annotators or third parties. No third-party members were harmed or exploited during the construction and evaluation of the datasets.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11511.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. In *arXiv preprint arXiv:2308.13418*, 2023. URL https://arxiv.org/abs/2308.13418.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023. URL https://arxiv.org/abs/2306.15595.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. LongLoRA: Efficient fine-tuning of long-context large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.12307.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013. URL https://papers.nips.cc/paper/4927.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2025. URL https://arxiv.org/abs/2407.01449.

ShunLiang Fu, Yanxin Zhang, Yixin Xiang, Xiaoyu Du, and Jinhui Tang. DMAP: Human-aligned structural document map for multimodal document understanding, 2026. URL https://arxiv.org/abs/2601.18203.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://arxiv.org/abs/2212.10496.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024. URL https://arxiv.org/abs/2312.10997.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023. URL https://arxiv.org/abs/2311.11944.

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*, 2020. URL https://arxiv.org/abs/2004.12832.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2111.15664.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2005.11401.

Mo Li, L.H. Xu, Qitai Tan, Long Ma, Flood Sung, Ting Cao, and Yunxin Liu. Sculptor: Empowering LLMs with cognitive agency via active context management. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HPeiH7da0Z.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. URL https://arxiv.org/abs/2307.03172.

Yubo Ma, Yuhang Zang, Liangyu Xu, Meiqi Chen, Yizhu Huang, Xinyu Tang, Jiajun Yang, et al. MMLongBench-Doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.01523.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. URL https://arxiv.org/abs/2007.00398.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999. URL http://ilpubs.stanford.edu:8090/422/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2103.00020.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.18059.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025. URL https://arxiv.org/abs/2501.09136.

Vishesh Tripathi, Tanmay Odapally, Indraneel Das, Uday Allu, and Biddwan Ahmed. Vision-guided chunking is all you need: Enhancing rag with multimodal document understanding, 2025. URL https://arxiv.org/abs/2506.16035.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.17453.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. URL https://arxiv.org/abs/1912.13318.

Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22600–22632, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1259. URL https://aclanthology.org/2024.emnlp-main.1259/.

A Additional Information on the Methodology

A.1 Theoretical Interpretation under the Modality Gap

We provide a theoretical interpretation of why SURGE can retrieve relevant non-textual chunks that are missed by direct cosine-similarity retrieval. Rather than assuming that the modality gap arises from a particular geometric source such as manifold curvature, we model it directly as a mismatch between two notions of relevance: direct query–chunk similarity and document-level co-association. This better matches the actual mechanism of SURGE, which first constructs a query-agnostic Sinkhorn co-association graph and then propagates query relevance through Personalized PageRank.

Setup. Let $\mathcal{V} = \{v_1, \dots, v_n\}$ be the chunk set of a document, containing text, table, and figure chunks. Let $\mathbf{s} \in \mathbb{R}^n$ denote the direct query–chunk similarity scores computed by the multimodal encoder. Since Personalized PageRank requires a probability vector, we assume that the scores are transformed into a nonnegative personalization vector

$$\mathbf{r} \in \Delta^n, \qquad r_i \geq 0, \qquad \sum_i r_i = 1, \tag{4}$$

for example by shifting, clipping, or softmax-normalizing the cosine similarities.

Let $\boldsymbol{\pi}$ be the Sinkhorn transport plan computed from the pairwise chunk similarity matrix, with uniform marginals

$$\boldsymbol{\mu} = \boldsymbol{\nu} = \frac{1}{n}\mathbf{1}_n. \tag{5}$$

The row-normalized transition matrix is denoted by $\mathbf{W}$. When no sparsification is applied, the Sinkhorn marginal constraints imply

$$W_{ij} = n\pi_{ij}, \qquad \sum_j W_{ij} = 1, \qquad \sum_i W_{ij} = 1. \tag{6}$$

Thus, $\mathbf{W}$ is doubly stochastic.

The Personalized PageRank score vector $\mathbf{p}$ is defined as the unique fixed point

$$\mathbf{p} = (1-\alpha)\mathbf{W}^\top \mathbf{p} + \alpha \mathbf{r}, \qquad \alpha \in (0,1). \tag{7}$$

Modality gap. We say that a relevant chunk $c^\star$ suffers from a modality gap with respect to a query q if its direct query score is lower than that of an irrelevant same-modality or text-dominant chunk z:

$$r_z > r_{c^\star}. \tag{8}$$

In this case, direct cosine retrieval prefers z over $c^\star$, even though $c^\star$ may contain the information required to answer the query. The goal of the graph propagation step is to overcome this direct-score deficit by allowing $c^\star$ to receive relevance mass from related chunks that are easier for the query encoder to retrieve.

Lemma 1 (Sinkhorn flow balancing)**.** *Assume the Sinkhorn OT problem is solved with uniform marginals and no sparsification is applied. Let $W_{ij} = n\pi_{ij}$. Then $\mathbf{W}$ is doubly stochastic. Moreover, for any distribution $\mathbf{x} \in \Delta^n$,*

$$\|\mathbf{W}^\top \mathbf{x}\|_\infty \leq \|\mathbf{x}\|_\infty. \tag{9}$$

Proof. The Sinkhorn marginal constraints give

$$\sum_j \pi_{ij} = \frac{1}{n}, \qquad \sum_i \pi_{ij} = \frac{1}{n}.$$

Multiplying by n gives

$$\sum_j W_{ij} = 1, \qquad \sum_i W_{ij} = 1.$$

Therefore, $\mathbf{W}$ is doubly stochastic. For any coordinate j,

$$(\mathbf{W}^\top \mathbf{x})_j = \sum_i W_{ij} x_i \le \|\mathbf{x}\|_\infty \sum_i W_{ij} = \|\mathbf{x}\|_\infty.$$

□

Lemma 1 formalizes the balancing effect of Sinkhorn OT. Unlike a raw cosine graph, where a small number of high-similarity hubs may receive disproportionately large incoming mass, the doubly stochastic Sinkhorn graph limits such hub domination by constraining both outgoing and incoming flow.

