1 need for multimodal RAG capabilities. In **Scientific Research**, experimental results are primarily communicated through plots, diagrams, and statistical visualizations. These contain core discoveries that remain invisible to text-only systems. **Financial Analysis** relies heavily on market charts, correlation matrices, and performance tables. Investment insights are encoded in visual patterns rather than textual descriptions. Additionally, **Medical Literature Analysis** depends on radiological images, diagnostic charts, and clinical data tables. These contain life-critical information essential for accurate diagnosis and treatment decisions. Current RAG frameworks systematically exclude these vital knowledge sources across all three scenarios. This creates fundamental gaps that render them inadequate for real-world applications requiring comprehensive information understanding. Therefore, multimodal RAG emerges as a critical advancement. It is necessary to bridge these knowledge gaps and enable truly comprehensive intelligence across all modalities of human knowledge representation.

2 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.

3 **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.

4 **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.

5 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.

6 **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