MMA-RAG: A Survey on Multimodal Agentic Retrieval-Augmented Generation

Vladana Perlić, Stéphane Lebailly, Vadim Malvone, Van-Tam Nguyen, Pascal Urard

MMA-RAG: A Survey on Multimodal Agentic Retrieval-Augmented Generation

Vladana Perlić*[1,2], Stéphane Lebailly[1], Vadim Malvone[2], Van-Tam Nguyen[2], and Pascal Urard[1]

[1]STMicroelectronics, Grenoble, France
[2]Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

Abstract

Multimodal Agentic Retrieval-Augmented Generation (MMA-RAG) marks a significant advancement in AI, empowering large language models to integrate and reason over diverse data types, including text, images, audio, and structured data. This survey provides the first comprehensive overview of the MMA-RAG paradigm, tracing its evolution from traditional text-based RAG to sophisticated multimodal and agentic frameworks. We systematically review foundational literature, analyze key architectures and dominant design patterns, and survey applications across domains such as scientific question answering, document understanding, and healthcare. We conduct a comparative analysis of system components, evaluation benchmarks, and agentic capabilities like planning and tool use, offering a holistic view of the current landscape. Our key insights highlight how multimodal integration and autonomous agents mitigate hallucinations and enhance contextual reasoning, while also surfacing persistent challenges in cross-modal alignment, evaluation, and scalability. We conclude by outlining open research directions and practical implications for next-generation AI systems.

1 Introduction

Large language models (LLMs) represent a major milestone in artificial intelligence, demonstrating remarkable fluency and generalization across a broad range of natural language processing tasks (Zhao et al., 2025; Minaee et al., 2025; Brown et al., 2020). However, despite their significantly improved performance, LLMs still exhibit structural limitations that hinder their deployment in real-world, knowledge-intensive scenarios. These limitations include: (i) reliance on static training corpora, and thus an inability to incorporate knowledge beyond their training cutoff date (Cheng et al., 2024); (ii) a tendency to produce *hallucinated* content, i.e., plausible but factually incorrect completions (L. Huang et al., 2025; Xu et al., 2025); (iii) the absence of explicit citations or verifiable sources (Wu et al., 2024; J. Huang and Chang, 2024; Park and Choi, 2024; Byun et al., 2024; Schreieder et al., 2025); (iv) constraints imposed by finite context windows, preventing reasoning over long documents (X. Wang et al., 2024; Cao et al., 2025); and (v) lack of direct access to private or continuously updated data streams. Collectively, these factors constrain the reliability,

*Corresponding author: vladana.perlic@telecom-paris.fr

transparency, and adaptability of LLMs in domains such as medicine, law, or scientific research, where factual accuracy and temporal relevance are crucial.

To address these shortcomings, researchers proposed **retrieval-augmented generation (RAG)**, which grounds language model outputs in external knowledge sources (Lewis et al., 2021; Guu et al., 2020; Shuster et al., 2021). Instead of relying only on static model parameters, a RAG pipeline first retrieves relevant documents and then conditions generation on that retrieved context. This design improves factual accuracy, expands the effective memory of the system, enables verifiable citations, and provides access to timely or domain-specific data.

Figure 1, reproduced from Y. Huang and J. Huang (2024), illustrates this advantage. When asked: *"Which country will host the 2032 Olympics?"*, a vanilla LLM, trained only up to 2022, cannot provide the correct answer. In contrast, a RAG system retrieves up-to-date information ("Australia will host the 2032 Olympics") and incorporates it into its generation.

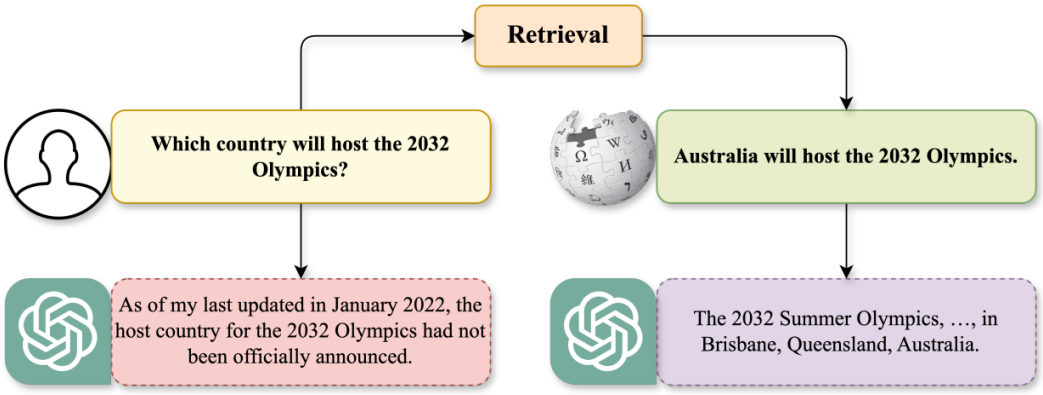

Figure 1: An example of RAG in action: by retrieving external information, the system overcomes training-data limitations and produces a correct answer (reproduced from Y. Huang and J. Huang (2024)).