Lemma 2 (Path expansion of Personalized PageRank)**.** *The solution to Eq. equation 7 admits the expansion*

$$\mathbf{p} = \alpha \sum_{\ell=0}^{\infty} (1-\alpha)^\ell (\mathbf{W}^\top)^\ell \mathbf{r}. \tag{10}$$

Consequently, for any chunk v,

$$p_v = \alpha r_v + \alpha(1-\alpha) \sum_u W_{uv} r_u + \alpha(1-\alpha)^2 \sum_{u_1,u_2} W_{u_2 v} W_{u_1 u_2} r_{u_1} + \cdots. \tag{11}$$

Proof. Rearranging Eq. equation 7 gives

$$\mathbf{p} = \alpha \left(\mathbf{I} - (1-\alpha)\mathbf{W}^\top\right)^{-1} \mathbf{r}.$$

Since $\mathbf{W}$ is row-stochastic, the spectral radius of $(1-\alpha)\mathbf{W}^\top$ is at most $1-\alpha < 1$. The inverse therefore has the Neumann-series expansion

$$\left(\mathbf{I} - (1-\alpha)\mathbf{W}^\top\right)^{-1} = \sum_{\ell=0}^{\infty} (1-\alpha)^\ell (\mathbf{W}^\top)^\ell,$$

which proves the result. □

Lemma 2 shows that SURGE does not rely only on direct query similarity. A chunk receives score from all paths in the Sinkhorn graph that terminate at it, with paths of length ℓ discounted by $(1-\alpha)^\ell$. Thus, a relevant non-textual chunk can be recovered even when its direct score $r_{c^\star}$ is small, provided that it is strongly connected to chunks with high query relevance.

Theorem 1 (Recovery condition under a modality gap)**.** *Let $c^\star$ be a relevant chunk and let z be a competing irrelevant chunk such that direct retrieval prefers z:*

$$r_z > r_{c^\star}. \tag{12}$$

Then SURGE *ranks $c^\star$ above z, i.e. $p_{c^\star} > p_z$, if and only if*

$$\sum_{\ell=1}^{\infty} (1-\alpha)^\ell \left[\left((\mathbf{W}^\top)^\ell \mathbf{r}\right)_{c^\star} - \left((\mathbf{W}^\top)^\ell \mathbf{r}\right)_z \right] > r_z - r_{c^\star}. \tag{13}$$

Proof. Using Lemma 2,

$$p_{c^\star} - p_z = \alpha(r_{c^\star} - r_z) + \alpha \sum_{\ell=1}^{\infty} (1-\alpha)^\ell \left[\left((\mathbf{W}^\top)^\ell \mathbf{r}\right)_{c^\star} - \left((\mathbf{W}^\top)^\ell \mathbf{r}\right)_z \right].$$

Since $\alpha > 0$, the inequality $p_{c^\star} > p_z$ is equivalent to Eq. equation 13. □

Theorem 1 gives a direct theoretical boundary for SURGE. The left-hand side is the graph-propagated advantage of the relevant chunk, while the right-hand side is its direct cosine-similarity deficit. Therefore, SURGE corrects a modality-gap error exactly when the Sinkhorn graph supplies enough propagated relevance mass to overcome the direct retrieval disadvantage.

Corollary 1 (One-hop bridging condition). *Let $c^\star$ and z be as in Theorem 1. For any integer $L \geq 1$, define the finite-horizon graph advantage*

$$A_L(c^\star, z) = \sum_{\ell=1}^{L} (1-\alpha)^\ell \left[\left((\mathbf{W}^\top)^\ell \mathbf{r} \right)_{c^\star} - \left((\mathbf{W}^\top)^\ell \mathbf{r} \right)_z \right]. \tag{14}$$

If

$$A_L(c^\star, z) > r_z - r_{c^\star} + \frac{(1-\alpha)^{L+1}}{\alpha}, \tag{15}$$

then $p_{c^\star} > p_z$. In particular, for $L = 1$, a sufficient condition is

$$(1-\alpha) \left[\sum_u W_{uc^\star} r_u - \sum_u W_{uz} r_u \right] > r_z - r_{c^\star} + \frac{(1-\alpha)^2}{\alpha}. \tag{16}$$

Proof. The infinite tail after length L satisfies

$$\left| \sum_{\ell=L+1}^{\infty} (1-\alpha)^\ell \left[\left((\mathbf{W}^\top)^\ell \mathbf{r} \right)_{c^\star} - \left((\mathbf{W}^\top)^\ell \mathbf{r} \right)_z \right] \right| \leq \sum_{\ell=L+1}^{\infty} (1-\alpha)^\ell = \frac{(1-\alpha)^{L+1}}{\alpha},$$

because each $(\mathbf{W}^\top)^\ell \mathbf{r}$ is a probability vector. Therefore, Eq. equation 15 implies the infinite-horizon condition in Theorem 1. □

Corollary 1 formalizes the role of bridging chunks. If a text chunk or nearby evidence chunk receives high direct query mass and assigns sufficient Sinkhorn flow to a relevant non-textual chunk, then the non-textual chunk can overtake irrelevant chunks that were preferred by direct cosine similarity. This is precisely the regime where graph-based propagation is expected to improve over Top-k RAG.

