BayesRAG: Probabilistic Mutual Evidence Corroboration for Multimodal Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has become a pivotal paradigm for Large Language Models (LLMs), yet current approaches struggle with visually rich documents by treating text and images as isolated retrieval targets. Existing methods relying solely on cosine similarity often fail to capture the semantic reinforcement provided by cross-modal alignment and layout-induced coherence. To address these limitations, we propose **BayesRAG**, a novel multimodal retrieval framework grounded in Bayesian inference and Dempster-Shafer evidence theory. Unlike traditional approaches that rank candidates strictly by similarity, BayesRAG models the intrinsic consistency of retrieved candidates across modalities as probabilistic evidence to refine retrieval confidence. Specifically, our method computes the posterior association probability for combinations of multimodal retrieval results, prioritizing text-image pairs that mutually corroborate each other in terms of both semantics and layout. Extensive experiments demonstrate that BayesRAG significantly outperforms state-of-the-art (SOTA) methods on challenging multimodal benchmarks. This study establishes a new paradigm for multimodal retrieval fusion that effectively resolves the isolation of heterogeneous modalities through an evidence fusion mechanism and enhances the robustness of retrieval outcomes. Anonymous code is available at https://anonymous.4open.science/r/BayesRAG-4C6B.

1 Introduction

RAG has fundamentally enhanced the generalization capabilities of LLMs by grounding them in external domain knowledge. However, real-world knowledge, ranging from academic papers to industrial reports, is predominantly presented in visually rich documents. In these documents, visual elements such as charts and diagrams are not merely decorative add-ons, but are intrinsically intertwined with textual explanations to convey complex semantics. Consequently, relying solely on unimodal text retrieval is insufficient. Verifying the semantic consistency between visual and textual evidence is paramount for building robust RAG systems.

In practice, RAG systems navigate vast repositories containing thousands of pages. While the advent of vision-language models has enabled multimodal retrieval, current SOTA approaches primarily focus on improving recall by expanding the Top-k window and performing a naive union of text and image retrieval results. We argue that this "bag-of-evidence" approach is fundamentally flawed. By treating modalities as independent channels and simply concatenating their representations, existing methods fail to capture the critical *semantic interaction* between modalities. Specifically, they cannot distinguish whether a retrieved image and a text chunk are logically corroborated or merely statistically similar (e.g., sharing common keywords but divergent contexts). This inability leads to a collection of high-similarity yet semantically disconnected fragments, which introduces noise and confounds the downstream answer generator.

To address modal disparity, some approaches employ Image-to-Text transduction or Knowledge Graphs (KGs) to unify information into a textual modality. However, these methods suffer from severe information and structural loss. Critical insights in knowledge-intensive domains are often encoded in non-textual formats that are "ineffable", which is difficult to translate losslessly into descriptions (Lumer et al., 2025). Furthermore, these linearization processes remove document layout information (Xiang et al., 2025). We consider that document layout is not merely a stylistic choice but serves as an implicit logical navigation map designed by the author. Standard chunking strategies disrupt this topological structure, severing the logical ties between spatially adjacent evidence.

In summary, developing a trustworthy multi-

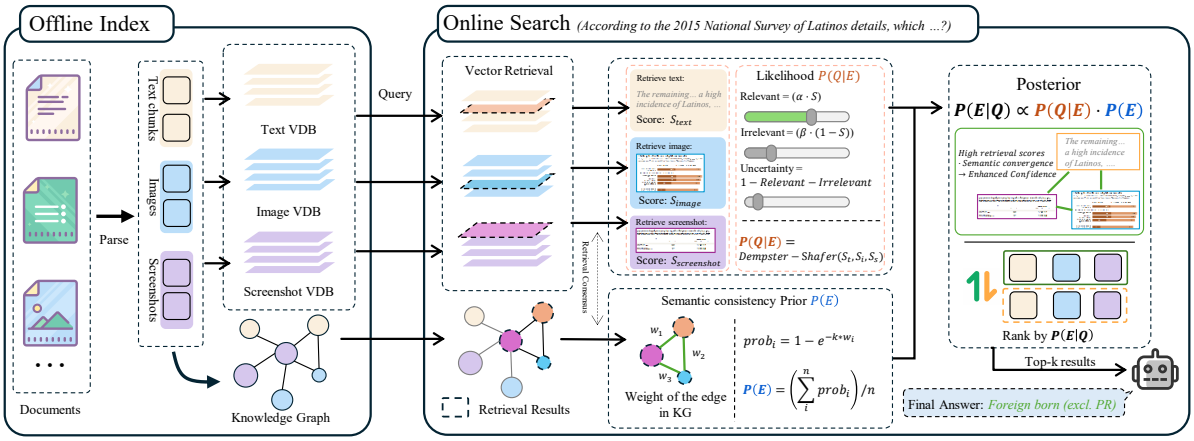

Figure 1: The architecture of BayesRAG. We reconceptualize multimodal retrieval as a Bayesian inference process. Our method computes the semantic likelihood $P(Q|E)$ and derives the consistency prior $P(E)$ via knowledge graph topology. The final evidence tuples are re-ranked based on posterior probability $P(E|Q)$, ensuring semantic relevance, mutual corroboration, and logical self-consistency.

modal RAG system necessitates addressing three pivotal challenges: (1) **Preserving Multimodal Fidelity:** The system should move beyond lossy translation to perform native representation learning (Mei et al., 2025). (2) **Cross-Modal Semantic Corroboration:** Achieving high confidence requires a mechanism for evidence verification. The system needs to facilitate interaction between retrieved modalities, identifying instances where visual and textual candidates semantically align. Such alignment should be treated as reinforcing evidence, thereby assigning higher posterior confidence to mutually corroborating retrieval results. (3) **Layout-Induced Structural Understanding:** Beyond content semantics, the system should be able to recover the contextual relationships of retrieved elements within the original document, which allows for decoding the implicit semantic topology within the document layout (Tong et al., 2025), thereby restoring connections that are frequently severed by traditional chunking strategies.

To address these challenges, we propose **BayesRAG**. Diverging from traditional approaches that rely solely on unimodal similarity matching, we reconceptualize multimodal retrieval as a process of *probabilistic evidence fusion*. Our core intuition addresses the prevalent issue of semantic inconsistency in heterogeneous retrieval; this is exemplified by scenarios where text retrieval suggests the fruit "apple" while visual retrieval identifies the corporate logo of "Apple Inc." Under such circumstances, independent similarity scores remain high for both yet fail to capture the fatal discordance between them. BayesRAG leverages Bayesian inference to elegantly resolve this dilemma by utilizing similarity scores from embedding models as the *likelihood* and modeling cross-modal semantic consistency as the *Prior Belief*. This formulation mathematically embodies the intuition that *"only text-image pairs that are semantically consistent deserve higher trust,"* and it thereby dynamically penalizes conflicting evidence during the fusion.

By computing the posterior probability, BayesRAG automatically down-weights retrieval candidates that exhibit high unimodal scores yet contradict each other. This mechanism effectively isolates high-fidelity evidence combinations that are mutually corroborating, simultaneously preserving native multimodal information and ensuring the logical rigor of cross-modal reasoning. Our main contributions are summarized as follows:

- We establish a unified probabilistic framework that reconceptualizes multimodal retrieval as an evidence fusion process; this paradigm effectively preserves the native layout structure often lost in traditional strategies.
- We introduce a Bayesian inference mechanism grounded in Dempster-Shafer theory to model semantic consistency; this approach resolves cross-modal conflicts by prioritizing mutually corroborating evidence.
- Extensive experiments demonstrate that BayesRAG achieves SOTA performance on challenging benchmarks; our analysis confirms that modeling evidence consistency effectively eliminates modal conflicts and semantic inconsistencies prevalent in existing baselines.

2 Method

In this section, we present **BayesRAG**, a probabilistic framework for multimodal retrieval.

2.1 Retrieval as Evidence Fusion

Traditional multimodal RAG systems typically treat retrieval as independent feature-matching tasks, where text and images are ranked solely by their semantic similarity to the query. This often leads to modal conflict, where a high-scoring text and a high-scoring image may contradict each other or originate from unrelated contexts.

BayesRAG redefines retrieval as a **Probabilistic Evidence Fusion** process. We aim to identify an evidence tuple $E = (e_{txt}, e_{vision}, e_{screenshot})$, consisting of evidence from text, image, and structural information, that is not only relevant to the query but also *internally consistent.* Our intuition is modeled via Bayesian inference: **Prior** (Internal Consistency): Before observing the query, how likely is it that this text and image belong together? This is determined by the inherent consistency of the tuple E, exemplified by factors such as semantic alignment or narrative proximity in the source document. **Likelihood** (Semantic Match): Given this evidence tuple, how likely is it to generate the user's query? We model this using vector similarity enhanced by belief functions. **Posterior** (Final Confidence): The updated probability that the tuple is the optimal evidence given the query.