To understand the architecture of retrieval-augmented generation, consider Figure 2, reproduced from Gupta et al. (2024). A retriever first queries an external knowledge base, ranks candidate documents, and forwards the most relevant evidence to the generator. The generator — most often a large language model[1] — then synthesizes this retrieved evidence into an answer. This modular pipeline ensures that responses are not only fluent but also accurate, current, and contextually informed. Crucially, it enables updating an LLM's knowledge without retraining the base model.

As the field has evolved, two complementary axes have crystallized: multimodal integration — grounding across text, images, tables, audio, and video—and agentic reasoning—autonomous planning, tool use, and iterative workflows. While powerful in isolation, these axes address different limitations of traditional RAG. Multimodal RAG enriches context but often lacks dynamic reasoning, whereas Agentic RAG introduces autonomy but is typically confined to textual data. Their convergence, which we term Multimodal Agentic RAG (MMA-RAG), enables systems that can plan, retrieve, and synthesize heterogeneous evidence in a coordinated manner. This survey positions MMA-RAG within the broader RAG landscape and provides a structured account of its capabilities, evaluation practices, and open challenges.

[1]While other sequence-to-sequence architectures are theoretically possible, the dominance of LLMs in natural language generation makes them the standard choice in RAG systems.

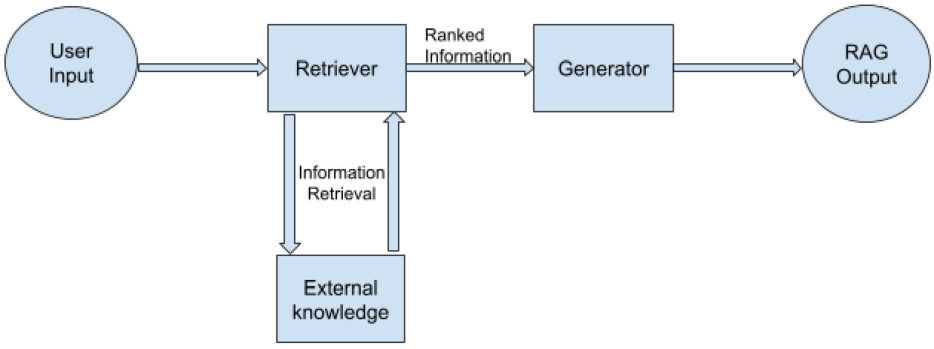

Figure 2: Core workflow of a retrieval-augmented generation (RAG) system: user input is routed through a retriever, which queries external knowledge and provides results to the generator (reproduced from Gupta et al. (2024)).

Our Contributions This survey makes the following contributions:

- Introduces a unified taxonomy that differentiates MMA-RAG from prior paradigms;
- Provides a comparative analysis of representative frameworks and architectures;
- Dissects the roles of multimodal integration and agentic reasoning within MMA-RAG;
- Compiles current benchmarks and evaluation protocols;
- Surfaces challenges distinctive to multimodal-agentic retrieval and reasoning;
- Outlines future directions toward more scalable and interpretable MMA-RAG systems.

Paper Organization The remainder of the paper proceeds as follows. Section 2 reviews the evolution of RAG paradigms from naïve text-only pipelines to multimodal and agentic variants and formalizes the taxonomy used throughout. Section 3 positions MMA-RAG within this taxonomy and details core architectural patterns, representative frameworks, and fusion/alignment techniques. Section 4 surveys application domains (document understanding, sports analytics, healthcare, scientific exploration, trust-aware vision classification, and embodied AI). Section 5 presents evaluation datasets, metrics, and comparative performance. Section 6 analyzes open challenges in cross-modal alignment, scalability, and evaluation standardization. Section 7 outlines future directions on robustness, interactivity, and integration of emerging modalities. Section 8 discusses the limitations of this survey. Section 9 concludes by synthesizing insights and highlighting the trajectory of MMA-RAG.