Parameter interpretation. The teleportation probability α controls the tradeoff between direct retrieval and graph propagation. As $\alpha \to 1$, Eq. equation 7 gives $\mathbf{p} \to \mathbf{r}$, so SURGE reduces to direct cosine retrieval and cannot correct modality-gap errors. As $\alpha \to 0$, long graph paths dominate and the query signal may be diluted. Thus, the useful regime is an intermediate one, where short document-structure paths receive enough weight to recover cross-modal evidence while the ranking remains anchored to the query.

The Sinkhorn regularization parameter controls the softness of the document co-association graph. If the regularization is too large, the transport plan approaches a near-uniform coupling and loses document-specific structure. If it is too small, the plan becomes overly concentrated and may fail to preserve useful secondary associations. The empirical benefit of SURGE is therefore expected to be strongest when the Sinkhorn graph captures reliable cross-modal associations without collapsing to either a uniform graph or a nearly deterministic matching.

A.2 PERSONALIZED PAGERANK STABILITY UNDER GRAPH SPARSIFICATION

For efficiency, SURGE may sparsify the row-normalized transition matrix $\mathbf{W}$ derived from the Sinkhorn plan. We analyze this operation directly as a perturbation of the PPR transition matrix. This avoids requiring the sparsified graph to satisfy a full spectral approximation guarantee, which is stronger than what simple thresholding or top-r retention necessarily provides.

Let $\mathbf{W}'$ be the sparsified and row-renormalized transition matrix. Let $\mathbf{p}$ and $\mathbf{p}'$ denote the PPR vectors computed using $\mathbf{W}$ and $\mathbf{W}'$, respectively, with the same personalization vector $\mathbf{r}$ and the same teleportation probability α:

$$\mathbf{p} = (1-\alpha)\mathbf{W}^\top \mathbf{p} + \alpha \mathbf{r}, \qquad \mathbf{p}' = (1-\alpha)(\mathbf{W}')^\top \mathbf{p}' + \alpha \mathbf{r}. \tag{17}$$

Theorem 2 (PPR stability under transition perturbation). *Suppose $\mathbf{W}$ and $\mathbf{W}'$ are row-stochastic. Then*

$$\|\mathbf{p} - \mathbf{p}'\|_1 \leq \frac{1-\alpha}{\alpha} \left\| (\mathbf{W}^\top - (\mathbf{W}')^\top) \mathbf{p}' \right\|_1. \tag{18}$$

Consequently,

$$\|\mathbf{p} - \mathbf{p}'\|_1 \leq \frac{1-\alpha}{\alpha}\|\mathbf{W} - \mathbf{W}'\|_\infty, \tag{19}$$

where

$$\|\mathbf{W} - \mathbf{W}'\|_\infty = \max_i \sum_j |W_{ij} - W'_{ij}|.$$

Proof. Subtracting the two fixed-point equations gives

$$\mathbf{p} - \mathbf{p}' = (1-\alpha)\mathbf{W}^\top(\mathbf{p} - \mathbf{p}') + (1-\alpha)(\mathbf{W}^\top - (\mathbf{W}')^\top)\mathbf{p}'.$$

Thus,

$$\mathbf{p} - \mathbf{p}' = (1-\alpha)\left(\mathbf{I} - (1-\alpha)\mathbf{W}^\top\right)^{-1}(\mathbf{W}^\top - (\mathbf{W}')^\top)\mathbf{p}'.$$

Since $\mathbf{W}$ is row-stochastic, $\|\mathbf{W}^\top\|_1 = 1$, and therefore

$$\left\|\left(\mathbf{I} - (1-\alpha)\mathbf{W}^\top\right)^{-1}\right\|_1 \leq \sum_{t=0}^{\infty}(1-\alpha)^t = \frac{1}{\alpha}.$$

This proves Eq. equation 18. The second inequality follows from

$$\left\|(\mathbf{W}^\top - (\mathbf{W}')^\top)\mathbf{p}'\right\|_1 \leq \|\mathbf{W}^\top - (\mathbf{W}')^\top\|_1\|\mathbf{p}'\|_1 = \|\mathbf{W} - \mathbf{W}'\|_\infty,$$

and $\|\mathbf{p}'\|_1 = 1$. □

Corollary 2 (Stability of threshold sparsification)**.** *Suppose $\mathbf{W}'$ is obtained by deleting, in each row i, a set of edges with total mass at most ϵ_i, and then renormalizing the remaining row. Let*

$$\epsilon_{\max} = \max_i \epsilon_i.$$

Then

$$\|\mathbf{p} - \mathbf{p}'\|_1 \leq \frac{2(1-\alpha)}{\alpha}\epsilon_{\max}. \tag{20}$$

More generally, the sharper weighted bound

$$\|\mathbf{p} - \mathbf{p}'\|_1 \leq \frac{2(1-\alpha)}{\alpha}\sum_i p'_i\epsilon_i \tag{21}$$

holds.