2.2 Problem Formulation

Let $\mathcal{D}$ be a multimodal document corpus. We define the sample space Θ as the set of all possible multimodal evidence tuples. A single evidence candidate is denoted as $E = (t, v, s) \in \Theta$, where t represents a textual segment, v represents a visual component, and s denotes the structural information from screenshots.

Let Q denote the observed user query. Our goal is to compute the posterior probability $P(E|Q)$, which represents the confidence that the tuple E is the optimal evidence given the query Q. According to Bayes' Theorem:

$$P(E|Q) = \frac{P(Q|E) \cdot P(E)}{P(Q)} \tag{1}$$

Since $P(Q)$ (the marginal likelihood of the query) is constant for all candidates during the ranking phase, we simplify the objective function to:

$$P(E|Q) \propto \underbrace{P(Q|E)}_{\text{Embedding Likelihood}} \cdot \underbrace{P(E)}_{\text{Consistency Prior}} \tag{2}$$

2.3 Bayesian Evidence Modeling

Modeling Likelihood with Belief Functions. The likelihood $P(Q|E)$ estimates the probability that the query Q is derived from the evidence E. To robustly handle the uncertainty in embedding cosine similarity scores (e.g., distinguishing between "irrelevant" and "uncertain"), we incorporate the Dempster-Shafer Theory of Evidence.

For each modality $i \in \{t, v, s\}$, let $S_i \in [0, 1]$ be the normalized cosine similarity score with the query. We construct a mass function m_i over the frame of discernment $\Theta = \{\text{Y}, \text{N}\}$ (Relevant, Irrelevant), as shown in Table 1.

Hypothesis	Mass Function Formulation
Evidence (e)	$m(\text{Y}) = \alpha \cdot S_i$
Not Evidence ($\bar{e}$)	$m(\text{N}) = \beta \cdot (1 - S_i)$
Uncertainty (Ω)	$m(\Omega) = 1 - m(\text{Y}) - m(\text{N})$

Table 1: Mass function definitions based on cosine similarity scores. $\alpha, \beta \in [0, 1]$ are hyperparameters representing trust in the embedding model.

We aggregate the mass functions from text, vision, and screenshot using Dempster's Rule of Combination ($m_{joint} = m_i \oplus m_j$). The likelihood is then projected from the combined belief function:

$$P(Q|E) \approx BetP_{joint}(\text{Y}) \tag{3}$$

This formulation ensures that the likelihood is high only when multiple modalities provide conflict-free evidence supporting the query (see Appendix A.1).

Modeling the Consistency Prior $P(E)$. The prior $P(E)$ represents the probability that the tuple $E = (t, v, s)$ constitutes a meaningful, coherent unit of information *independent* of the user's query. This captures the "topological" or "physical" reality of the document structure. If the text t and image v are semantically unrelated or spatially distant in the source document, $P(E)$ should be low. We model this as:

$$P(E) = f_{\text{consistency}}(t, v, s) \tag{4}$$

where $f_{\text{consistency}}$ is modeled by the semantic consistency or spatial distance between the text, image, and screenshot.

To quantify the intrinsic coherence $P(E)$ of an evidence tuple $E = (t, v, s)$, we propose two modeling strategies:(1) **Graph-Topology Prior**, which constructs a multimodal knowledge graph from document chunks to model *semantic consistency*. This approach calculates connection strengths between retrieved nodes to prioritize tuples that form strongly connected paths. Specifically, we check the topological connectivity of the retrieved triplet (t, v, s) within the graph and model their mutual semantic consistency based on the weights of the edges connecting these three nodes. (2) **Layout Prior**, which leverages bounding box coordinates to model *geometric proximity*, based on the assumption that spatially adjacent elements on the same or nearby pages have a higher probability of association. Both strategies map physical or topological signals into a probability space $P(E) \in [0, 1]$ (details in Appendix A.2). As shown in the "Spatial vs. Semantic Priors" analysis in Section 3.3, incorporating either prior significantly improves QA performance compared to standard embedding-only retrieval. Notably, the semantic constraints offered by the Graph-Topology Prior outperform the purely geometric constraints of the Layout Prior.

3 Experiments

3.1 Experimental Settings

Evaluation Benchmarks. We evaluate BayesRAG on two challenging multimodal Document QA benchmarks: DocBench (Zou et al., 2025) and MMLongBench-Doc (Ma et al., 2024). DocBench comprises 229 documents across five domains (e.g., Finance, Law) with 1,102 QA pairs, featuring substantial length (avg. 66 pages, ∼46k tokens) to test long-context capabilities. MMLongBench-Doc complements this with 135 documents spanning 7 diverse types and 1,082 questions. Together, these datasets cover over 2,000 questions in varied real-world scenarios, providing a robust testbed for multimodal long-document understanding.

Baselines. To evaluate the effectiveness of BayesRAG, we compare it against a diverse set of SOTA multimodal RAG systems, covering modular, visual-centric, agentic, and graph-based paradigms: **RAGFlow**[1]: A widely adopted open-source RAG engine that integrates retrieval pipelines with agentic capabilities. It serves as a representative baseline for standard, modularized RAG implementations used in industry. **VisRAG** (Yu et al., 2025a): A visual-centric framework that bypasses traditional OCR parsing. Instead of extracting text, VisRAG treats document pages directly as images, utilizing VLMs for both visual embedding and generation. **ViDoRAG** (Wang et al., 2025b): An agent-based approach that employs a Gaussian Mixture Model (GMM) for multimodal fusion. It introduces an iterative agentic workflow (consisting of Explore, Summarize, and Reflect stages) to verify the reliability of retrieved evidence dynamically. **RAGAnything** (Guo et al., 2025): A knowledge-graph-based system that reconceptualizes multimodal content as interconnected entities. It employs a dual-graph construction mechanism and hybrid retrieval, combining structural navigation within the KG with semantic vector matching to locate cross-modal evidence. We also evaluate the performance of directly feeding PDFs into **GPT-4o-mini**. For documents with fewer than 50 pages, we use page-level screenshots as input. For longer documents, we randomly sample 50 screenshots as input.

[1] https://github.com/infiniflow/ragflow

Method	Domains					Overall
	Aca.	Fin.	Gov.	Law	News	
GPT-4o-mini	42.5	36.8	53.1	51.4	51.1	44.1
VisRAG	21.9	28.2	20.3	29.6	11.6	25.3
ViDoRAG	29.0	**69.1**	35.8	39.7	37.2	43.5
RAGFlow	45.2	34.0	41.8	48.1	23.8	39.0
RAGAnything	51.4	42.3	43.2	50.2	57.5	48.7
BayesRAG (ours)	**52.4**	37.5	**56.0**	**52.3**	**66.8**	**51.2**

Table 2: Accuracy (%) on DocBench. Best performance is highlighted in **bold** and second-best is underlined. Domain categories include Academia (Aca.), Finance (Fin.), Government (Gov.), Legal Documents (Law), and News Articles (News).

Evaluation Metrics. For **DocBench**, we strictly follow the official evaluation protocol, utilizing the standard LLM-based scoring mechanism to ensure fair comparison with existing baselines. For **MMLongBench-Doc**, in addition to reporting the official rule-based accuracy (Acc), we introduce a supplementary LLM-based evaluation. This is designed to mitigate false negatives inherent in rigid string matching, specifically addressing cases of semantic equivalence (e.g., abbreviations or paraphrasing) that rule-based metrics fail to capture.

