RAG-ANYTHING: ALL-IN-ONE RAG FRAMEWORK

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, Chao Huang*
The University of Hong Kong
zrguo101@hku.hk xubinrencs@gmail.com chaohuang75@gmail.com

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://github.com/HKUDS/RAG-Anything.

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 document-intensive domains where multimodal content carries essential meaning. Academic research, financial analysis, and technical documentation represent prime examples of knowledge-rich environments. These domains fundamentally depend on visual and structured information. Critical insights are often encoded exclusively in non-textual formats. Such formats resist meaningful conversion to plain text.

The consequences of this limitation become particularly severe in knowledge-intensive domains where multimodal content carries essential meaning. Three representative scenarios illustrate the critical

*Corresponding Author: Chao Huang