# RAG-ANYTHING: ALL-IN-ONE RAG FRAMEWORK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Retrieval-Augmented Generation (RAG) has emerged as a fundamental paradigm for expanding Large Language Models beyond their static training limitations. However, a critical misalignment exists between current RAG capabilities and real-world information environments. Modern knowledge repositories are inherently multimodal, containing rich combinations of textual content, visual elements, structured tables, and mathematical expressions. Yet existing RAG frameworks are limited to textual content, creating fundamental gaps when processing multimodal documents. We present RAG-Anything, a unified framework that enables comprehensive knowledge retrieval across all modalities. Our approach reconceptualizes multimodal content as interconnected knowledge entities rather than isolated data types. The framework introduces dual-graph construction to capture both cross-modal relationships and textual semantics within a unified representation. We develop cross-modal hybrid retrieval that combines structural knowledge navigation with semantic matching. This enables effective reasoning over heterogeneous content where relevant evidence spans multiple modalities. RAG-Anything demonstrates superior performance on challenging multimodal benchmarks, achieving significant improvements over state-of-the-art methods. Performance gains become particularly pronounced on long documents where traditional approaches fail. Our framework establishes a new paradigm for multimodal knowledge access, eliminating the architectural fragmentation that constrains current systems. Our framework is open-sourced at: https://anonymous.4open.science/r/RAG-Anything-WWW-2B64/.

## 1 INTRODUCTION

Retrieval-Augmented Generation (RAG) has emerged as a fundamental paradigm for expanding the knowledge boundaries of Large Language Models (LLM) beyond their static training limitations Zhang et al. (2025). By enabling dynamic retrieval and incorporation of external knowledge during inference, RAG systems transform static language models into adaptive, knowledge-aware systems. This capability has proven essential for applications requiring up-to-date information, domain-specific knowledge, or factual grounding that extends beyond pre-training corpora.

However, existing RAG frameworks focus exclusively on text-only knowledge while neglecting the rich multimodal information present in real-world documents. This limitation fundamentally misaligns with how information exists in authentic environments. Real-world knowledge repositories are inherently heterogeneous and multimodal Abootorabi et al. (2025). They contain rich combinations of textual content, visual elements, structured tables, and mathematical expressions across diverse document formats. This textual assumption forces existing RAG systems to either discard non-textual information entirely or flatten complex multimodal content into inadequate textual approximations.

The consequences of this limitation become particularly severe in knowledge-intensive domains where multimodal content carries essential meaning. Three representative scenarios illustrate the critical need for multimodal RAG capabilities. In **Scientific Research**, experimental results are primarily communicated through plots, diagrams, and statistical visualizations that contain core discoveries invisible to text-only systems. **Financial Analysis** relies heavily on market charts, correlation matrices, and performance tables, where investment insights are encoded in visual patterns rather than textual descriptions. **Medical Literature Analysis** depends on radiological images, diagnostic charts, and clinical data tables that contain life-critical information for accurate diagnosis and treatment. Current RAG frameworks systematically exclude these vital knowledge sources,



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.

creating fundamental gaps that render them inadequate for real-world applications. Multimodal RAG is therefore a critical advancement, necessary to bridge these knowledge gaps and enable truly comprehensive intelligence across all modalities of human knowledge representation.

Addressing multimodal RAG presents three fundamental technical challenges that demand principled solutions. This makes it significantly more complex than traditional text-only approaches. The naive solution of converting all multimodal content to textual descriptions introduces severe information loss. Visual elements such as charts, diagrams, and spatial layouts contain semantic richness that cannot be adequately captured through text alone. These inherent limitations necessitate the design of effective technical components. Such components must be specifically designed to handle multimodal complexity and preserve the full spectrum of information contained within diverse content types.

**Technical Challenges**. • **First**, the **unified multimodal representation** challenge requires seamlessly integrating diverse information types. The system must preserve their unique characteristics and cross-modal relationships. This demands advanced multimodal encoders that can capture both intra-modal and inter-modal dependencies without losing essential visual semantics. • **Second**, the **structure-aware decomposition** challenge demands intelligent parsing of complex layouts. The system must maintain spatial and hierarchical relationships crucial for understanding. This requires specialized layout-aware parsing modules that can interpret document structure and preserve contextual positioning of multimodal elements. • **Third**, the **cross-modal retrieval** challenge necessitates sophisticated mechanisms that can navigate between different modalities. These mechanisms must reason over their interconnections during retrieval. This calls for cross-modal alignment systems capable of understanding semantic correspondences across text, images, and structured data. These challenges are amplified in long-context scenarios. Relevant evidence is dispersed across multiple modalities and sections, requiring coordinated reasoning across heterogeneous information sources.

**Our Contributions**. To address these challenges, we introduce RAG-Anything, a unified framework that fundamentally reimagines multimodal knowledge representation and retrieval. Our approach employs a **dual-graph construction strategy** that elegantly bridges the gap between cross-modal understanding and fine-grained textual semantics. Rather than forcing diverse modalities into text-centric pipelines, RAG-Anything constructs **complementary knowledge graphs** that preserve both multimodal contextual relationships and detailed textual knowledge. This design enables seamless integration of visual elements, structured data, and mathematical expressions within a unified retrieval framework. The system maintains **semantic integrity** across modalities while ensuring efficient **cross-modal reasoning capabilities** throughout the process.

Our **cross-modal hybrid retrieval** mechanism strategically combines **structural knowledge navigation** with **semantic similarity matching**. This architecture addresses the fundamental limitation of existing approaches that rely solely on embedding-based retrieval or keyword matching. RAG-Anything leverages explicit graph relationships to capture multi-hop reasoning patterns. It simultaneously employs dense vector representations to identify semantically relevant content that lacks direct structural connections. The framework introduces **modality-aware query processing** and **cross-modal alignment systems**. These enable textual queries to effectively access visual and structured information. This unified approach eliminates the architectural fragmentation that plagues current multimodal RAG systems. It delivers superior performance particularly on long-context documents where relevant evidence spans multiple modalities and document sections.

**Experimental Validation**. To validate the effectiveness of our proposed approach, we conduct comprehensive experiments on two challenging multimodal benchmarks: DocBench and MMLongBench. Our evaluation demonstrates that RAG-Anything achieves superior performance across diverse domains. The framework represents substantial improvements over state-of-the-art baselines. Notably, our performance gains become increasingly significant as content length increases. We observe particularly pronounced advantages on long-context materials. This validates our core hypothesis that dual-graph construction and cross-modal hybrid retrieval are essential for handling complex multimodal materials. Our ablation studies reveal that graph-based knowledge representation provides the primary performance gains. Traditional chunk-based approaches fail to capture the structural relationships critical for multimodal reasoning. Case studies further demonstrate that our framework excels at precise localization within complex layouts. The system effectively disambiguates similar terminology and navigates multi-panel visualizations through structure-aware retrieval mechanisms.