Experimental Setup. To ensure a rigorous and standardized evaluation, we employ GPT-4o-mini as the final response generator across all experiments. We utilize MinerU (Niu et al., 2025) for

Method	Reports		Tutorials		Academic		Guidebooks		Brochures		Industry		Financial		Overall	
	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score
GPT-4o-mini	29.8	36.7	29.0	38.2	16.8	24.0	24.0	32.4	29.5	34.6	**41.3**	**50.6**	38.3	41.0	28.2	35.2
VisRAG	27.9	30.0	19.4	20.8	21.5	24.0	19.8	23.7	19.8	23.7	19.7	24.6	11.1	11.1	21.3	23.8
ViDoRAG	29.8	36.1	33.3	39.5	16.5	21.0	27.6	32.0	19.6	24.7	21.8	25.9	35.3	36.7	26.5	31.4
RAGFlow	36.8	38.9	31.5	35.2	27.8	29.9	30.6	35.2	33.4	36.6	28.7	30.8	31.2	34.1	32.0	34.9
RAGAnything	40.6	45.3	39.5	46.0	32.0	34.6	37.3	43.5	36.4	**39.6**	30.5	33.3	39.2	42.7	37.1	41.5
BayesRAG (ours)	**42.6**	**46.7**	**41.8**	**53.6**	**32.7**	**34.8**	**41.6**	**46.4**	**37.0**	39.4	31.1	38.6	**41.2**	**48.6**	**38.8**	**44.1**

Table 3: Accuracy (%) and GPT-Score (%) on MMLongBench-Doc across different domains and overall performance. Best performance is highlighted in **bold** and second-best is underlined. Domain categories include Research Reports/Introductions (Reports), Tutorials/Workshops (Tutorials), Academic Papers (Academic), Guidebooks (Guidebooks), Brochures (Brochures), Administration/Industry Files (Industry), and Financial Reports (Financial). Acc is calculated by the verification method in the MMLongBench-Doc paper. Score is the score verified by LLM.

Method	Recall@1	Recall@3	Recall@5	Recall@10	Recall@20
Vector Retrieval	24.4	41.7	46.4	51.6	56.6
BayesRAG	27.5	44.3	54.7	61.7	76.6

Table 4: Retrieval performance comparison on the MMLongBench-Doc dataset. **Vector Retrieval** denotes the baseline using raw similarity scores from embedding models and text reranker without probabilistic re-ranking. **BayesRAG** incorporates our proposed Bayesian evidence fusion.

document parsing. We employ Qwen3-Embedding-4B (Zhang et al., 2025) to encode textual chunks. For image elements, we utilize PE-Core-G14-448 (Bolya et al., 2025). Additionally, we use colnomic-embed-multimodal-3b[2] to generate page-level embeddings for global layout understanding. All retrieved text candidates are refined using bge-reranker-v2-m3 (Chen et al., 2024). For the construction of the Knowledge Graph, we leverage the Qwen3-VL-32B (Bai et al., 2025) to extract entities and relations from both text and images. We build all documents in the benchmark into one vector database/knowledge graph for retrieval and evaluation. Detailed hyperparameters and environment configurations are provided in Appendix B to facilitate reproducibility. Crucially, due to the absence of mature visual reranking models and the computational cost of VLM-as-rerankers, the baseline is constrained to a shallow pool of visual candidates (Top-20 or the adaptive top-k in its method). In contrast, BayesRAG's probabilistic fusion is computationally lightweight, allowing us to expand the initial visual pool to Top-512 without incurring significant latency (see Appendix F).

3.2 Main Results

Performance comparison. Table 2 further validates the versatility of our approach. BayesRAG achieves an SOTA overall score of **51.2%**, surpassing the strongest baseline RAGAnything by 2.5%. It demonstrates dominance in text-intensive and high-noise domains, such as *Government* (+12.8%) and *News* (+9.3%). This confirms that our Bayesian evidence fusion not only handles multimodal alignment but also enhances textual retrieval by penalizing irrelevant evidence through consistency priors. As shown in Table 3, BayesRAG secures the leading position in composite metrics, achieving **38.8%** accuracy and an LLM score of **44.1%**. Our framework exhibits exceptional robustness in heterogeneous domains, notably *Guidebooks* (+4.3% Acc over the runner-up) and *Tutorials/Workshops* (+7.6% Score). These documents typically feature intricate interplays between text, charts, and layout. By explicitly modeling evidence consistency, BayesRAG effectively filters out modal mismatches, whereas visual-centric methods like VisRAG suffer from noise introduced by full-page embedding (21.3% overall Acc).

Retrieval performance. We compare BayesRAG against the Vector Retrieval baseline. The baseline merely aggregates raw cosine scores from text, image, and screenshot embedding models with a text reranker, lacking mechanisms for semantic consistency checks or structural priors. Utilizing the page-level evidence annotations in MMLongBench-Doc, we approximate retrieval recall to evaluate performance. Table 4 demonstrates that BayesRAG achieves substantial gains on all Recall@K metrics: Precision Enhancement (R@1): Recall@1 improves significantly by 3.1%. This result confirms that our Bayesian formulation effectively mitigates high-similarity hallucinations. By penalizing conflicting evidence, where $P(Q|E)$ drops due to modal inconsistencies, the framework successfully prioritizes truly corroborating evidence. Recall Robustness (R@20): Crucially, the advantage be-

[2] https://huggingface.co/nomic-ai/colnomic-embed-multimodal-3b

Method	Bayesian Inference	Likelihood	Prior	Reports	Tutorials	Academic	Guidebooks	Brochures	Industry	Financial	Overall
Embedding-RAG	✗	-	-	39.2	27.9	26.4	31.3	32.6	25.3	19.8	32.7
BayesRAG (Linear Fusion)	✓	Linear Weighted	Graph	45.0	41.3	33.3	39.3	35.0	26.2	35.7	40.6
BayesRAG (Layout Prior)	✓	Dempster-Shafer	Layout	38.4	33.5	30.5	40.6	36.5	21.2	27.2	36.8
BayesRAG (Full)	✓	Dempster-Shafer	Graph	46.7	53.6	34.8	46.4	39.4	38.6	48.6	44.1

Table 5: Ablation study on Bayesian components. *Embedding-RAG*: Baseline with independent retrieval. *BayesRAG (Layout Prior)*: Incorporates spatial consistency constraints ($P(E)$) based on bounding box proximity (assigning 1.0 for adjacent elements, 0.1 otherwise). *BayesRAG (Linear Fusion)*: Estimates likelihood ($P(Q|E)$) via linear weighted averaging of normalized scores. *BayesRAG (Full)*: Synergizes Graph-Topology priors with Dempster-Shafer theory to robustly handle evidence conflict and uncertainty.

comes more pronounced as K increases, with Recall@20 surging by 20.0% (56.6% $\rightarrow$ 76.6%). This reveals that standard vector retrieval often buries ground-truth evidence in the long tail due to low semantic scores caused by modality misalignment or OCR noise. BayesRAG leverages consistency priors $P(E)$, to "rescue" these relevant but low-scoring candidates.

Attribution Analysis. While BayesRAG demonstrates strong retrieval capabilities, achieving a Recall@20 of 76.6%, the final generation performance (Score) stands at 44.1%. This discrepancy reveals a substantial *Retrieval-Generation Gap*. However, rather than a limitation, we interpret this as BayesRAG successfully unlocking the retrieval bottleneck that constrains previous systems. The high recall indicates that our framework provides a rich, high-fidelity context that current compact generators (like GPT-4o-mini) may struggle to fully digest, but which lays a solid foundation for more capable Multimodal LLMs. To investigate the current generation bottleneck, we conduct a manual analysis on a sample of 100 failure cases. We find that 51% of the errors stem from query ambiguity inherent in MMLongBench-Doc. Specifically, questions often appear in generic forms such as "What year is the report for?" In the RAG setting that involves retrieving information across multiple distinct documents, such queries lack the necessary context to disambiguate the target document. This observation elucidates the phenomenon reported in Table 2 and Table 3, where directly feeding 50 page-level screenshots of the specific target document into GPT-4o-mini yields superior performance.

3.3 Ablation Experiments

To assess the individual contributions of our structural priors and evidential fusion mechanism, we compare BayesRAG against three variants on MMLongBench-Doc: Embedding-RAG, which relies solely on embedding-based retrieval and a text reranker; a variant utilizing only Layout Priors; and a variant employing Linear Fusion.

Bridging the Visual Reranking Gap. To isolate the contribution of our Bayesian inference mechanism, we compare BayesRAG against Embedding-RAG. As shown in Table 5, the baseline employs a hybrid retrieval strategy: text candidates are refined using bge-reranker-v2-m3 (retrieving Top-1024 and selecting Top-20), while visual candidates (images and page screenshots) rely solely on raw embedding scores (Top-20) due to the absence of mature visual reranking models or the computational cost of VLM-as-rerankers. The results reveal a significant performance disparity. BayesRAG (Full) outperforms the baseline by a substantial margin of **11.4%** in overall score. The performance leap can be attributed to how BayesRAG handles visual noise. In the baseline, visual retrieval is limited to the Top-20 raw candidates. However, embedding models may assign high scores to visually similar but semantically irrelevant images. BayesRAG addresses this by expanding the initial visual pool to Top-512 and utilizing the posterior $P(E|Q)$ as a **soft reranker**. By enforcing consistency constraints, BayesRAG effectively filters out high-scoring visual hallucinations that dominate the baseline's Top-20 list, ensuring that the final retrieved context is semantically coherent.