Proof. For row i, let ϵ_i be the total probability mass removed by sparsification. After renormalization, the retained entries are scaled by $1/(1-\epsilon_i)$. The ℓ_1 difference between the original and sparsified row is therefore

$$\sum_j |W_{ij} - W'_{ij}| \leq 2\epsilon_i.$$

Applying Theorem 2 gives the result. □

Consequence for SURGE. Corollary 2 shows that sparsification is safe when the removed Sinkhorn mass is small on rows that are important under the PPR distribution. The minimum-degree constraint prevents dangling or isolated nodes, while thresholding removes low-mass associations that contribute little to the transition distribution. With $\alpha = 0.70$, the prefactor $(1-\alpha)/\alpha$ is moderate, so small row-wise perturbations of $\mathbf{W}$ lead to controlled perturbations of the final PPR scores. This explains why SURGE is empirically robust to aggressive sparsification when the Sinkhorn mass is concentrated on a small number of meaningful document associations.

B Additional Implementation Details

B.1 Data Preprocessing Pipeline

Our preprocessing pipeline aims to produce near-optimal chunk boundaries with minimal confounding from segmentation noise, allowing us to evaluate retrieval in isolation. The pipeline proceeds in three stages.

Stage 1: Automated extraction. Raw PDFs are processed with PyMuPDF to extract text blocks and detect bounding boxes for candidate table and figure regions. Text blocks are extracted directly where the PDF layer is intact, with OCR applied as a fallback for scanned or image-rendered pages. Non-textual regions are cropped from the page using the detected bounding boxes and saved as image patches; captions are extracted separately from the surrounding text. Each chunk at this stage is assigned a provisional index corresponding to its position in reading order.

Stage 2: LLM-assisted correction. An LLM modifier reviews each page in sequence to detect non-textual elements missed by the automated extractor—such as tables whose borders were not detected, figures embedded within text columns, or cross-page elements split across bounding boxes. For each detected discrepancy, the modifier updates the chunk set and reassigns indices to maintain consistent ordering throughout the document. This stage substantially reduces errors of omission that would otherwise silently degrade the Sinkhorn co-association graph by severing links between text and their referents.

Stage 3: Human validation. Human annotators review the output of Stage 2, checking chunk boundaries, bounding-box accuracy, and correct pairing of captions with their referent chunks. Annotators correct residual extraction errors and resolve layout ambiguities that the LLM modifier did not catch. The resulting chunks serve as the human-annotated condition used in all main experiments and the noise ablation study in Section 4.3.

Dataset-specific variation. The degree of human involvement differs substantially across the three datasets by document homogeneity and scale.

MMLongBench-MQ contains heterogeneous documents spanning diverse layouts, fonts, and visual styles, making automated extraction unreliable for non-textual elements. Human annotators effectively took over the annotation of figures and tables in full, manually drawing bounding boxes, verifying caption associations, and resolving OCR failures at the element level. The LLM modifier served primarily as a pre-filter to surface candidate regions for human review rather than as a corrector in its own right.

FinanceBench consists of corporate financial filings, which are more typographically homogeneous and use standardized table formats. However, the corpus spans a large number of documents that exceed the capacity for detailed per-element human review. We therefore adopt a coarser strategy: an LLM is deployed to classify each page as containing or not containing tabular or visual content, and pages identified as multimodal are saved as full-page image chunks. This approach intentionally introduces some redundancy between the image chunk and the extracted text for the same page, but ensures that no tabular evidence is lost due to bounding-box failures. Human annotators performed minimal sanity checks, focusing on a held-out validation subset rather than the full corpus.

Chip Dataset documents—microprocessor specifications and architecture manuals—are homogeneous in layout, and the hardcoded PyMuPDF pipeline reliably identifies tables and figures with few errors of omission. Human annotators were deployed to validate the multimodal chunks produced by Stage 1, performing bounding-box corrections where necessary but without needing to introduce new elements from scratch. The overall annotation burden was substantially lower than for MMLongBench-MQ.

Depending on the dataset, the human annotators are given multimodal pages that they need to crop the tables/figures, or identify if a page contains multimodal chunks. The instruction is as below:

> *For identification:* `The presented is a page extracted from a technical document. Read it and answer the following question:`
> `Does it contain tables or figures? Answer either table, figure, both or none.`

For annotation: `The presented is a page extracted from a technical document with tables or figures. Double check if it contains such components. If so, crop the figure/table and upload it below. If not, leave it blank.`
`[Upload button]`

B.2 Additional Experimental Setups

Shared settings. All experiments use Voyage AI `voyage-multimodal-3` as the multimodal encoder. We retrieve $k = 5$ chunks for the main comparison; Section 4.3 reports the sensitivity analysis over k. Additional analysis of the embedding representation is provided in Appendix C.5.

SURGE (Ours). The Sinkhorn OT graph is constructed with entropic regularization coefficient $\lambda = 0.05$ and run for $T = 1{,}000$ Sinkhorn iterations until numerical convergence. The resulting plan is sparsified by discarding edges with weight below 0.05, with a minimum of 2 edges retained per node to prevent dangling nodes. When a text chunk explicitly references a non-textual chunk via patterns such as `Table~X`, `Fig.~X`, or `tab.~X`, the corresponding pairwise similarity is set to $\tilde{S}_{ij} = 1$ before computing the cost matrix. The Personalized PageRank step uses teleportation probability $\alpha = 0.70$, which keeps scores anchored to query relevance while permitting structural propagation. Its convergence tolerance is $\delta = 10^{-6}$, and it terminates in at most 10 power iterations across all documents in our evaluation.

Top-k RAG (Baseline). Chunks are ranked by cosine similarity between the query embedding and each chunk embedding; the top k are returned. No additional hyperparameters are introduced.