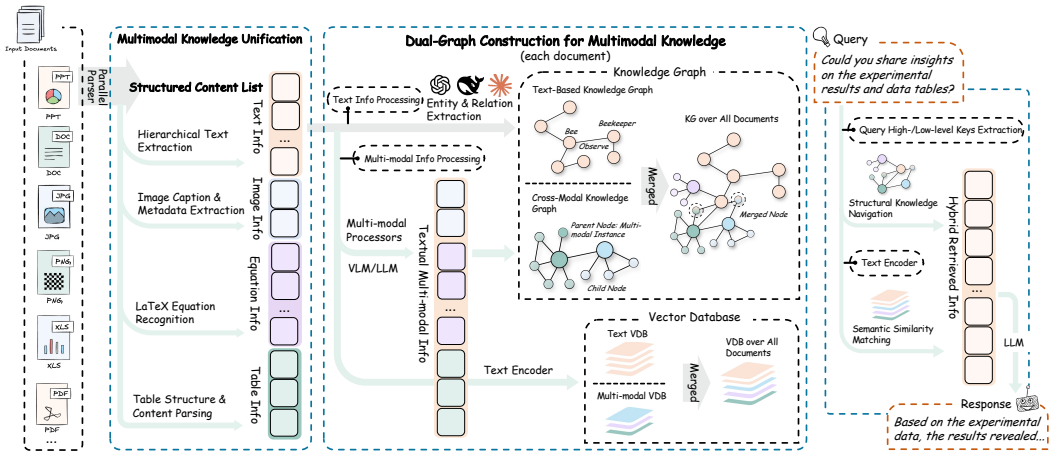


Figure 1: Overview of our proposed universal RAG framework RAG-Anything.

## 2 The RAG-Anything Framework

RAG-Anything provides a unified approach for retrieving and processing knowledge from heterogeneous multimodal information sources, addressing the fundamental challenge of handling diverse data modalities and document formats within a single retrieval pipeline. The framework comprises three core components: universal indexing for multimodal knowledge, cross-modal adaptive retrieval, and knowledge-enhanced response generation. This integrated design enables effective knowledge utilization across modalities while maintaining computational efficiency.

### 2.1 Universal Representation for Heterogeneous Knowledge

A key requirement for universal knowledge access is the ability to represent heterogeneous multimodal content in a unified, retrieval-oriented abstraction. Unlike existing pipelines that simply parse documents into text segments, RAG-Anything introduces **Multimodal Knowledge Unification**. This process decomposes raw inputs into atomic knowledge units while preserving their structural context and semantic alignment. For instance, RAG-Anything ensures that figures remain grounded in their captions, equations remain linked to surrounding definitions, and tables stay connected to explanatory narratives. This transforms heterogeneous files into a coherent substrate for cross-modal retrieval.

Formally, each knowledge source $k_i \in \mathcal{K}$ (e.g., a web page) is decomposed into atomic content units:

$$k_i \xrightarrow{\text{Decompose}} \{c_j = (t_j, x_j)\}_{j=1}^{n_i}, \tag{1}$$

where each unit $c_j$ consists of a modality type $t_j \in$ text, image, table, equation, ... and its corresponding raw content $x_j$. The content $x_j$ represents the extracted information from the original knowledge source, processed in a modality-aware manner to preserve semantic integrity.

To ensure high-fidelity extraction, RAG-Anything leverages specialized parsers for different content types. Text is segmented into coherent paragraphs or list items. Figures are extracted with associated metadata such as captions and cross-references. Tables are parsed into structured cells with headers and values. Mathematical expressions are converted into symbolic representations. The resulting $x_j$ preserves both content and structural context within the source. This provides a faithful, modality-consistent representation. The decomposition abstracts diverse file formats into atomic units while maintaining their hierarchical order and contextual relationships. This canonicalization enables uniform processing, indexing, and retrieval of multimodal content within our framework.

#### 2.1.1 Dual-Graph Construction for Multimodal Knowledge

While multimodal knowledge unification provides a uniform abstraction across modalities, directly constructing a single unified graph often risks overlooking modality-specific structural signals. The proposed RAG-Anything addresses this challenge through a dual-graph construction strategy. The system first builds a **cross-modal knowledge graph** that faithfully grounds non-textual modalities

within their contextual environment. It then constructs a **text-based knowledge graph** using established text-centric extraction pipelines. These complementary graphs are merged through entity alignment. This design ensures accurate cross-modal grounding and comprehensive coverage of textual semantics, enabling richer knowledge representation and robust retrieval.

- **Cross-Modal Knowledge Graph**: Non-textual content like images, tables, and equations contains rich semantic information that traditional text-only approaches often overlook. To preserve this knowledge, RAG-Anything constructs a multimodal knowledge graph where non-text atomic units are transformed into structured graph entities. RAG-Anything leverages multimodal large language models to derive two complementary textual representations from each atomic content unit. The first is a **detailed description** $d_j^{\text{chunk}}$ optimized for cross-modal retrieval. The second is an **entity summary** $e_j^{\text{entity}}$ containing key attributes such as entity name, type, and description for graph construction. The generation process is context-aware, processing each unit with its local neighborhood $C_j = \{c_k \mid |k - j| \leq \delta\}$, where $\delta$ controls the contextual window size. This ensures representations accurately reflect each unit's role within the broader document structure.

  Building on these textual representations, RAG-Anything constructs the graph structure using non-text units as anchor points. For each non-text unit $c_j$, the graph extraction routine $R(\cdot)$ processes its description $d_j^{\text{chunk}}$ to identify fine-grained entities and relations:

$$(\mathcal{V}_j, \mathcal{E}_j) = R(d_j^{\text{chunk}}), \tag{2}$$

  where $\mathcal{V}_j$ and $\mathcal{E}_j$ denote the sets of intra-chunk entities and their relations, respectively. Each atomic non-text unit is associated with a multimodal entity node $v_j^{\text{mm}}$ that serves as an anchor for its intra-chunk entities through explicit `belongs_to` edges:

$$\tilde{V} = \{v_j^{\text{mm}}\}_j \cup \bigcup_j \mathcal{V}_j, \tag{3}$$

$$\tilde{E} = \bigcup_j \mathcal{E}_j \cup \bigcup_j \{(u \xrightarrow{\texttt{belongs_to}} v_j^{\text{mm}}) : u \in \mathcal{V}_j\}. \tag{4}$$

  This construction preserves modality-specific grounding while ensuring non-textual content is contextualized by its textual neighborhood. This enables reliable cross-modal retrieval and reasoning.

- **Text-based Knowledge Graph**: For text modality chunks, we construct a traditional text-based knowledge graph following established methodologies similar to LightRAG (Guo et al., 2024) and GraphRAG (Edge et al., 2024). The extraction process operates directly on textual content $x_j$ where $t_j =$ text, leveraging named entity recognition and relation extraction techniques to identify entities and their semantic relationships. Given the rich semantic information inherent in textual content, multimodal context integration is not required for this component. The resulting text-based knowledge graph captures explicit knowledge and semantic connections present in textual portions of documents, complementing the multimodal graph's cross-modal grounding capabilities.

#### 2.1.2 Graph Fusion and Index Creation

The separate cross-modal and text-based knowledge graphs capture complementary aspects of document semantics. Integrating them creates a unified representation leveraging visual-textual associations and fine-grained textual relationships for enhanced retrieval.

• (i) **Entity Alignment and Graph Fusion**. To create a unified knowledge representation, we merge the multimodal knowledge graph $(\tilde{V}, \tilde{E})$ and text-based knowledge graph through entity alignment. This process uses entity names as primary matching keys to identify semantically equivalent entities across both graph structures. The integration consolidates their representations, creating a comprehensive knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. This graph captures both multimodal contextual relationships and text-based semantic connections. The merged graph provides a holistic view of the document collection. This enables effective retrieval by leveraging visual-textual associations from the multimodal graph and fine-grained textual knowledge relationships from the text-based graph.