Comparison of fusion strategy. We further investigate the fusion mechanism by comparing our Dempster-Shafer-based approach against a standard *Linear Fusion*, which calculates $P(Q|E)$ as a weighted average of normalized embedding scores. While Linear Fusion improves upon the baseline (40.6% score), it still underperforms the BayesRAG (Full) (44.1%) by a distinct margin. The deficiency of linear weighting lies in its inability to handle **modal conflicts**. In scenarios where text and image retrievers contradict each other, a linear sum blindly averages the scores, potentially retaining the noise. In contrast, Dempster-Shafer formulation explicitly calculates a *Conflict Coef-*

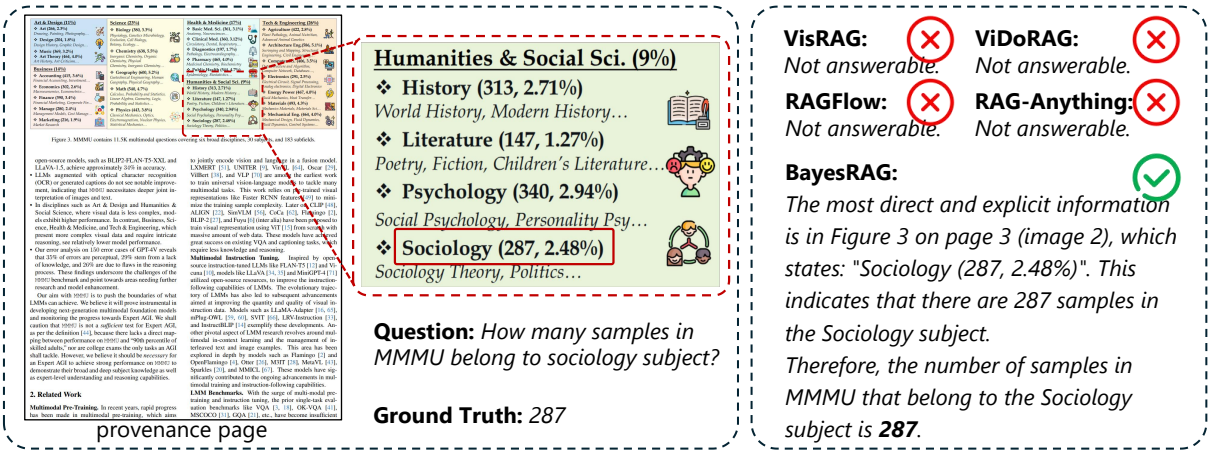

Figure 2: A qualitative comparison illustrating the efficacy of BayesRAG in broadening the scope of visual retrieval and filtering inconsistent evidence. Our statistical analysis reveals that approximately 34% of failure cases in baseline methods are attributed to this retrieval deficiency.

ficient K. When divergence between modalities is detected, the belief mass is redistributed to the uncertainty state or penalized, ensuring that only mutually corroborating evidence survives. This is particularly evident in the *Tutorials* domain (41.3% $\rightarrow$ 53.6%), where precise alignment between complex charts and descriptions is critical.

Spatial vs. Semantic Priors. The *Layout Prior* introduces a geometric constraint $P(E)$ derived from document layout analysis. It functions as a spatial filter, assigning a high prior probability (1.0) to evidence tuples where the visual and textual components are spatially adjacent (within a specific bounding box distance), and a penalty (0.1) otherwise. As shown in Table 5, incorporating this physical constraint alone boosts the overall score from 32.7% to 36.8%. This result validates our hypothesis that *spatial consistency* serves as a reliable proxy for *semantic consistency*. By strictly enforcing layout coherence, the model effectively reduces uncertainty, filtering out visually similar but structurally unrelated hallucinations. While bounding boxes provide a strong baseline, this motivates our further exploration into Knowledge Graphs (in the full model) to capture deeper semantic links beyond mere physical proximity.

3.4 Case Study

To explicitly demonstrate how BayesRAG handles complex retrieval challenges, we visualize two representative cases in Figure 2 and Figure 3, comparing our method against leading baselines.

Case 1: Figure 2 depicts target text ("Sociology") buried in a complex layout. Baselines fail to capture this low-salience signal, returning "Not answerable". BayesRAG addresses this by expanding the retrieval candidate pool (Top-k) and leveraging the semantic consistency between the query and the image text to filter the resulting noise effectively. This capability is critical, as our analysis reveals that about 34% of baseline errors stem from similar failures to retrieve ground-truth evidence, directly resulting in incorrect "Not answerable" predictions.

Case 2: Figure 3 presents a challenge where baselines are misled by the visual similarity of a "Mobile phone" distractor chart located on a different slide. BayesRAG successfully retrieves the correct evidence through its probabilistic scoring mechanism. Specifically, the explicit match between the query "Smartphone" and the textual component of the target evidence tuple yields a higher **likelihood** score. Simultaneously, the semantic consistency between the tuple's text and image results in a robust **prior** score. These factors synergize during the Bayesian update to rank the correct "Smartphone" tuple significantly higher than the visually similar distractor, enabling the accurate calculation.

In summary, BayesRAG not only expands recall for obscure details in images but also robustly filters confusing distractors through rigorous probabilistic verification.

4 Related Works

Multimodal Vector Retrieval Paradigms. RAG significantly enhances LLMs by integrating external knowledge bases (Lewis et al., 2020), with embedding models (Chen et al., 2024) playing a pivotal role in this process. Recently, vision-language models such as CLIP and ColPali (Faysse et al., 2025) have extended RAG capabilities to

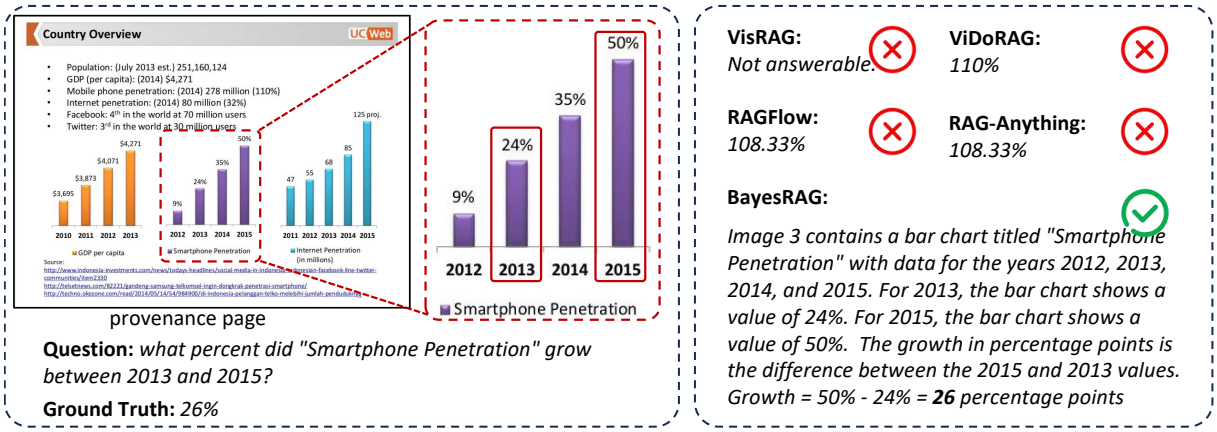

Figure 3: A sample is for semantic consistency between text and charts to find the correct evidence

retrieve images or document pages directly. Building on these advancements, RegionRAG (Li et al., 2025) introduces a region-level contrastive learning objective, thereby refining retrieval granularity from coarse-grained whole pages to specific semantic regions. Broadly, current approaches fall into two categories: (1) Visual-Centric Unification: Methods like VisRAG (Yu et al., 2025b) and M3DocRAG (Cho et al., 2024) unify modalities by treating entire document pages as images for retrieval, effectively preserving layout but potentially sacrificing fine-grained textual details. (2) Dual-Stream Retrieval: Systems like VisDoM (Suri et al., 2025) store and retrieve text and images separately. They typically employ a late fusion strategy, where separate visual and textual models generate intermediate answers that are subsequently merged by a fusion model. However, these separate streams often lack interaction during the retrieval phase, leading to potential semantic misalignment.

Knowledge-Structured Multimodal RAG. To capture complex relationships within documents, several approaches integrate Knowledge Graphs. Methods such as HM-RAG, MMGraphRAG (Wan and Yu, 2025), and mKG-RAG (Yuan et al., 2025) utilise Multimodal LLMs (MLLMs) to extract entities and relationships from both text and images. RAG-Anything (Guo et al., 2025) employs a hybrid search mechanism that combines explicit KG links with vector semantic similarity. It utilizes a multi-signal scoring mechanism that weighs both the topological importance of the graph structure and the semantic similarity scores. While effective, these graph-based methods rely heavily on the quality of upstream information extraction.