GraphRAG. A knowledge graph is constructed from document text by extracting up to 15 named entities per text chunk. At query time, up to 5 entities are extracted from the top-5 text chunks by cosine similarity to the query. Entity centrality in the knowledge graph is computed via standard PageRank with damping factor 0.85. Chunk scores are a weighted combination of three signals: query–chunk cosine similarity (50%), entity-matching score (30%), and entity-neighbor context score (20%). The entity-matching sub-score is itself a weighted combination of entity overlap with query entities (60%) and average entity centrality (40%); entity-neighbor context uses a 2-hop neighborhood. Since GraphRAG operates on text only, non-textual chunks are provided to it as enriched text descriptions.

DMAP. The original DMAP pipeline uses a ColPali image tower and a ColBERTv2 text tower, constructs a human-aligned hierarchical document map with a Structured-Semantic Understanding Agent, and uses a Reflective Reasoning Agent to assess evidence sufficiency and interact iteratively with that map. Our implementation preserves this process while using the document representations already available in our benchmark pipeline. It replaces ColBERTv2 with the cached Voyage `voyage-multimodal-3` text ranking and uses cached ColPali page retrieval as the image tower. Reciprocal-rank fusion ($k_{\mathrm{RRF}} = 60$) provides only the initial evidence seeds. For each document, we construct a hierarchy over inferred sections, pages, text chunks, figures, and tables. The map contains typed edges for adjacent pages, same-page multimodal elements, explicit figure/table references, and the two nearest text-semantic neighbours under Voyage embeddings. At query time, the evaluated backbone acts as the Reflective Reasoning Agent: it receives the current evidence, navigable map neighbours, and relevant images; judges evidence sufficiency; and may retain, replace, or add chunks. We permit at most two reflective iterations before generating the final answer from $k = 5$ selected chunks. This retains DMAP's map-construction–dual-retrieval–reflection–answering process, while substituting our parser, embedding representation, and constrained JSON reasoning prompt for the authors' components.

B.3 Context Assembly

After retrieval, the top-k chunks are assembled into a single prompt for the backbone VLM. Text chunks are concatenated under a labelled `Text References` section, and image chunks (figures and tables) are appended in their original modality as image inputs accompanied by a `Visual References` section that lists their filenames. Each text chunk receives an inline index (`[Text 1]`, `[Text 2]`, ...) so the model can refer back to specific evidence. Throughout all main experiments we use the prompt below, which deliberately omits any instruction permitting the model to refuse to answer; we found that allowing an explicit "impossible to answer" option

Component	Phase	Measured time
ColPali page encoding	Offline	18.88 s/document
DMAP structural-map construction	Offline	0.06 s/document
ColPali late-interaction retrieval	Query time	0.029 s/query
DMAP reflective reasoning (GPT-4o)	Query time	3.01 s/query
DMAP reflective reasoning (GPT-5.2)	Query time	4.93 s/query
DMAP reflective reasoning (Gemini Flash/Pro)	Query time	13.00 / 25.89 s/query
DMAP reflective reasoning (Qwen 8B/32B)	Query time	39.14 / 58.97 s/query

Table 6: Stage-separated computational profile. ColPali page encoding and retrieval were measured on FinanceBench using a GPU; DMAP map construction loads cached Voyage representations, while reflective latency is averaged over all three datasets and excludes deliberate rate-limit sleeps and the final answer call. DMAP makes one or two sequential reflective calls per query.

Method	Preprocessing/document	Retrieval/query
SURGE	12.32 s	163.6 ms
ColPali	18.88 s	29.3 ms
DMAP	31.15 s	196.1 ms

Table 7: Observed embedding-inclusive mean latency on FinanceBench under our deployment. Document preprocessing includes representation encoding and method-specific index construction; query latency includes query encoding and retrieval. DMAP's two representation towers are counted sequentially. Reflective-agent and final-answer VLM calls are excluded.

caused models to abstain on a non-trivial fraction of answerable queries, biasing the comparison between retrieval methods.

```
Based on the following context from a technical document,
answer the question.  Note that some documents may include
similar contents so pay attention to details.
Carefully examine the question and provide a precise answer
based on the given context.
Context:
{context}
Question:  {query}
Answer:
```

The {context} placeholder is replaced with the assembled context as described above; the {query} placeholder is the user's question verbatim. For backbone models that natively support multi-image inputs (GPT-4o, GPT-5.2, Gemini 2.5 Pro/Flash, Qwen3-VL), the figure and table chunks are passed as separate image attachments alongside the text prompt; for text-only baselines (GraphRAG), the same chunks are first verbalized by an LLM as enriched textual descriptions and inserted in the `Text References` block.

C Additional Experimental Results

C.1 Additional Information on Latency Computation

Table 7 complements the stage-separated comparison in Table 2 with the wall-clock cost observed when representation encoding is included. These measurements characterize the evaluated deployments, not a hardware-normalized comparison of the retrieval algorithms: SURGE and DMAP obtain Voyage embeddings through a managed remote API, whereas ColPali's ColQwen2 encoder runs locally on a GPU. Under this setup, ColPali has the lowest query latency (29.3 ms), while SURGE has the lowest document preprocessing cost (12.32 s). DMAP uses both representation towers, giving sequential costs of 31.15 s per document and 196.1 ms per query before reflective reasoning. If its independent image and text towers are executed concurrently, the estimated critical-path costs fall to approximately 18.91 s per document and 166.8 ms per query. All query measurements exclude

final answer generation, and the DMAP entry additionally excludes its one or two reflective-agent VLM calls so that the table isolates embedding and retrieval work.