• (ii) **Dense Representation Generation**. To enable efficient similarity-based retrieval, we construct a comprehensive embedding table $\mathcal{T}$ that encompasses all components generated during the indexing process. We encode dense representations for all graph entities, relationships, and atomic content

chunks across modalities using an appropriate encoder. This creates a unified embedding space where each component $s \in$ entities, relations, chunks is mapped to its corresponding dense representation:

$$\mathcal{T} = \text{emb}(s) : s \in \mathcal{V} \cup \mathcal{E} \cup c_{j_j}, \tag{5}$$

where $\text{emb}(\cdot)$ denotes the embedding function tailored for each component type. Together, the unified knowledge graph $\mathcal{G}$ and the embedding table $\mathcal{T}$ constitute the complete retrieval index $\mathcal{I} = (\mathcal{G}, \mathcal{T})$. This provides both structural knowledge representation and dense vector space for efficient cross-modal similarity search during the subsequent retrieval stage.

### 2.2 Cross-Modal Hybrid Retrieval

The retrieval stage operates on the index $\mathcal{I} = (\mathcal{G}, \mathcal{T})$ to identify relevant knowledge components for a given user query. Traditional RAG methods face significant limitations when dealing with multimodal documents. They typically rely on semantic similarity within single modalities and fail to capture the rich interconnections between visual, mathematical, tabular, and textual elements. To address these challenges, our framework introduces a cross-modal hybrid retrieval mechanism. This mechanism leverages structural knowledge and semantic representations across heterogeneous modalities.

**Modality-Aware Query Encoding.** Given a user query $q$, we first perform modality-aware query analysis to extract lexical cues and potential modality preferences (e.g., terms like "figure," "chart," "table," or "equation" signal the expected modality of relevant information). We then compute a unified text embedding $\mathbf{e}_q$ using the same encoder employed during indexing, ensuring consistency between query and knowledge representations and enabling textual queries to access multimodal content through shared representations.

**Hybrid Knowledge Retrieval Architecture**. Recognizing that knowledge relevance manifests through both explicit structural connections and implicit semantic relationships, we design a hybrid retrieval architecture that strategically combines two complementary mechanisms.

• (i) **Structural Knowledge Navigation**. This mechanism addresses the challenge of capturing explicit relationships and multi-hop reasoning patterns. Traditional keyword-based retrieval often fails to identify knowledge connected through intermediate entities or cross-modal relationships. To overcome this limitation, we exploit the structural properties encoded within our unified knowledge graph G. We employ keyword matching and entity recognition to locate relevant graph components. The retrieval process begins with exact entity matching against query terms.

We then perform strategic neighborhood expansion to include related entities and relationships within a specified hop distance. This structural approach proves particularly effective at uncovering high-level semantic connections and entity-relation patterns that span multiple modalities. It capitalizes on the rich cross-modal linkages established in our multimodal knowledge graph. The structural navigation yields candidate set $\mathcal{C}_{\text{stru}}(q)$ containing relevant entities, relationships, and their associated content chunks that provide comprehensive contextual information.

• (ii) **Semantic Similarity Matching**. This mechanism addresses the challenge of identifying semantically relevant knowledge that lacks explicit structural connections. While structural navigation excels at following explicit relationships, it may miss relevant content that is semantically related but not directly connected in the graph topology. To bridge this gap, we conduct dense vector similarity search between the query embedding $\mathbf{e}_q$ and all components stored in embedding table $\mathcal{T}$.

This approach encompasses atomic content chunks across all modalities, graph entities, and relationship representations, enabling fine-grained semantic matching that can surface relevant knowledge even when traditional lexical or structural signals are absent. The learned embedding space captures nuanced semantic relationships and contextual similarities that complement the explicit structural signals from the navigation mechanism. This retrieval pathway returns the top-k most semantically similar chunks $\mathcal{C}_{\text{seman}}(q)$ ranked by cosine similarity scores, ensuring comprehensive coverage of both structurally and semantically relevant knowledge.

**Candidate Pool Unification**. Both retrieval pathways may return overlapping candidates with differing relevance signals. This necessitates a principled approach to unify and rank results. Retrieval candidates from both pathways are unified into a comprehensive candidate pool: $\mathcal{C}(q) = \mathcal{C}_{\text{stru}}(q) \cup \mathcal{C}_{\text{seman}}(q)$. Simply merging candidates would ignore distinct evidence each pathway provides. It would fail to account for redundancy between retrieved content.

Table 1: Statistics of Experimental Datasets.

| Dataset | # Documents | # Avg. Pages | # Avg. Tokens | # Doc Types | # Questions |
|---|---|---|---|---|---|
| DocBench | 229 | 66 | 46377 | 5 | 1102 |
| MMLongBench | 135 | 47.5 | 21214 | 7 | 1082 |

• (i) **Multi-Signal Fusion Scoring**. To address these challenges, we apply a sophisticated fusion scoring mechanism integrating multiple complementary relevance signals. These include structural importance derived from graph topology, semantic similarity scores from embedding space, and query-inferred modality preferences obtained through lexical analysis. This multi-faceted scoring approach ensures that final ranked candidates $\mathcal{C}^\star(q)$ effectively balance structural knowledge relationships with semantic relevance while appropriately weighting different modalities based on query characteristics.

• (ii) **Hybrid Retrieval Integration**. The resulting hybrid retrieval mechanism enables our framework to leverage the complementary strengths of both knowledge graphs and dense representations. This provides comprehensive coverage of relevant multimodal knowledge for response generation.

### 2.3 From Retrieval to Synthesis

Effective multimodal question answering requires preserving rich visual semantics while maintaining coherent grounding across heterogeneous knowledge sources. Simple text-only approaches lose crucial visual information, while naive multimodal methods struggle with coherent cross-modal integration. Our synthesis stage addresses these challenges by systematically combining retrieved multimodal knowledge into comprehensive, evidence-grounded responses.

• (i) **Building Textual Context**. Given the top-ranked retrieval candidates $\mathcal{C}^\star(q)$, we construct a structured textual context. We concatenate textual representations of all retrieved components, including entity summaries, relationship descriptions, and chunk contents. The concatenation incorporates appropriate delimiters to indicate modality types and hierarchical origins. This approach ensures the language model can effectively parse and reason over heterogeneous knowledge components.

• (ii) **Recovering Visual Content**. For multimodal chunks corresponding to visual artifacts, we perform dereferencing to recover original visual content, creating $\mathcal{V}^\star(q)$. This design maintains consistency with our unified embedding strategy. Textual proxies enable efficient retrieval while authentic visual content provides rich semantics necessary for sophisticated reasoning during synthesis.

The synthesis process jointly conditions on both the assembled comprehensive textual context and dereferenced visual artifacts using a vision-language model:

$$\text{Response} = \text{VLM}(q, \mathcal{P}(q), \mathcal{V}^\star(q)), \tag{6}$$

where the VLM integrates information from query, textual context, and visual content. This unified conditioning enables sophisticated visual interpretation while maintaining grounding in retrieved evidence. The resulting responses are both visually informed and factually grounded.

## 3 Evaluation

### 3.1 Experimental Settings

**Evaluation Datasets**. We conduct comprehensive evaluations on two challenging multimodal Document Question Answering (DQA) benchmarks that reflect real-world complexity and diversity. DocBench (Zou et al., 2024) provides a rigorous testbed with 229 multimodal documents spanning five critical domains: Academia, Finance, Government, Laws, and News. The dataset includes 1,102 expert-crafted question-answer pairs. These documents are notably extensive, averaging 66 pages and approximately 46,377 tokens, which presents substantial challenges for long-context understanding.