Evidence Verification and Reranking. Ensuring the reliability of retrieved content is critical. For text modality, cross-encoder rerankers are standard practice. MIRAGE (Wei et al., 2025) ensures trustworthy evidence by cross-validation of inference chains in enhanced retrieval-augmented graphs. For images, frameworks like ViDoRAG (Wang et al., 2025b), MDocAgent (Han et al., 2025), HKRAG (Tong et al., 2025), mRAG (Hu et al., 2025), and MADAM-RAG (Wang et al., 2025a) adopt an agent-based approach, utilizing agents such as the Inspector or Critical Agent to perform iterative verification loops to filter evidence. However, these agent-based methods are often computationally intensive due to repeated LLM calls, resulting in substantial token consumption and high latency. Furthermore, they lack a unified probabilistic framework to explicitly model the mutual corroboration between heterogeneous modalities—a gap we aim to fill with BayesRAG.

5 Conclusion

In this work, we introduce BayesRAG to advance multimodal RAG for visually rich documents by transforming retrieval into a dynamic evidence fusion process. By synergizing Bayesian inference with Dempster-Shafer theory, our framework offers a theoretically rigorous solution to the semantic inconsistencies prevalent in heterogeneous data. We demonstrate that explicitly modeling modal conflicts allows the system to filter out high-scoring but semantically misaligned noise. This capability effectively closes the gap between visual retrieval and semantic verification. Our extensive evaluations confirm that this probabilistic approach not only achieves state-of-the-art performance but also establishes a new paradigm for trustworthy multimodal RAG. We hope this shift towards evidence-based corroboration paves the way for future systems that demand high-fidelity reasoning.

Limitations

Despite the substantial advancements achieved, BayesRAG is currently subject to certain limitations that point towards future research directions. First, we observe a *Retrieval-Generation Gap*. While our Bayesian retriever achieves high effective recall (61.7% @ Top-10 and 76.6% @ Top-20 in Table 4), the downstream generation accuracy is bounded by the current reasoning capabilities of MLLMs when interpreting information-dense visual evidence. We anticipate that as generator models evolve, they will better utilize the high-fidelity evidence provided by our framework. Second, as our methodology is specialized for *multimodal* retrieval, its advantages are naturally less pronounced in domains dominated by pure text or massive continuous tables, such as certain financial reports. In such contexts, the standard chunking strategy often fragments tabular continuity. This remains a common challenge inherent to most RAG pipelines. BayesRAG prioritizes cross-modal consistency, which is most effective when documents contain rich interplay between texts and images.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. 2025. Perception encoder: The best visual embeddings are not at the output of the network. *Preprint*, arXiv:2504.13181.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025. Rag-anything: All-in-one rag framework. *Preprint*, arXiv:2510.12323.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *Preprint*, arXiv:2503.13964.

Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, Suofei Feng, Ryan Rossi, and Zhengzhong Tu. 2025. mrag: Elucidating the design space of multi-modal retrieval-augmented generation. *Preprint*, arXiv:2505.24073.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9459–9474.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. 2025. Regionrag: Region-level retrieval-augmented generation for visual document understanding. *Preprint*, arXiv:2510.27261.

Elias Lumer, Alex Cardenas, Matt Melich, Myles Mason, Sara Dieter, Vamse Kumar Subbiah, Pradeep Honaganahalli Basavaraju, and Roberto Hernandez. 2025. Comparison of text-based and image-based retrieval in multimodal retrieval augmented generation large language model systems. *Preprint*, arXiv:2511.16654.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems*, volume 37, pages 95963–96010. Curran Associates, Inc.

Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. 2025. A survey of multimodal retrieval-augmented generation. *Preprint*, arXiv:2504.08748.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6088–6109, Albuquerque, New Mexico. Association for Computational Linguistics.

Anyang Tong, Xiang Niu, ZhiPing Liu, Chang Tian, Yanyan Wei, Zenglin Shi, and Meng Wang. 2025. Hkrag: Holistic knowledge retrieval-augmented generation over visually-rich documents. *Preprint*, arXiv:2511.20227.

Xueyao Wan and Hang Yu. 2025. Mmgraphrag: Bridging vision and language with interpretable multimodal knowledge graphs. *Preprint*, arXiv:2507.20804.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025b. ViDoRAG: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9124–9145, Suzhou, China. Association for Computational Linguistics.

Kaiwen Wei, Rui Shan, Dongsheng Zou, Jianzhong Yang, Bi Zhao, Junnan Zhu, and Jiang Zhong. 2025. Mirage: Scaling test-time inference with parallel graph-retrieval-augmented reasoning chains. *Preprint*, arXiv:2508.18260.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation. *Preprint*, arXiv:2506.05690.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025a. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025b. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. 2025. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *Preprint*, arXiv:2508.05318.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2025. DocBench: A benchmark for evaluating LLM-based document reading systems. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 359–373, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

A Implementation of Bayesian inference

A.1 Likelihood Estimation via Dempster-Shafer Theory

In this appendix, we detail the calculation of the likelihood term $P(Q|E)$, which represents the semantic support of the evidence tuple $E = (e_{txt}, e_{img}, e_{screenshot})$ for the query Q. Our implementation utilizes Dempster-Shafer Theory to robustly fuse heterogeneous retrieval scores and explicitly model uncertainty.

A.1.1 Score Normalization and Sharpening

Since retrieval scores from different embedding models (e.g., dense text embeddings vs. visual embeddings) often inhabit different numerical ranges, direct combination is infeasible. We first apply a normalization function to map raw scores into a probabilistic interval $[0, 1]$.

For a raw score s from modality i, given the empirically observed maximum (s_{max}) and minimum (s_{min}) scores for that modality, the normalized score $\tilde{s}_i$ is computed as:

$$\tilde{s}_i = \frac{s - s_{min}}{s_{max} - s_{min}} \tag{5}$$

A.1.2 Mass Function Construction

We define the *Frame of Discernment* as $\Theta = Y, N$, representing whether the evidence is "Relevant" (Y) or "Irrelevant" (N). For each normalized score $\tilde{s}_i$, we construct a Basic Probability Assignment, denoted as mass function $m_i(\cdot)$, based on hyperparameters α and β (representing our trust in the retriever's positive and negative capabilities):

$$m_i(Y) = \alpha \cdot \tilde{s}_i \tag{6}$$

$$m_i(N) = \beta \cdot (1 - \tilde{s}_i) \tag{7}$$

$$m_i(\Omega) = 1 - m_i(Y) - m_i(N) \tag{8}$$

Here, $m_i(\Omega)$ represents the *uncertainty* (or ignorance) remaining after observing the score. We enforce $m_i(\Omega) \geq 0$ by clipping. Based on experimental experience, we fix $\alpha = 0.7$ and $\beta = 0.6$, respectively, for all experiments.

A.1.3 Recursive Evidence Combination

To fuse evidence from multiple sources (Text, Image, Screenshot), we apply Dempster's Rule of Combination recursively. Let m_{curr} be the accumulated mass and m_{new} be the mass of the next incoming modality. The combined mass $m_{comb} = m_{curr} \oplus m_{new}$ is calculated as follows:

First, we compute the conflict coefficient K, which measures the degree of contradiction between the two evidence sources:

$$\begin{aligned} K = & m_{curr}(Y) \cdot m_{new}(N) + \\ & m_{curr}(N) \cdot m_{new}(Y) \end{aligned} \tag{9}$$

If $K \to 1$ (e.g., $K \geq 0.9$), indicating extreme conflict (one source is certain it's relevant, the other certain it's not), we forcefully set the likelihood to 0. Otherwise, we proceed with the update rules:

$$m_{comb}(H) = \frac{\sum_{X \cap Z = H} m_{curr}(X) \cdot m_{new}(Z)}{1 - K}, \quad \forall H \in Y, N, \Omega \tag{10}$$

Specifically for our binary frame, the update logic for the relevant hypothesis Y is:

$$\begin{aligned} m_{comb}(Y) = \frac{1}{1-K} \Big(& m_{curr}(Y) \cdot m_{new}(Y) \\ & + m_{curr}(Y) \cdot m_{new}(\Omega) \\ & + m_{curr}(\Omega) \cdot m_{new}(Y) \Big) \end{aligned} \tag{11}$$

A.1.4 Final Score

After fusing all evidence sources, we obtain the final mass distribution m_{final}. To convert this belief function into a probability measure suitable for ranking (i.e., $P(Q|E)$), we apply the *Pignistic Probability Transformation*, which distributes the uncertainty mass Ω equally among the singleton hypotheses:

$$P(Q|E) \approx \text{BetP}(Y) = m_{final}(Y) + \frac{m_{final}(\Omega)}{2} \tag{12}$$

This final score reflects the aggregated semantic likelihood, accounting for both the strength of individual matches and the consistency across modalities.