C.2 Full Backbone-Wise Answer Accuracy

Methods	MMLongBench-MQ	FinanceBench	Chip Dataset	Average
Top-k RAG				
GPT-5.2	48.1	58.0	52.8	53.0
GPT-4o	44.9	52.0	43.1	46.7
Gemini 2.5 Pro	54.1	52.7	47.2	51.3
Gemini 2.5 Flash	48.1	50.0	43.1	47.1
Qwen3-VL-32B	50.3	52.0	37.5	46.6
Qwen3-VL-8B	40.0	54.0	31.9	42.0
GraphRAG				
GPT-5.2	34.6	56.7	50.0	47.1
GPT-4o	27.6	50.7	34.7	37.7
Gemini 2.5 Pro	33.5	50.7	38.9	41.0
Gemini 2.5 Flash	29.7	48.0	43.1	40.3
Qwen3-VL-32B	28.6	54.0	36.1	39.6
Qwen3-VL-8B	22.2	54.7	34.7	37.2
ColPali				
GPT-5.2	24.2	38.0	40.0	34.1
GPT-4o	27.5	44.7	45.6	39.3
Gemini 2.5 Pro	39.0	43.3	38.9	40.4
Gemini 2.5 Flash	31.9	43.3	32.2	35.8
Qwen3-VL-32B	32.4	32.0	36.7	33.7
Qwen3-VL-8B	20.3	42.0	33.3	31.9
DMAP				
GPT-5.2	28.7	56.0	40.0	41.6
GPT-4o	27.6	58.7	41.1	42.5
Gemini 2.5 Pro	39.2	60.7	45.6	48.5
Gemini 2.5 Flash	35.4	54.7	45.6	45.2
Qwen3-VL-32B	29.8	63.3	37.8	43.6
Qwen3-VL-8B	25.4	58.0	26.7	36.7
Ours (SURGE)				
GPT-5.2	50.3	**64.0**	**56.9**	**57.1**
GPT-4o	46.5	52.7	45.8	48.3
Gemini 2.5 Pro	**56.2**	53.3	52.8	54.1
Gemini 2.5 Flash	51.4	54.0	48.6	51.3
Qwen3-VL-32B	53.3	53.3	40.3	49.0
Qwen3-VL-8B	42.2	54.0	30.6	42.3

Table 8: Full answer accuracy (%) for every retrieval method and backbone VLM. FinanceBench uses the human-annotated `financebench-manual` split. Average is the arithmetic mean over the three datasets. The global best is marked **bold**, and the best result within each backbone is underlined.

Table 8 disaggregates the best-backbone summary in Table 4. SURGE attains the highest accuracy in 12 of the 18 dataset–backbone combinations, including at least one dataset for every evaluated backbone, and remains strongest overall despite less consistent gains with Qwen3-VL-8B. The table

also shows the dataset-specific strength of DMAP on FinanceBench, where it leads for five of the six backbones, while its results on MMLongBench-MQ and the Chip Dataset remain below the strongest non-agentic methods. This breakdown identifies which backbone produces each headline result and shows how consistently each retrieval method transfers across model families.

C.3 Qualitative Analysis of Alternative Evidence Paths

Conventional retrieval metrics compare retrieved chunks against a fixed annotated evidence set. This is necessarily incomplete in long documents, where the same fact may appear in a table, its surrounding prose, a nearby specification summary, or a cross-referenced section. We refer to these semantically equivalent routes to an answer as *alternative evidence paths*. A retrieval result can therefore have zero overlap with the annotated chunks while still providing sufficient evidence for a correct answer.

```
Chip dataset · entry 6 prompt 8

Question: For the MIPI DSI Host interface, what is the typical High-Speed
     transmit differential voltage (|VOD|) according to the MIPI D-PHY
    characteristics table? Additionally, what is the maximum DSI physical
     link speed per lane?

Ground truth
 - Answer: Typical |VOD|: 200 mV. Maximum DSI speed per lane: 1 Gbps.
 - Critical chunks: Table_51 chunk 1 & 2 (MIPI D-PHY characteristics
    table)

## SURGE retrieval, k=5

* Chunk image#51 --- Content: LCD timing diagram
* Chunk image#50 --- Content: USART timing diagram
* Chunk text#65 --- Content: "Pinout, pin description and alternate
    functions ..."
* Chunk text#84 --- Content: "STM32H757xI Pinout, pin description and
    alternate functions ... ."
* Chunk text#144 --- Content: "STM32H757xI Electrical characteristics 228
     VCMTX HS transmit common mode voltage"

**Recall@5 vs annotated GT (Table_51): 0 / 1**

SURGE's answer
- **Typical High-Speed transmit differential voltage (|VOD|):** **200 mV
    **
  (from the MIPI D-PHY characteristics table).
- **Maximum DSI physical link speed per lane:** **1 Gbps per lane**
  (from the DSI Host interface description).

Correctness: 1. Answer matches gold on both numeric values (200 mV, 1
    Gbps).

Note: text#144 is on page 145, adjacent to the GT Table_51 (pages 143--
    144), and is the surrounding MIPI D-PHY electrical-characteristics
    prose. It carries the |VOD| number and DSI speed info even though it
    is not the annotated table image itself.
```

This example is a strict false negative for evidence-overlap evaluation: the annotated path is missed, but an adjacent textual path supports the complete gold answer. It illustrates why similar MRR, recall, or nDCG values need not imply similar downstream accuracy, and why answer-based evaluation is essential alongside retrieval metrics for long multimodal documents. More generally, benchmarks with a single evidence annotation can conflate failure to retrieve one designated chunk with failure to retrieve sufficient evidence; exhaustively annotating alternative paths would make the retrieval evaluation more faithful but is substantially more costly.