MMLongBench (Ma et al., 2024) complements this evaluation by focusing specifically on long-context multimodal document comprehension. It features 135 documents across 7 diverse document types with 1,082 expert-annotated questions. Together, these benchmarks provide comprehensive coverage of the multimodal document understanding challenges that RAG-Anything aims to address.

Table 2: Accuracy (%) on DocBench Dataset. Performance results with best scores highlighted in **dark blue** and second-best in **light blue**. Domain categories include Academia (Aca.), Finance (Fin.), Government (Gov.), Legal Documents (Law), and News Articles (News). Document types are categorized as Text-only (Txt.), Multimodal (Mm.), and Unanswerable queries (Una.).

| Method | Domains | | | | | Types | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | Aca. | Fin. | Gov. | Law. | News | Txt. | Mm. | Una. | |
| GPT-4o-mini | 40.3 | 46.9 | 60.3 | 59.2 | 61.0 | 61.0 | 43.8 | **49.6** | 51.2 |
| LightRAG | 53.8 | 56.2 | 59.5 | **61.8** | **65.7** | **85.0** | 59.7 | 46.8 | 58.4 |
| MMGraphRAG | **64.3** | 52.8 | **64.9** | 40.0 | 61.5 | 67.6 | 66.0 | **60.5** | **61.0** |
| RAGAnything | **61.4** | **67.0** | **61.5** | **60.2** | **66.3** | **85.0** | **76.3** | 46.0 | **63.4** |

Table 3: Accuracy (%) on MMLongBench across different domains and overall performance. Best results are highlighted in **dark blue** and second-best in **light blue**.. Domain categories include Research Reports/Introductions (Res.), Tutorials/Workshops (Tut.), Academic Papers (Acad.), Guidebooks (Guid.), Brochures (Broch.), Administration/Industry Files (Admin.), and Financial Reports (Fin.).

| Method | Domains | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|
| | Res. | Tut. | Acad. | Guid. | Broch. | Admin. | Fin. | |
| GPT-4o-mini | 35.5 | **44.0** | 24.6 | 33.1 | 29.5 | **46.8** | 31.1 | 33.5 |
| LightRAG | **40.8** | 34.1 | **36.2** | **39.4** | **41.0** | 44.4 | 38.3 | **38.9** |
| MMGraphRAG | **40.8** | 36.5 | 35.7 | 35.8 | 28.2 | **46.9** | **38.5** | 37.7 |
| RAGAnything | **46.6** | **43.5** | **38.7** | **43.9** | **34.0** | 45.7 | **43.6** | **42.8** |

They ensure our evaluation captures both breadth across domains and depth in document complexity. Detailed dataset statistics and characteristics are provided in Appendix A.1.

**Baselines**. We compare RAG-Anything against the following methods for performance evaluation:

- **GPT-4o-mini**: A powerful multimodal language model with native text and image understanding capabilities. Its 128K token context window enables direct processing of entire documents. We evaluate this model as a strong baseline for long-context multimodal understanding.
- **LightRAG** (Guo et al., 2024): A graph-enhanced RAG system that integrates structured knowledge representation with dual-level retrieval mechanisms. It captures both fine-grained entity-relation information and broader semantic context, improving retrieval precision and response coherence.
- **MMGraphRAG** (Wan & Yu, 2025): A multimodal retrieval framework that constructs unified knowledge graphs spanning textual and visual content. This method employs spectral clustering for multimodal entity analysis and retrieves context along reasoning paths to guide generation.

**Experimental Settings**. In our experiments, we implement all baselines using GPT-4o-mini as the backbone LLM. Documents are parsed using MinerU (Wang et al., 2024) to extract text, images, tables, and equations for downstream RAG processing. For the retrieval pipeline, we employ the `text-embedding-3-large` model with 3072-dimensional embeddings. We use the `bge-reranker-v2-m3` model for reranking. For graph-based RAG methods, we enforce a combined entity-and-relation token limit of 20,000 tokens and a chunk token limit of 12,000 tokens. Outputs are constrained to a one-sentence format. For the baseline GPT-4o-mini in our QA scenario, documents are concatenated into image form with a maximum of 50 pages per document, rendered at 144 dpi. Finally, all query results are evaluated for accuracy by `GPT-4o-mini`.

### 3.2 Performance Comparison

**Superior Performance and Cross-Domain Generalization.** RAG-Anything demonstrates superior overall performance over baselines through its unified multimodal framework. Unlike LightRAG, which is restricted to text-only content processing, RAG-Anything treats text, images, tables, and

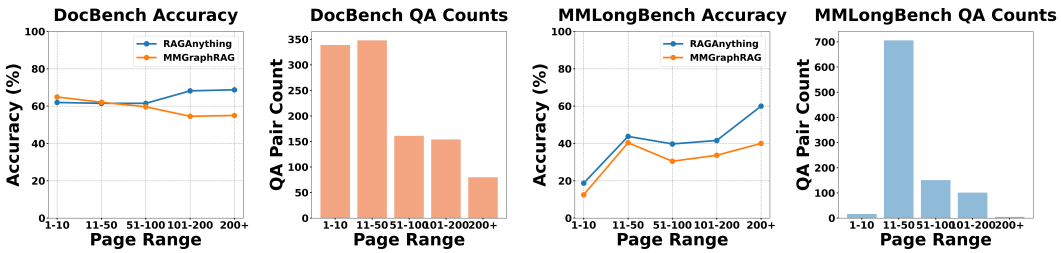


Figure 2: Performance evaluation across documents of varying lengths.

Table 4: Ablation study results on DocBench. The "Chunk-only" variant bypasses dual-graph construction and relies solely on traditional chunk-based retrieval, while "w/o Reranker" eliminates cross-modal reranking but preserves the core graph-based architecture.

| Method | Domains | | | | | Types | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | Aca. | Fin. | Gov. | Law. | News | Txt. | Mm. | Una. | |
| Chunk-only | 55.8 | 61.5 | **60.1** | **60.7** | 64.0 | 81.6 | 66.2 | 43.5 | 60.0 |
| w/o Reranker | **60.9** | **63.5** | 58.8 | **60.2** | **68.6** | **81.7** | **74.7** | **45.4** | **62.4** |
| RAGAnything | **61.4** | **67.0** | **61.5** | **60.2** | **66.3** | **85.0** | **76.3** | **46.0** | **63.4** |

equations as first-class entities. MMGraphRAG only adds basic image processing while treating tables and equations as plain text, missing crucial structural information. RAG-Anything introduces a comprehensive dual-graph construction strategy that preserves structural relationships across all modalities. This unified approach enables superior performance across both evaluation benchmarks.

**Enhanced Long-Context Performance**. RAG-Anything demonstrates superior performance on long-context documents where relevant evidence is dispersed across modalities and sections, achieving the best results in information-dense domains such as Research Reports and Financial Reports on MMLongBench. These improvements stem from a structured context injection mechanism that integrates dual-graph construction for cross-page entity alignment, combines semantic retrieval with structural navigation, and applies modality-aware processing for efficient context-window utilization. As illustrated in Figure 2, RAG-Anything and MMGraphRAG perform comparably on short documents, but RAG-Anything's advantage grows with length: on DocBench the gap exceeds 13 points for documents over 100 pages (68.2% vs. 54.6% for 101–200 pages; 68.8% vs. 55.0% for 200+ pages), and on MMLongBench RAG-Anything improves accuracy by 3.4 / 9.3 / 7.9 points on 11–50 / 51–100 / 101–200 page documents. These findings confirm that dual-graph construction and cross-modal hybrid retrieval are particularly effective for long-document reasoning.