We provide the pseudocode for the likelihood estimation process $P(Q|E)$ in Algorithm 1.

A.2 Implementation of Priors

We detail the exact algorithms used to compute the structural prior $P(E)$ for both document layout and knowledge graph scenarios. The prior $P(E)$ measures the intrinsic coherence of an evidence tuple $E = (e_{txt}, e_{img}, e_{screenshot})$ derived from geometric or semantic constraints.

Algorithm 1 Likelihood Estimation via Dempster-Shafer Fusion

Require: Raw retrieval scores $\mathcal{S} = \{s_{txt}, s_{img}, s_{screenshot}\}$
Require: Hyperparameters $\alpha, \beta \in [0, 1]$ (Trust factors)
Require: Normalization function $\Phi(\cdot)$
Ensure: Likelihood score $P(Q|E)$

Initialize:
$m_{curr} \leftarrow \{Y : 0, N : 0, \Omega : 1\}$ {Initial state: total ignorance}
Preprocessing:
$\mathcal{S}' \leftarrow [\Phi(s)$ for s in $\mathcal{S}]$ {Normalize scores to $[0, 1]$}
for each score $s \in \mathcal{S}'$ **do**
  Construct Mass Function (BPA):
  $m_{new}(Y) \leftarrow \alpha \cdot s$
  $m_{new}(N) \leftarrow \beta \cdot (1 - s)$
  $m_{new}(\Omega) \leftarrow \max(0, 1 - m_{new}(Y) - m_{new}(N))$
  Compute Conflict Coefficient K:
  $K \leftarrow m_{curr}(Y) \cdot m_{new}(N) + m_{curr}(N) \cdot m_{new}(Y)$
  if $K \geq 0.999$ **then**
    return 0.0 {Extreme conflict detected}
  end if
  Apply Dempster's Rule:
  $D \leftarrow 1 - K$
  $m_{next}(Y) \leftarrow (m_{curr}(Y) \cdot m_{new}(Y) + m_{curr}(Y) \cdot m_{new}(\Omega) + m_{curr}(\Omega) \cdot m_{new}(Y))/D$
  $m_{next}(N) \leftarrow (m_{curr}(N) \cdot m_{new}(N) + m_{curr}(N) \cdot m_{new}(\Omega) + m_{curr}(\Omega) \cdot m_{new}(N))/D$
  $m_{next}(\Omega) \leftarrow (m_{curr}(\Omega) \cdot m_{new}(\Omega))/D$
  $m_{curr} \leftarrow m_{next}$
end for
Final Score:
$Score \leftarrow m_{curr}(Y) + m_{curr}(\Omega)/2$
return $Score$

A.2.1 Layout-Aware Prior Calculation

For document-native formats (e.g., PDFs), we utilize bounding box (BBox) coordinates to determine spatial proximity. The calculation of $P(E)_{\text{layout}}$ is implemented as a strict geometric consistency check.

Given a text node t, an image node v, and a layout shortcut node s, we first verify the consistency of the metadata. The tuple is considered valid only if all three components originate from the same source document D. Let $pg(\cdot)$ denote the page index and $d(\cdot, \cdot)$ denote the Euclidean distance between the centers of two bounding boxes.

The prior is computed as a binary indicator function with a relaxation for adjacent pages:

$$P(E)_{\text{bbox}} = \begin{cases} 1.0 & \text{if } \mathcal{C}_{\text{dist}} \wedge \mathcal{C}_{\text{page}} \\ \epsilon & \text{otherwise} \end{cases} \tag{13}$$

where ϵ is a small constant (e.g., $\epsilon = 0.1$). The conditions are defined as:

1. **Spatial Condition** ($\mathcal{C}_{\text{dist}}$): The spatial distance must be within a fraction of the page diagonal length L_{diag}:

$$d(t, v) < \tau \cdot L_{diag} \tag{14}$$

   where τ is a distance threshold parameter (e.g., $\tau = 2$).

2. **Pagination Condition** ($\mathcal{C}_{\text{page}}$): The elements must appear on the same or immediately adjacent pages relative to the screenshot reference:

$$\max(|pg(t) - pg(s)|, |pg(v) - pg(s)|) < \tau_{page} \tag{15}$$

This implementation ensures that a high prior probability is assigned solely to figure-caption pairs or spatially clustered information.

A.2.2 Knowledge Graph Topology Prior

For the knowledge graph, we construct a weighted connectivity map and compute the coherence of the retrieved triplet. This process consists of three steps: Relation Aggregation, Edge Normalization, and Triplet Coherence Scoring.

Step 1: Relation Aggregation. First, we map retrieved entities to their source document chunks. For any pair of chunks (u, v), we aggregate the weights of all relations r connecting them in the Knowledge Graph:

$$S_{uv} = \sum_{r \in \mathcal{R}_{u,v}} w_r \tag{16}$$

where w_r is the weight of a specific relation (defaulting to 1.0 or weighted by relation type).

Step 2: Probabilistic Edge Normalization. To convert the unbounded raw score S_{uv} into a probability $P_{edge}(u, v) \in [0, 1)$, we apply an exponential saturation function. This models the intuition that diminishing returns apply as evidence accumulates:

$$P_{edge}(u, v) = 1 - \exp(-\kappa \cdot S_{uv}) \tag{17}$$

where κ is a scaling factor (set to 0.1 in our experiments). This defines the pairwise connectivity map.

Step 3: Triplet Coherence Scoring. For a candidate tuple $E = (e_{id_1}, e_{id_2}, e_{id_3})$, we retrieve the three pairwise edge probabilities: p_{12}, p_{23}, p_{13}. We calculate their average as the final probability

$$P(E)_{\text{graph}} = \frac{p_{12} + p_{23} + p_{13}}{3} \tag{18}$$

B Reproducibility

In this section, we provide detailed configurations for the baseline models to ensure reproducibility. For most parameters, we use the default settings in the original repository. Across all RAG systems, we employ identical embedding and reranker models.

B.1 VisRAG

We follow to the reproducible baseline methodology provided by UltraRAG[3]. We adapt the data structures of MMLongBench-Doc and DocBench to align with the benchmark schema required by UltraRAG. Question answering is conducted using the official configuration template[4]. Finally, answer extraction and scoring are performed using the evaluation scripts provided by the respective benchmarks.

B.2 ViDoRAG

We follow the official implementation of ViDoRAG[5] with specific modifications to the embedding module. Specifically, we utilize `MinerU` for document parsing. For vector database construction, we employ `Qwen3-Embedding-4B` for text segments and `nomic-embed-multimodal-3b` for page-level PDF screenshots. We apply the proposed Gaussian Mixture Model (GMM)-based hybrid strategy to effectively handle multi-modal retrieval. `GPT-4o-mini` server as the backbone model for the Seeker, Inspector, and Answer agents.

B.3 RAGFlow

We deploy RAGFlow using the official Docker image[6]. We employ `Qwen3-Embedding-4B` and `bge-reranker-v2-m3` (host via VLLM) for text embedding and reranking, respectively. For multimodal content, `Qwen3-VL-32B` is utilized to generate descriptions. All other parsing parameters retain their default settings.

During the question-answering phase, we use `GPT-4o-mini` and enable the following features: *Keyword Analysis*, *TOC Enhance*, *Use Knowledge Graph*, and *Reasoning*. The specific chat configuration parameters are detailed in Listing 1.