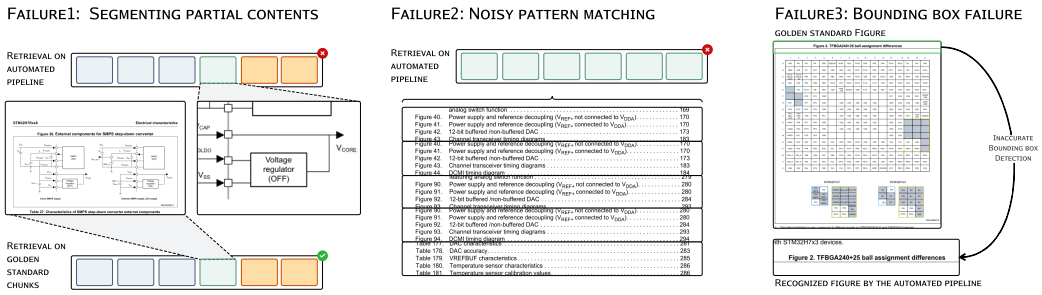


Figure 4: Three representative failure modes of the automated chunking pipeline. **Failure 1** (left): partial content segmentation splits a multi-element chunk, leaving the retrieved unit incomplete. **Failure 2** (centre): noisy pattern matching populates a chunk with a large number of irrelevant table-of-contents entries. **Failure 3** (right): an inaccurate bounding box crops only a fragment of the target figure, discarding the informative content.

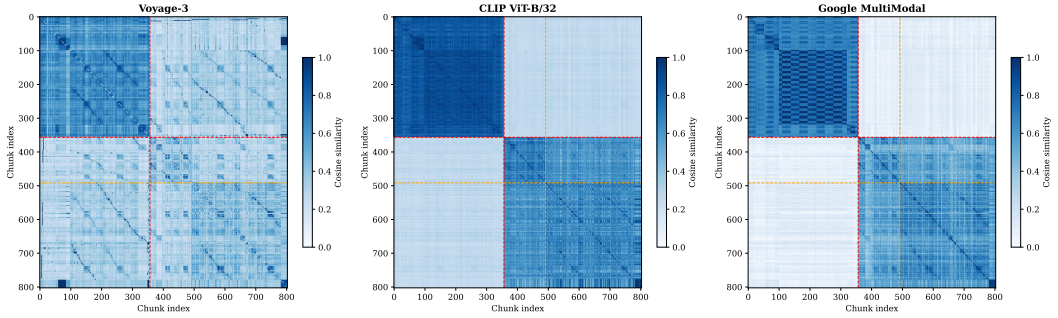


Figure 5: Pairwise cosine similarity maps for the same set of 803 chunks (Chip Dataset Document 1) under three multimodal embedders. Red and orange dashed lines mark the text/figure and figure/table boundaries respectively. Cross-modal quadrants vary substantially across embedders, exposing different magnitudes of the modality gap.

C.4 CHUNKING NOISE AND RETRIEVAL PERFORMANCE

The noise ablation in Section 4.4 shows that automated chunking depresses retrieval accuracy across all methods. Figure 4 illustrates three concrete failure modes that explain this degradation.

Failure 1: Segmenting partial contents. A document element that spans multiple layout regions—such as a figure accompanied by a caption block, or a table whose header sits in a separate bounding box from its body—may be split by the automated pipeline into independent chunks. When retrieval is performed on the resulting partial chunk, the retrieved unit contains only a fragment of the original element: a figure without its caption, or a table body without column headers. The missing portion may carry exactly the information required to answer the query, and because it has been separated into a sibling chunk that scores lower under both cosine similarity and PPR, it is systematically absent from the top-k retrieved set. Under human-annotated chunking, the element is kept intact and retrieved as a single unit, making the evidence accessible.

Failure 2: Noisy pattern matching. Automated pipelines commonly rely on layout heuristics or regular-expression patterns to identify tables and figures—for instance, detecting text lines that match the format of a table-of-contents entry or a figure index. When these patterns fire on incidental matches, they aggregate a large number of semantically unrelated entries into a single chunk. The resulting chunk embeds as a diffuse, low-information vector whose cosine similarity to any focused query is poor, and whose presence in the OT co-association graph introduces spurious edges that dilute PPR propagation. In the example shown, the automated pipeline captures an entire table-of-contents listing as a single figure chunk, rendering it non-retrievable for any specific factual query despite containing numerous relevant figure labels.

Block	Voyage-3	CLIP ViT-B/32	Google MM
text–text	**0.627**	0.863	0.757
figure–figure	**0.420**	0.701	0.580
table–table	**0.430**	0.662	0.622
figure–table	**0.320**	0.606	0.522
text–figure	**0.304**	0.262	0.082
text–table	**0.344**	0.262	0.102

Table 9: Mean cosine similarity within and across modality blocks for the three embedders. Text-to-visual cross-modal entries (rows 4–5) reveal the modality gap. The relations in the upper half is the lower the better as they imply content-agnostic intra-chunk similarity, and for the lower half vice versa. The figure–table relationship is put to the upper half because figure and table in our method shares the same encoding modality (image).

Document	Baseline	Ours (SURGE)
Document 1	41.7	50.0
Document 2	25.0	58.3
Document 3	58.3	**16.7**
Document 4	25.0	58.3
Document 5	83.3	66.7
Document 6	83.3	91.7

Table 10: Per-document accuracy (%) on Chip Dataset (human-annotated, $k = 5$, GPT-5.2). Document 3 is the sole case where SURGE substantially underperforms Baseline.