### 3.3 Architectural Validation with Ablation Studies

To isolate and quantify the contributions of key architectural components in RAG-Anything, we conducted systematic ablation studies examining two critical design choices. Given that our approach fundamentally differs from existing methods through dual-graph construction and hybrid retrieval, we specifically evaluated: i) **Chunk-only**, which bypasses graph construction entirely and relies solely on traditional chunk-based retrieval, and ii) **w/o Reranker**, which eliminates the cross-modal reranking component while preserving the core graph-based architecture.

As demonstrated in Table 4, the results validate our architectural design through striking performance variations. • **Graph Construction is Essential.** The chunk-only variant achieves merely 60.0% accuracy with substantial cross-domain drops. This demonstrates that traditional chunking fails to capture structural and cross-modal relationships essential for multimodal documents. • **Reranking Provides Marginal Gains.** Removing the reranker yields only a modest decline to 62.4%, while the full model achieves 63.4% accuracy. This indicates that cross-modal reranking provides valuable refinement, but primary gains stem from our graph-based retrieval and cross-modal integration.

### 3.4 Case Studies

Beyond aggregate metrics, we further dissect RAG-Anything's behavior through representative cases drawn from DocBench. These illustrate how preserving *intra-modal* structure—panel↔caption↔axis edges in figures and header↔cell↔unit edges in tables—enables fine-grained, modality-specific grounding that text-flattening baselines (including MMGraphRAG, which models only image entities) cannot achieve. Detailed walkthroughs covering multi-panel figure interpretation, financial table navigation, visual reasoning over bar plots, and disambiguation of complex tabular structures are provided in Appendix A.2.

## 4 Related Work

• **Graph-Enhanced Retrieval-Augmented Generation**. Graph-based RAG explores graph construction and knowledge aggregation (Zhang et al., 2025; Bei et al., 2025). GraphRAG organizes entity relations and community summaries for query-focused generation (Edge et al., 2024), while LightRAG combines graph and vector indexing with dual-level retrieval (Guo et al., 2024). GNN-RAG (Mavromatis & Karypis, 2024) and GFM-RAG (Luo et al., 2025) explore graph-aware neural retrieval, and HippoRAG (Jimenez Gutierrez et al., 2024) uses graph structure as memory. Hierarchical and community-based approaches, including RAPTOR (Sarthi et al., 2024), Youtu-GraphRAG (Dong et al., 2025), and ArchRAG (Wang et al., 2025), organize information at multiple levels to support retrieval and aggregation. These directions establish relational and hierarchical representations as complementary ways to structure external knowledge. RAG-Anything builds on structured retrieval by integrating text-derived and cross-modal knowledge graphs, representing text, figures, tables, and equations as interconnected document content.

• **Multimodal Retrieval-Augmented Generation**. Multimodal RAG retrieves external knowledge from heterogeneous sources to support multimodal response generation (Abootorabi et al., 2025). VideoRAG (Ren et al., 2025) combines visual and textual channels for video understanding, while MM-VID (Lin et al., 2023) represents video content through textual scripts. For document understanding, VisRAG (Yu et al., 2025) embeds document pages as images and retrieves them for a vision-language generator, preserving page layouts. ColPali (Faysse et al., 2025) represents page images with multiple vectors and uses late interaction for query–page matching. VDocRAG (Tanaka et al., 2025) trains a retriever and generator over image-based document representations. These approaches emphasize page-level visual evidence; ColPali studies page retrieval, while VisRAG and VDocRAG also examine downstream answer generation. RAG-Anything uses parsed content units for element-level indexing and restores relevant visual content during generation. The principal design choices concern the representation used for indexing, the granularity of retrieved evidence, and the organization of context for generation. Page-image systems retain visual structure directly, whereas content-unit systems can associate individual elements with textual descriptions and relations. These choices offer distinct ways to organize evidence within visually rich documents.

Multimodal knowledge graphs provide another approach to organizing heterogeneous evidence. MMGraphRAG (Wan & Yu, 2025) derives scene graphs from visual content, connects them with a textual knowledge graph, and retrieves reasoning paths. MegaRAG (Hsiao et al., 2026) incorporates visual cues into knowledge-graph construction, retrieval, and generation for long-form content. RAG-Anything combines cross-modal and text-derived graph information with vector retrieval, using document elements as its content units. UniversalRAG (Yeo et al., 2026) selects among modality-specific corpora and retrieval granularities for each query. Its corpus-level routing complements the intra-document evidence integration studied here.

## 5 Conclusion

RAG-Anything provides a unified graph-based framework for multimodal retrieval. Its dual-graph construction integrates cross-modal and text-based knowledge graphs, treating multimodal content as interconnected entities rather than forcing it into text-centric pipelines, while the hybrid retrieval mechanism combines structural navigation with semantic matching to deliver superior performance on long-context documents ($>$100 pages) in the reported evaluations. Limitations and future directions are discussed in Appendix A.5.

## AI USE STATEMENT

Generative AI tools assisted with language editing, interpretation of experimental results, and identifying and summarizing related literature. The authors take responsibility for verifying the citations, claims, and final text.

## REFERENCES


Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025.

Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, Jiajun Bu, Shirui Pan, Yizhou Yu, et al. Graphs meet ai agents: Taxonomy, progress, and future opportunities. *arXiv preprint arXiv:2506.18019*, 2025.

Junnan Dong, Siyu An, Yifei Yu, Qian-Wen Zhang, Linhao Luo, Xiao Huang, Yunsheng Wu, Di Yin, and Xing Sun. Youtu-graphrag: Vertically unified agents for graph retrieval-augmented complex reasoning. *arXiv preprint arXiv:2508.19855*, 2025.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Chi-Hsiang Hsiao, Yi-Cheng Wang, Tzung-Sheng Lin, Yi-Ren Yeh, and Chu-song Chen. Megarag: Multimodal knowledge graph-based retrieval augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 48031–48059, 2026. doi: 10.18653/v1/2026.acl-long.2218. URL https://aclanthology.org/2026.acl-long.2218/.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *NeurIPS*, 37:59532–59569, 2024.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, Ce Liu, and Lijuan Wang. Mm-vid: Advancing video understanding with gpt-4v(ision). *arXiv preprint arXiv:2310.19773*, 2023.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. Gfm-rag: graph foundation model for retrieval augmented generation. *arXiv preprint arXiv:2502.01113*, 2025.

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

Costas Mavromatis and George Karypis. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*, 2024.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.

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

Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24827–24837, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tanaka_VDocRAG_Retrieval-Augmented_Generation_over_Visually-Rich_Documents_CVPR_2025_paper.html.

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

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. Archrag: Attributed community-based hierarchical retrieval-augmented generation. *arXiv preprint arXiv:2502.09891*, 2025.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3843–3871, 2026. doi: 10.18653/v1/2026.acl-long.177. URL https://aclanthology.org/2026.acl-long.177/.

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

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Yi Chang, and Xiao Huang. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. Docbench: A benchmark for evaluating llm-based document reading systems. *arXiv preprint arXiv:2407.10701*, 2024.

# A APPENDIX

This appendix provides comprehensive supporting materials for our experimental evaluation and implementation details. Section A.1 presents detailed dataset statistics for the DocBench and MMLongBench multi-modal benchmarks, including document type distributions and complexity metrics. Section A.2 showcases additional case studies that demonstrate RAG-Anything's structure-aware capabilities across diverse multimodal content understanding tasks. Section A.3 documents the complete set of multimodal analysis prompts for vision, table, and equation processing that enable context-aware interpretation. Section A.4 provides the standardized accuracy evaluation prompt used for consistent response assessment across all experimental conditions.