Listing 1: Configuration parameters for RAGFlow.

```
{
  "language": "English",
  "llm": {
    "max_tokens": 4096,
    "model_name": "azure-gpt-4o-mini___OpenAI-
        ↪ API@OpenAI-API-Compatible",
    "temperature": 0.1
  },
  "meta_data_filter": {
    "method": "automatic"
  },
  "prompt": {
    "empty_response": "",
    "keyword": true,
    "keywords_similarity_weight": 0.7,
    "opener": "Hi! I'm your assistant. What can I do
        ↪  for you?",
    "prompt": "You are an intelligent assistant.
        ↪ Please summarize the content of the
        ↪ knowledge base to answer the question.
        ↪ Please list the data in the knowledge
        ↪ base and answer in detail. When all
        ↪ knowledge base content is irrelevant to
        ↪ the question, your answer must include
        ↪ the sentence \'The answer you are
        ↪ looking for is not found in the
        ↪ knowledge base!\' Answers need to
        ↪ consider chat history.\n      Here is
        ↪ the knowledge base:\n      {knowledge}\n
        ↪        The above is the knowledge base.",
    "reasoning": true,
    "refine_multiturn": false,
    "rerank_model": "bge-reranker-v2-m3___OpenAI-
        ↪ API@OpenAI-API-Compatible",
    "show_quote": true,
    "similarity_threshold": 0.2,
    "toc_enhance": true,
    "top_n": 8,
    "tts": false,
    "use_kg": true,
    "variables": [
      {
        "key": "knowledge",
        "optional": true
      }
    ]
  },
  "prompt_type": "simple",
  "top_k": 1024
}
```

[3] https://github.com/OpenBMB/UltraRAG

[4] https://github.com/OpenBMB/UltraRAG/blob/main/examples/visrag.yaml

[5] https://github.com/Alibaba-NLP/ViDoRAG

[6] https://hub.docker.com/r/infiniflow/ragflow

B.4 RAGAnything

Consistent with our other setups, we employ `MinerU` as the document parser. We utilize `Qwen3-VL-32B` to generate document descriptions and construct the knowledge graph. The vector retrieval component relies on `Qwen3-Embedding-4B` and `bge-reranker-v2-m3`. Finally, `GPT-4o-mini` is employed as the downstream question-answering model.

C Prompts in BayesRAG

The specific prompts employed in BayesRAG are detailed below:

Prompt

You are an intelligent assistant. Please summarize the content of the knowledge base (both in text, images, and screenshots of pdf pages) to answer questions, think step by step.

First, analyze the keywords in the question, and then check whether all texts and images of the given knowledge base contain content related to the keywords in the question. Finally summarize and answer questions.

When all knowledge base content is not related to the question, your answer should include the sentence "The answer you want is not found in the knowledge base!"

Here is the question:

Question: {question}

The above is the question.

Here is the knowledge base:

{knowledge}

The above is the knowledge base.

Figure 4: Prompts used by the model to predict answers.

Prompt

You are required to determine if a predicted answer is correct or can reasonably answer the question compared to the ground truth. The question will be placed within <question></question> tags, answer type will be placed within <type></type> tags, predicted answer will be placed within <predict></predict> tags, and the ground truth answer will be placed within <gt></gt> tags.

You must output the final score as a floating-point number (0.0 to 1.0) enclosed within '<answer>' tags.

Output Format: Output **only** the final numerical score inside the tags. Do not provide reasoning.

Example:
<answer> 1.0 </answer>

Question: <question> {question} </question>

Predict Answer: <predict> {predict answer} </predict>

Ground Truth: <gt> {reference answer} </gt>

Figure 5: Prompts used by evaluating the answer in MMLongBench-Doc.

D Additional Case Studies

In the main text, we demonstrate cases where BayesRAG improves retrieval recall for low-salience details and disambiguates visually similar charts. In this appendix, we provide additional qualitative examples highlighting a critical advantage of our framework: the ability to solve complex, multi-hop reasoning tasks where baseline methods fail to retrieve the correct evidence entirely.

A common failure mode in existing multimodal RAG systems is the inability to locate evidence that is intricately embedded within structured documents, such as complex tables, stacked charts, or dense infographics. Single-stream retrievers (re-

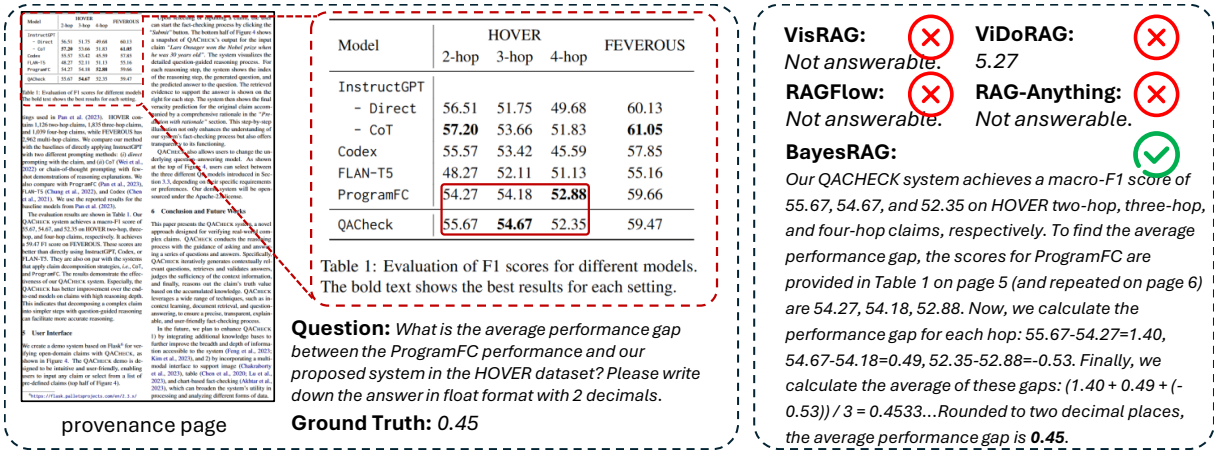

Figure 6: Complex tabular reasoning requiring multi-hop numerical extraction.

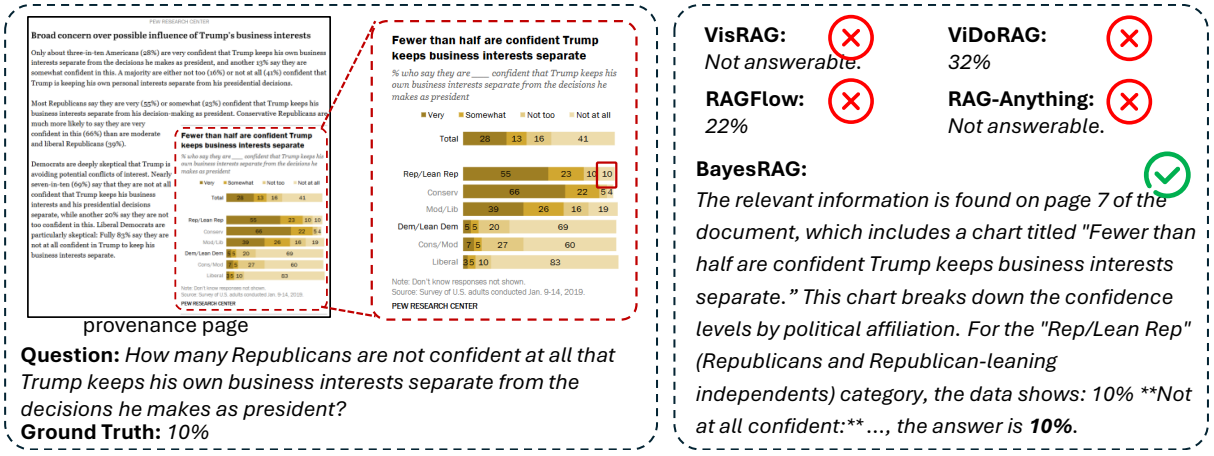

Figure 7: Fine-grained visual-textual alignment in stacked bar charts.

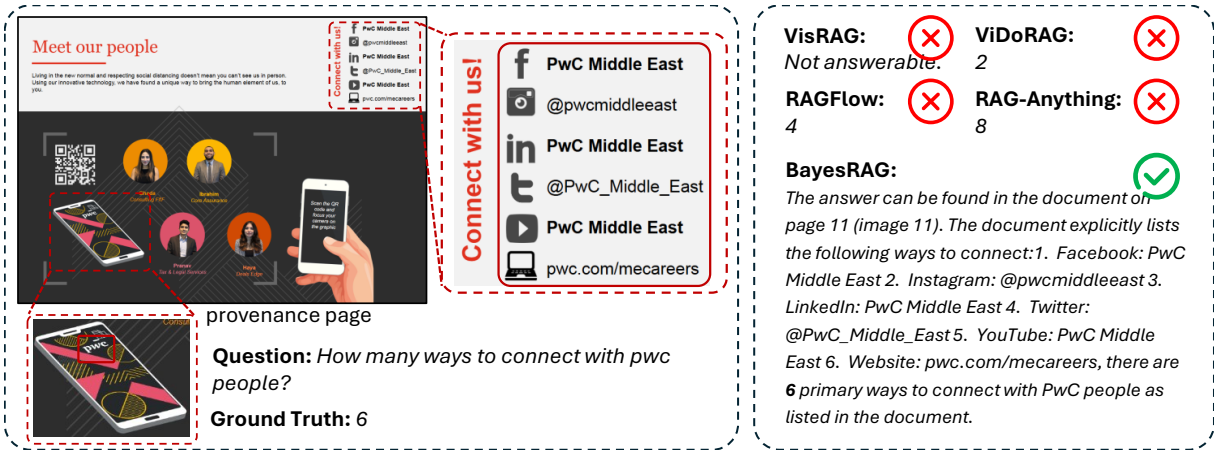

Figure 8: Information enumeration from unstructured infographics.

lying solely on text chunks or naive visual embeddings) often miss these cues due to a lack of fine-grained semantic alignment between the query and the visual structure.