Failure 3: Bounding box failure. Even when a table or figure is correctly identified as a non-textual element, the bounding box detected by the layout parser may crop only a portion of the element—clipping columns from a wide table, cutting off axis labels from a plot, or excluding the legend from a diagram. The cropped image chunk embeds into a region of the embedding space that does not accurately represent the full element, weakening both its direct query similarity and its co-association weights with surrounding text chunks. In the example shown, the automated pipeline captures only the lower-right corner of a ball-assignment table, discarding the column and row headers that carry the structured information the query targets.

The role of human annotation. All three failure modes share a common property: they are immediately apparent to a human reviewer but invisible to an automated quality metric. Partial content splits are detectable by visual inspection of adjacent chunks; noisy pattern matches are identifiable as clearly misclassified elements; bounding box failures are visible as truncated or misaligned crops. Human annotators and validators can detect and correct each of these errors directly—merging split elements, discarding spurious chunks, and adjusting bounding boxes—at a cost that is feasible for curating controlled evaluation benchmarks. This motivates the human-in-the-loop pipeline described in Appendix B.1 and underscores why automated chunking alone is insufficient as a foundation for rigorous multimodal retrieval evaluation.

C.5 EMBEDDING MODELS AND THE MODALITY GAP

To examine how the choice of embedding backbone shapes the inter-chunk similarity structure, we compute pairwise cosine similarity matrices on Document 1 of the Chip Dataset ($n_{\text{text}} = 357$, $n_{\text{figs}} = 135$, $n_{\text{tabs}} = 311$, total 803 chunks) under three multimodal embedders: Voyage Multimodal-3 (the embedder used in our main experiments), CLIP ViT-B/32 (Radford et al., 2021), and Google's `multimodalembedding@001` (Vertex AI). Figure 5 visualizes the resulting maps; Table 9 summarizes the mean similarity within each modality block and across each pair of modalities.

Observations. The three embedders exhibit qualitatively different modality structures. **CLIP** produces high within-modality similarity (text–text 0.86, figure–figure 0.70) but suppresses cross-modal text-to-visual similarity to 0.26, reflecting its contrastive objective which optimizes within-modality clustering and aligns text only to its paired image at training time. **Google** exhibits the

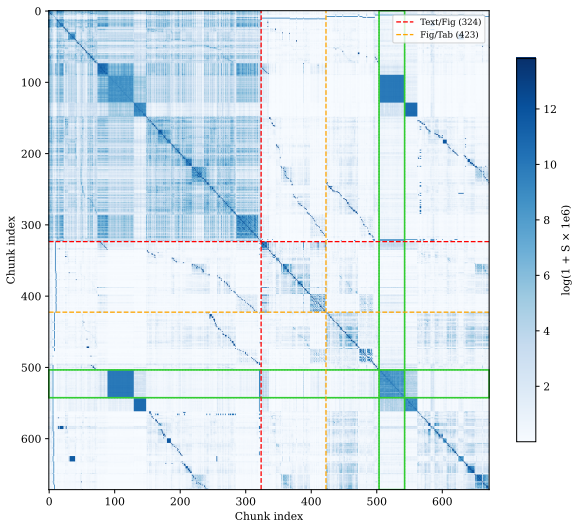


Figure 6: Sinkhorn transport map for Document 3 (log-scaled). The green band marks indices 504–542, one of two dense table clusters that accumulate disproportionate off-diagonal transport mass. The red and orange bonds marks the border of modalities.

most pronounced modality gap, with text–figure and text–table similarities collapsing to 0.08 and 0.10 respectively—visible as nearly white cross-modal quadrants in Figure 5—despite competitive within-modality similarities. **Voyage-3** has the smallest modality gap of the three: text–table cross-modal similarity (0.344) is in the same range as fig–fig within-modality similarity (0.420), making it the most modality-agnostic of the embedders we tested.

C.6 Failure Studies: Overweighted Table Clusters in the Sinkhorn Graph

We provide a case study of a failure mode observed in the Chip Dataset, where SURGE underperforms the Baseline by a large margin. Table 10 reports the accuracy of Baseline (Top-k) and SURGE broken down by individual document on the Chip Dataset (human-annotated variant, $k = 5$, GPT-5.2 backbone). While SURGE achieves consistent gains on most documents, it degrades severely on Document 3, dropping from 58.3% (Baseline) to 16.7% (SURGE).

Inspecting the Sinkhorn transport map of Document 3 (672 chunks: 324 text, 99 figures, 249 tables) reveals two densely connected table clusters at absolute indices 503–542 and 543–582 (Figure 6). These clusters correspond to back-to-back groups of timing and register tables in the datasheet. Together they account for $\mathbf{33.9}\%$ of all off-diagonal transport mass in the table block, which dominates the co-association structure and creates a strong attractor for PPR propagation.

Since Personalized PageRank propagates relevance along edges weighted by Sinkhorn co-association, nodes with high aggregate off-diagonal mass function as *probability sinks*: regardless of which text chunks are seeded by the initial query embedding, PPR probability flows preferentially into the two dense clusters. The result is that almost every query in Document 3 retrieves chunks from the same two table groups, which forms a query-independent retrieval distribution that is appropriate for at most a small fraction of the questions. By contrast, Baseline selects the top-k chunks by direct cosine similarity to the query embedding, which remains query-specific and correctly surfaces the relevant text or figure chunks for most questions.