## A.1 DATASET CHARACTERISTICS AND STATISTICS

Table 5: Document type distribution and statistics for the DocBench benchmark.

| **Type** | Acad. | Fin. | Gov. | Law. | News |
|---|---|---|---|---|---|
| # Docs | 49 | 40 | 44 | 46 | 50 |
| # Questions | 303 | 288 | 148 | 191 | 172 |
| Avg. Pages | 11 | 192 | 69 | 58 | 1 |

Table 6: Document type distribution and statistics for the MMLongBench benchmark.

| **Type** | Res. | Tut. | Acad. | Guid. | Broch. | Admin. | Fin. |
|---|---|---|---|---|---|---|---|
| # Docs | 34 | 17 | 26 | 22 | 15 | 10 | 11 |
| # Questions | 292 | 138 | 199 | 155 | 100 | 81 | 117 |
| Avg. Pages | 39 | 58 | 35 | 78 | 30 | 17 | 87 |

Tables 5 and 6 present the distribution of document types across the DocBench and MMLong-Bench benchmarks. • **DocBench** encompasses medium- to long-length documents spanning various domains, including legal, governmental, and financial files. Financial reports represent the most extensive category, averaging 192 pages per document, while the News category consists of concise single-page newspapers. • **MMLongBench** demonstrates a broader spectrum of document types and lengths. Research reports, tutorials, and academic papers maintain moderate lengths of 35–58 pages on average, while guidebooks extend to approximately 78 pages. Brochures and administrative files remain relatively compact, whereas financial reports again emerge as the longest category.

Collectively, these two benchmarks provide comprehensive coverage ranging from brief news articles to extensive technical and financial documentation. This establishes diverse and challenging evaluation contexts for multimodal document understanding tasks.

## A.2 ADDITIONAL CASE STUDIES

Multimodal documents contain rich structural information within each modality. Understanding these *intra-modal* structures is crucial for accurate reasoning. We analyze four representative cases from DocBench to demonstrate how RAG-Anything leverages these structures. These cases highlight a key limitation of existing methods. Baselines either rely on superficial textual cues or flatten complex visual elements into plain text. In contrast, RAG-Anything builds modality-aware graphs that preserve essential relationships (*e.g.*, table header↔cell↔unit edges; panel↔caption↔axis edges). This enables precise reasoning over complex document layouts.

• **Case 1: Multi-panel Figure Interpretation**. This case examines a common scenario in academic literature. Researchers often need to compare results across different experimental conditions. These results are typically presented in multi-panel visualizations. Figure 3 shows a challenging t-SNE visualization with multiple subpanels. The query requires distinguishing between two related but distinct panels. RAG-Anything constructs a visual-layout graph where panels, axis titles, legends, and captions become nodes. Key edges encode semantic relationships. Panels contain specific plots.

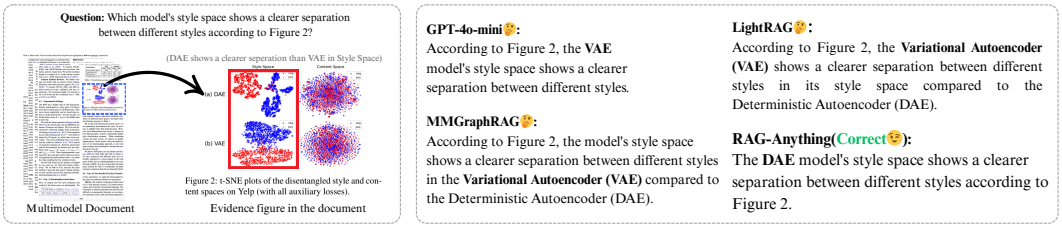


Figure 3: Multi-panel figure interpretation case. The query requires identifying cluster separation patterns from the style-space panel, while avoiding confusion from the adjacent content-space panel.

Captions provide contextual information. Subfigures relate hierarchically. This structure guides the retriever to focus on the *style-space* panel for comparing cluster separation patterns. The system avoids confusion from the adjacent content space panel. This panel shows less clear distinctions.

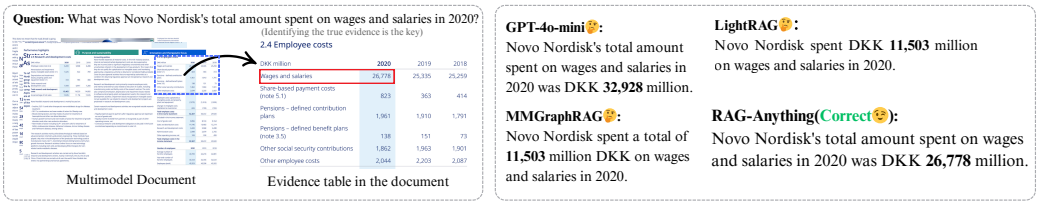


Figure 4: Financial table navigation case. The query involves locating the specific intersection of "Wages and salaries" row and "2020" column amid similar terminological entries.

• **Case 2: Financial Table Navigation**. This case addresses a common challenge in financial document analysis. Analysts must extract specific metrics from tables with similar terminology and multiple time periods. Figure 4 shows this scenario. The query involves resolving ambiguous financial terms and selecting the correct column for a specified year.

RAG-Anything transforms the financial report table into a structured graph. Each row header, column header (year), data cell, and unit becomes a node. The edges capture key relationships: *row-of*, *column-of*, *header-applies-to*, and *unit-of*. This structure enables precise navigation. The retriever focuses on the row "Wages and salaries" and the column for "2020". It directs attention to the target cell (26,778 million). The system successfully disambiguates nearby entries like "Share-based payments." Competing methods treat tables as linear text. They often confuse numerical spans and years. This leads to significantly inaccurate answers. RAG-Anything explicitly models relationships within the table. It achieves precise selection and numeric grounding. This ensures accurate responses.

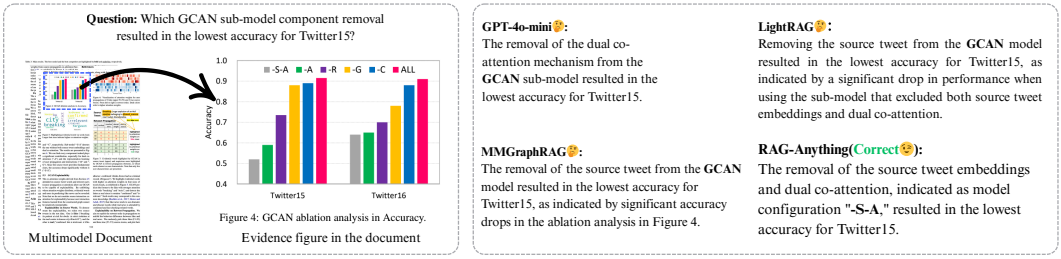


Figure 5: Visual reasoning case. RAG-Anything correctly identifies "-S-A" as the lowest accuracy configuration, while baselines misinterpret spatial relationships.

• **Case 3: Demonstrating Visual Reasoning Capabilities**. Figure 5 illustrates how RAG-Anything handles complex visual reasoning tasks involving chart interpretation. The query asks which GCAN sub-model component removal yields the lowest accuracy on Twitter15. Traditional approaches struggle with spatial relationships between visual elements. RAG-Anything addresses this challenge by constructing a structured graph representation of the bar plot. Bars, axis labels, and legends become interconnected nodes. These are linked by semantic relations such as *bar-of* and *label-applies-to*.