BayesRAG addresses this through its parallel retrieval architecture, simultaneously querying text chunks, cropped image regions, and full-page PDF screenshots. Crucially, the subsequent Bayesian fusion mechanism calculates a posterior probability representing the semantic consistency between the query and these heterogeneous evidence sources. This allows BayesRAG to identify and utilize evidence that requires establishing strong cross-modal links (e.g., mapping legend colors to textual categories, or aligning table headers with specific numerical cells), tasks where baselines frequently hallucinate or fail to answer.

We present three such challenging scenarios below: complex tabular reasoning (Figure 6), fine-grained chart interpretation (Figure 7), and infographic enumeration (Figure 8).

E Reliability of Automated Evaluation

To address concerns regarding the reliability of Qwen3-32B as an automated judge in MMLongBench-Doc, we conduct a human verification study on a stratified sample of 200 instances. Our analysis yields two critical observations that validate our evaluation protocol: **High Human-Model Agreement.** We observe an exceptionally high consistency between the Qwen3-32B judge and expert human annotators, achieving an agreement rate of **98%**. Disagreements are rare and mostly occur in highly ambiguous cases where even humans might diverge. This statistical evidence strongly supports the use of Qwen3-32B as a surrogate for human evaluation in this specific context. **Suitability for Factoid QA.** The high reliability stems from the intrinsic nature of the MMLongBench-Doc dataset. Unlike open-ended generation tasks, this benchmark primarily consists of factoid questions requiring specific numerical values or fixed entities as answers. In such closed-ended scenarios, the evaluation task simplifies to verification rather than subjective critique.

Crucially, our qualitative analysis highlights that the LLM judge effectively resolves the limitations of the original rule-based scoring scripts. Traditional regex-based methods often produce false negatives due to valid lexical variations. By recognizing semantic equivalence, Qwen3-32B provides a more accurate assessment of model performance than rigid string matching, ensuring that correct answers are not unfairly penalized due to minor formatting deviations.

F Efficiency and Computational Analysis

To evaluate the practical overhead of BayesRAG, we break down the time latency across the entire inference pipeline. Table 6 details the average time costs recorded on the MMLongBench-Doc. All experiments are conducted on a server with NVIDIA Tesla A800 80GB GPUs and a 12-core CPU.

Module	Time (s)
Vector Retrieval (Text)	5.81
Vector Retrieval (Image)	26.18
Vector Retrieval (Screenshot)	52.09
Text Reranking (bge-reranker-v2-m3)	9.36
Bayes Fusion & Reranking	**10.21**
LLM Generation (QA)	29.17

Table 6: Average time cost (seconds) per query breakdown. Note that *Vector Retrieval (Image/Screenshot)* is using an expanded search space (Top-k=512). *Bayes Fusion* represents the re-ranking overhead introduced by our method.

Retrieval Latency vs. Recall. As observed, the visual retrieval components (Image and Screenshot) dominate the latency, consuming approximately 52s. This is an intentional design choice rather than an algorithmic limitation. To validate the "Recall Enhancement" capabilities of BayesRAG (as discussed in Section 3.4), we expand the image retrieval window size to Top-512 (compared to the standard Top-10 or Top-20). This ensures that low-salience evidence is captured in the initial candidate pool. In production environments, this latency can be significantly reduced via parallelized asynchronous retrieval, etc, independent of our proposed fusion algorithm.

Efficiency of Bayesian Inference. A critical finding is the comparison between standard text reranking and our Bayesian fusion process. The **Bayes Fusion** step takes **10.21s**, which introduces only a marginal overhead ($\sim$0.85s) compared to the standard **Text Reranking** (9.36s). Theoretically, combining heterogeneous candidates (Text, Image, Screenshot) from a Top-512 pool could lead to a combinatorial explosion ($512^3 \approx 134$ million tuples). However, BayesRAG circumvents this issue

Method	Reports	Tutorials	Academic	Guidebooks	Brochures	Industry	Financial	Overall	Generator Model
RAGAnything	45.3	46.0	34.6	43.5	39.6	33.3	42.7	41.5	GPT-4o-mini
BayesRAG (ours)	46.7	53.6	34.8	46.4	39.4	38.6	48.6	44.1	
RAGAnything	53.4	47.4	40.2	51.9	39.6	37.0	50.4	47.1	Qwen3-VL-32B
BayesRAG (ours)	61.8	57.8	58.2	58.9	54.9	54.4	65.4	59.4	

Table 7: Performance comparison with different generator models on MMLongBench-Doc. The switch to the more capable Qwen3-VL-32B results in a significantly larger performance boost for BayesRAG compared to RAGAnything, highlighting our method's ability to retrieve high-quality evidence that stronger models can better exploit.

Method	Domains					Types				Overall	Generator Model
	Aca.	Fin.	Gov.	Law.	News	Text	Meta	MM.	Unans.		
RAGAnything	51.4	42.3	43.2	50.2	57.5	77.6	10.0	57.4	11.2	48.7	GPT-4o-mini
BayesRAG (ours)	52.4	37.5	56.0	52.3	66.8	84.2	19.7	43.8	25.8	51.2	
RAGAnything	57.4	44.7	49.3	50.7	61.0	82.0	11.2	60.7	19.3	52.4	Qwen3-VL-32B
BayesRAG (ours)	65.6	53.8	61.4	59.6	70.3	87.3	30.2	62.6	39.5	61.7	

Table 8: Performance comparison on DocBench with different generators. BayesRAG consistently outperforms baselines, with the performance gap widening significantly when using the stronger Qwen3-VL-32B generator, verifying superior evidence recall.

by imposing **metadata constraints** before probability calculation. Specifically, we restrict the construction of evidence tuples $E = (t, v, s)$ to those derived from the same source document. Furthermore, candidates yielding low probabilities—based on retrieval scores and knowledge graph distributions—are directly pruned. These constraints effectively transform the dense combinatorial space into a highly sparse matrix, significantly reducing computational complexity. This confirms that BayesRAG provides a computationally efficient solution for sophisticated multimodal reasoning, effectively shifting the heavy lifting from expensive LLM reasoning to a lightweight probabilistic verification layer.

G Impact of Generator Models

To disentangle the contribution of the retrieval module from the generation capabilities, we conduct a comparative analysis using different backbones: **GPT-4o-mini** and the more powerful **Qwen3-VL-32B**. The results on MMLongBench-Doc and DocBench are detailed in Table 7 and Table 8.

Performance Sensitivity to Generator Capacity. As expected, replacing the generator with the SOTA Qwen3-VL-32B leads to universal performance improvements. However, a critical observation is the *disproportionate gain* achieved by BayesRAG compared to the baseline RAGAnything. On MMLongBench-Doc, while RAGAnything improves by 5.6% (41.5% $\rightarrow$ 47.1%), BayesRAG achieves a dramatic surge of **15.3%** (44.1% $\rightarrow$ 59.4%). A similar trend is observed on DocBench, where the performance gap between BayesRAG and RAGAnything widens significantly from 2.5% under GPT-4o-mini to 9.3% under Qwen3-VL-32B.

Evidence for Superior Effective Recall. This phenomenon provides compelling evidence for the superior retrieval quality of BayesRAG. In RAG systems, the performance is bounded by the quality of the retrieved context. A stronger generator cannot hallucinate correct answers if the supporting evidence is missing (low recall). The fact that BayesRAG benefits substantially more from a stronger generator indicates that our Bayesian fusion mechanism successfully retrieves high-fidelity, complex multimodal evidence that is present in the context but under-utilized by the weaker GPT-4o-mini. In contrast, the baseline's smaller improvement suggests its retrieval bottleneck lies in missing evidence (low recall), which even a superior generator cannot mitigate. Thus, these results confirm that BayesRAG delivers significantly higher **Effective Recall**, providing a richer and more accurate evidence set that unlocks the full potential of advanced MLLMs.