This graph-based approach enables precise alignment between visual and textual elements. The system correctly identifies the bar labeled "-S-A" (removing source tweet embeddings and dual

co-attention) and its corresponding accuracy value as the lowest performer. Baseline methods that flatten visual information often misinterpret spatial relationships. They frequently conflate nearby components. RAG-Anything's structured representation preserves critical visual-textual associations. This leads to accurate query resolution and proper attribution of performance drops to "-S-A".

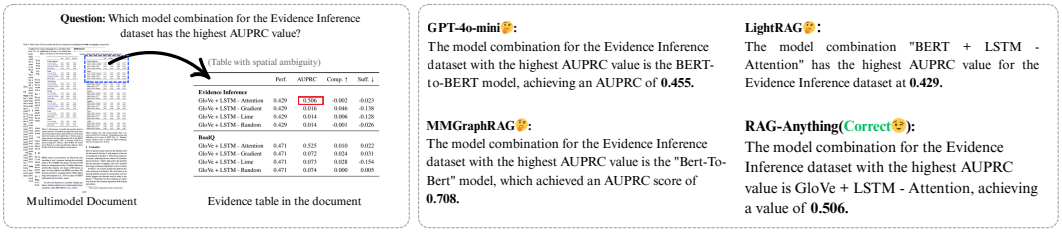


Figure 6: Tabular navigation case. RAG-Anything locates the highest AUPRC value (0.506), while the compared approaches struggle with structural ambiguity.

• **Case 4: Handling Complex Tabular Structures**. Figure 6 showcases RAG-Anything's ability to navigate intricate tabular data where structural disambiguation is crucial. The query seeks the model combination achieving the highest AUPRC value for the Evidence Inference dataset—a task complicated by repeated row labels across multiple datasets within the same table. This scenario highlights a fundamental limitation of conventional approaches that struggle with structural ambiguity in data.

RAG-Anything overcomes this by parsing the table into a comprehensive relational graph where headers and data cells become nodes connected through explicit *row-of* and *column-of* relationships. This structured representation enables the system to correctly isolate the Evidence Inference dataset context and identify "GloVe + LSTM – Attention" with a score of 0.506 as the optimal configuration. By explicitly preserving hierarchical table constraints that other methods often collapse or misinterpret, RAG-Anything ensures reliable reasoning across complex multi-dataset tabular structures.

**Key Insights.** These cases collectively demonstrate how RAG-Anything's structure-aware design delivers targeted advantages. Our approach transforms documents into explicit graph representations that capture intra-modal relationships traditional methods miss. In figures, connections between panels, captions, and axes enable panel-level comparisons rather than keyword matching. In tables, row–column–unit graphs ensure accurate identification through explicit modeling. This structure-aware retrieval reduces confusion from repeated terminology and complex layouts. Even MMGraphRAG fails on these scenarios because it only models image-modality entities while ignoring entities from tables (cells, row headers, column headers) and other modalities. RAG-Anything's comprehensive graph representation captures all modality-specific entities and their relationships, enabling precise grounding for document Q&A tasks requiring fine-grained localization.

### A.3 Context-Aware Multimodal Prompting

The following three prompts orchestrate structured, context-aware multimodal analysis with JSON-formatted outputs, systematically guiding the model to extract comprehensive descriptions of visual, tabular, and mathematical content while maintaining explicit alignment with surrounding information:

- **Vision Analysis Prompt** (Figure 7) directs the model to systematically capture compositional elements, object relationships, visual attributes, stylistic features, dynamic actions, and technical components (e.g., charts), while establishing explicit connections to accompanying text—enabling contextually-grounded interpretations beyond superficial description.
- **Table Analysis Prompt** (Figure 8) structures systematic tabular content decomposition across multiple dimensions—structural organization, column semantics, critical values, statistical patterns, and contextual relevance—with precise terminology and numerical accuracy requirements that ensure faithful preservation of key indicators.
- **Equation Analysis Prompt** (Figure 9) prioritizes semantic interpretation over syntactic restatement, instructing analysis of variable definitions, operational logic, theoretical foundations, inter-formula relationships, and practical applications so that mathematical content becomes integral to the broader argumentative framework.

### Vision Analysis Prompt

vision_analysis_prompt.png

```
Please analyze this image in detail, considering the surrounding context. Provide a JSON response with the
following structure:

{
"detailed_description": "A comprehensive and detailed visual description of the image following these
guidelines:
- Describe the overall composition and layout
- Identify all objects, people, text, and visual elements
- Explain relationships between elements and how they relate to the surrounding context
- Note colors, lighting, and visual style
- Describe any actions or activities shown
- Include technical details if relevant (charts, diagrams, etc.)
- Reference connections to the surrounding content when relevant
- Always use specific names instead of pronouns",
"entity_info": {
"entity_name": "{entity_name}",
"entity_type": "image",
"summary": "concise summary of the image content, its significance, and relationship to surrounding content
(max 100 words)"
}
}

Context from surrounding content:
{context}

Image details:
- Image Path: {image_path}
- Captions: {captions}
- Footnotes: {footnotes}

Focus on providing accurate, detailed visual analysis that incorporates the context and would be useful for
knowledge retrieval.
```

Figure 7: Vision analysis prompt for context-aware image interpretation and knowledge extraction.

### Table Analysis Prompt

table_analysis_prompt.png

```
Please analyze this table content considering the surrounding context, and provide a JSON response with the
following structure:

{
"detailed_description": "A comprehensive analysis of the table including:
- Table structure and organization
- Column headers and their meanings
- Key data points and patterns
- Statistical insights and trends
- Relationships between data elements
- Significance of the data presented in relation to surrounding context
- How the table supports or illustrates concepts from the surrounding content
Always use specific names and values instead of general references.",
"entity_info": {
"entity_name": "{entity_name}",
"entity_type": "table",
"summary": "concise summary of the table's purpose, key findings, and relationship to surrounding content (max
100 words)"
}
}

Context from surrounding content:
{context}

Table Information:
Image Path: {table_img_path}
Caption: {table_caption}
Body: {table_body}
Footnotes: {table_footnote}

Focus on extracting meaningful insights and relationships from the tabular data in the context of the
surrounding content.
```

Figure 8: Table analysis prompt for structured content decomposition and semantic understanding.

Equation Analysis Prompt

equation_analysis_prompt.png

```
Please analyze this mathematical equation considering the surrounding context, and provide a JSON response
with the following structure:

{
"detailed_description": "A comprehensive analysis of the equation including:
- Mathematical meaning and interpretation
- Variables and their definitions in the context of surrounding content
- Mathematical operations and functions used
- Application domain and context based on surrounding material
- Physical or theoretical significance
- Relationship to other mathematical concepts mentioned in the context
- Practical applications or use cases
- How the equation relates to the broader discussion or framework
Always use specific mathematical terminology.",
"entity_info": {
"entity_name": "{entity_name}",
"entity_type": "equation",
"summary": "concise summary of the equation's purpose, significance, and role in the surrounding context (max
100 words)"
}
}

Context from surrounding content:
{context}

Equation Information:
Equation: {equation_text}
Format: {equation_format}

Focus on providing mathematical insights and explaining the equation's significance within the broader
context.
```

Figure 9: Equation analysis prompt for mathematical expression interpretation and integration.

### A.4 Accuracy Evaluation Prompt Design

Figure 10 presents the standardized prompt specifically designed for systematic factual accuracy assessment of generated responses across multiple domains. The prompt establishes explicit evaluation criteria that prioritize content correctness over stylistic considerations, producing binary accuracy classifications accompanied by concise analytical justifications. All accuracy evaluations throughout our comprehensive experimental framework were conducted using **GPT-4o-mini**, ensuring consistent and reliable assessment standards across diverse question categories and specialized domains.

### A.5 Challenges and Future Directions for Multi-modal RAG

While current multimodal RAG systems demonstrate promising capabilities, their limitations emerge most clearly through systematic analysis of failure cases. Understanding where and why these systems break down is crucial for advancing the field beyond current performance plateaus. Examining failure patterns helps identify fundamental architectural bottlenecks and design principles for more robust multimodal systems. Our investigation reveals two critical failure patterns exposing deeper systemic issues in multimodal RAG architectures. These patterns are not merely edge cases but reflect fundamental challenges in cross-modal information integration and structural reasoning:

• **Text-Centric Retrieval Bias**: Systems exhibit strong preference for textual sources, even when queries explicitly demand visual information. This reveals inadequate cross-modal attention.

• **Document Structure Processing Challenges**: Systems struggle with complex layouts and non-linear information flows. This exposes limitations in spatial reasoning and contextual understanding.

These failure modes illuminate key insights about current multimodal AI. They provide concrete directions for architectural innovations that could substantially improve system robustness.

**Case 1: Cross-Modal Misalignment**. Figure 11 presents a particularly revealing failure scenario where all evaluated methods consistently produce incorrect answers despite having access to the necessary information. This universal failure across different architectures suggests fundamental limitations in how current systems handle noisy, heterogeneous multimodal data—a critical challenge as real-world applications inevitably involve imperfect, inconsistent information sources. The failure exposes two interconnected systemic issues that compound each other:

Accuracy Evaluation Prompt

accuracy_evaluation_prompt.png

```
You are an expert evaluator tasked with assessing the accuracy of answers generated by a RAG
(Retrieval-Augmented Generation) system.

**Task**: Evaluate whether the generated answer correctly responds to the given question based on the expected
answer.

**Question**: {question}

**Expected Answer**: {expected_answer}

**Generated Answer**: {generated_answer}


**Evaluation Criteria**:
1. **Accuracy (0 or 1)**: Does the generated answer match the factual content of the expected answer?
- 1: The generated answer is factually correct and aligns with the expected answer
- 0: The generated answer is factually incorrect or contradicts the expected answer

**Instructions**:
- Focus on factual correctness, not writing style or format
- Consider partial matches: if the generated answer contains the correct information but includes additional
context, it should still be considered accurate
- For numerical answers, check if the values match or are equivalent
- For list answers, check if all key elements are present
- If the expected answer is "Not answerable" and the generated answer indicates inability to answer, consider
it accurate

**Output Format**:
Please respond with a JSON object containing only:
{
"accuracy": 0 or 1,
"reasoning": "Brief explanation of your evaluation"
}
```

Figure 10: Accuracy evaluation prompt for consistent factual assessment across question types.

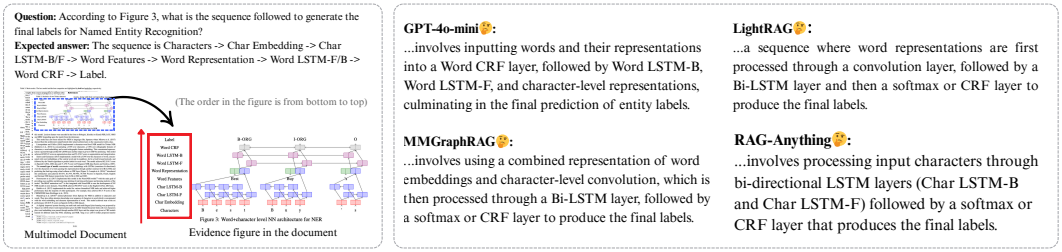


Figure 11: Cross-modal noise case. All methods fail to retrieve the correct answer from the specified image, instead retrieving noisy textual evidence that misaligns with the structured visual content.

**Issue 1: Retrieval Bias Toward Text**. Current RAG systems demonstrate pronounced bias toward textual passages. This occurs particularly when visual content lacks exact keyword matches. The bias persists even when queries contain explicit instructions to prioritize visual sources. This reveals a fundamental weakness in cross-modal attention mechanisms.

The retrieved textual information, while topically related, often operates at a different granularity level than visual content. Images may contain precise, structured data such as specific numerical values, detailed diagrams, or exact spatial relationships. Corresponding text typically provides general, conceptual descriptions. This semantic misalignment introduces noise that actively misleads the reasoning process. The system attempts to reconcile incompatible levels of detail and specificity.

**Issue 2: Rigid Spatial Processing Patterns**. Current visual processing models exhibit fundamental rigidity in spatial interpretation. Most systems default to sequential scanning patterns—top-to-bottom and left-to-right—that mirror natural reading conventions. While effective for simple text documents, this approach creates systematic failures with structurally complex real-world content. Many documents require non-conventional processing strategies. Tables demand column-wise interpretation, technical diagrams follow specific directional flows, and scientific figures embed critical information in unexpectedly positioned annotations. These structural variations are prevalent in professional documents, making adaptive spatial reasoning essential.

In the observed failure case, the correct answer required integrating visual elements in reverse order from the model's default processing sequence. The system's inability to recognize and adapt to this structural requirement led to systematic misinterpretation. This represents a fundamental architectural limitation where spatial reasoning remains static regardless of document context or query intent. When spatial processing patterns are misaligned with document structure, the extracted information becomes not merely incomplete but actively misleading. This structural noise compounds other processing errors and can lead to confident but entirely incorrect conclusions.

**Case 2: Structural Noise in Ambiguous Table Layouts**. As shown in Figure 12, all methods failed when confronted with a structurally ambiguous table. The primary failure stems from the table's confusing design: the GEM row lacks dedicated cell boundaries, and the "Joint" and "Slot" columns merge without clear separation. These structural irregularities create parsing ambiguities that systematically mislead extraction algorithms. This failure pattern reveals a critical vulnerability in current RAG systems. When table structures deviate from standard formatting conventions—through merged cells, unclear boundaries, or non-standard layouts—extraction methods consistently misinterpret cell relationships and conflate distinct data values. This exposes the brittleness of current approaches when faced with real-world document variations that deviate from clean, structured formats.

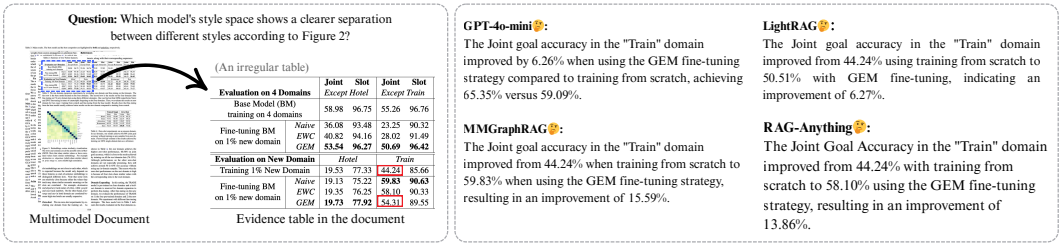


Figure 12: Ambiguous table structure case. All methods fail to correctly parse the confusing table layout with merged cells and unclear column boundaries, leading to incorrect data extraction.

The case highlights two essential directions for enhancing robustness. RAG systems require layout-aware parsing mechanisms that can recognize and adapt to structural irregularities rather than imposing rigid formatting assumptions. Additionally, integrating visual processing capabilities could significantly improve noise resilience, as visual models can leverage spatial relationships and contextual design cues that are lost in purely structural representations.