2 Architectures and Methodologies

Many applications need systems that work across different kinds of data—text, images, audio, video, and structured/graph data—and that stay current, cite their sources, handle long contexts, and (when authorized) access private or live feeds (Abootorabi et al., 2025; Mei et al., 2025; Liu et al., 2025). As discussed in the Introduction, RAG retrieves relevant material first and then conditions generation on that retrieved context. This improves factual accuracy, expands the system's effective memory, enables verifiable citations, and gives access to timely or domain-specific data.

Multimodal Agentic RAG (MMA-RAG) applies the same idea across modalities. Specialized retrievers gather text, images, tables, or audio/video from vector stores, graph knowledge bases, and web/API endpoints, and a coordinator plans, checks, and fuses these signals into a single,

source-backed answer (Liu et al., 2025). Agentic capabilities—planning, reflection, verification, and iterative tool use—help the system adapt as questions evolve or new evidence arrives (Singh et al., 2025; Abootorabi et al., 2025; Mei et al., 2025). In practice, hierarchical multi-agent frameworks split complex questions into sub-tasks handled by modality-specific agents, while a coordinator synthesizes evidence across sources—an approach that shows strong gains on multimodal benchmarks (Liu et al., 2025). Figure 3 illustrates this contrast, highlighting how multi-agent multimodal RAG architectures extend beyond single-agent single-modal designs. These properties are especially important in healthcare, robotics, and scientific research, where multi-step reasoning, traceable sources, and up-to-date knowledge are required.

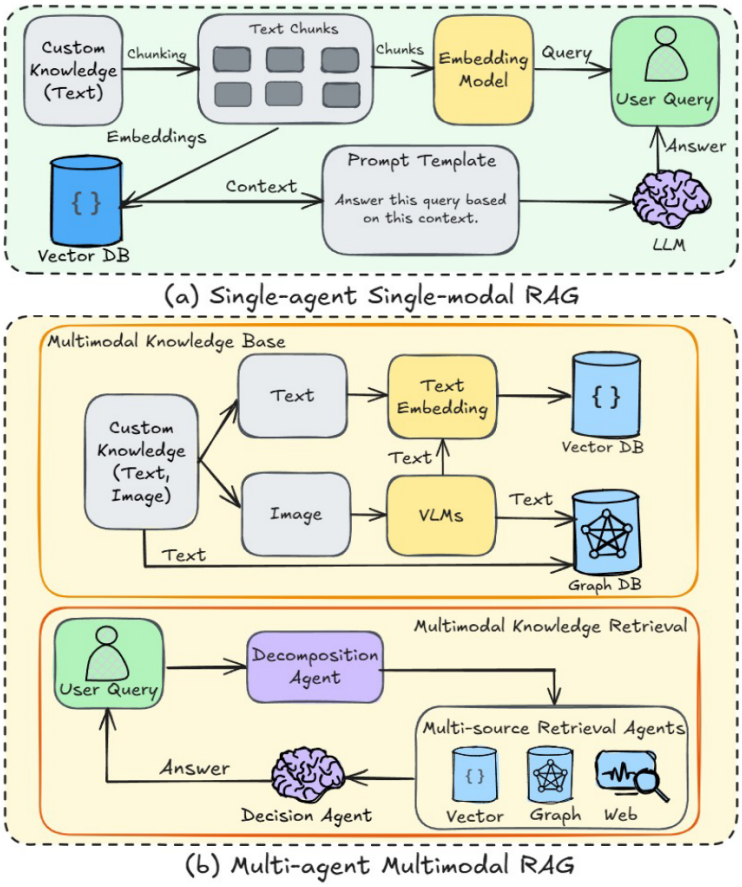

Figure 3: Comparison of (a) single-agent single-modal RAG and (b) multi-agent multimodal RAG. The latter integrates vector/graph/web sources across modalities under agentic orchestration, enabling complex, grounded answers (figure reproduced from Liu et al. (2025)).

2.1 Positioning Within the RAG Taxonomy

We follow the taxonomy introduced earlier (RAG, Multimodal RAG, Agentic RAG, MMA-RAG) and focus here on architecture-level differences rather than repeating the full historical evolution. To clarify the relationships and architectural distinctions among RAG paradigms, Figure 4 presents an improved taxonomy. This diagram highlights the evolution from naïve and advanced RAG to modular, agentic, and multimodal variants, culminating in MMA-RAG. It also emphasizes that

Graph RAG is an orthogonal enhancement, applicable across paradigms, and organizes the taxonomy along two complementary axes: modality (text-only vs. multimodal) and agency (static vs. agentic).

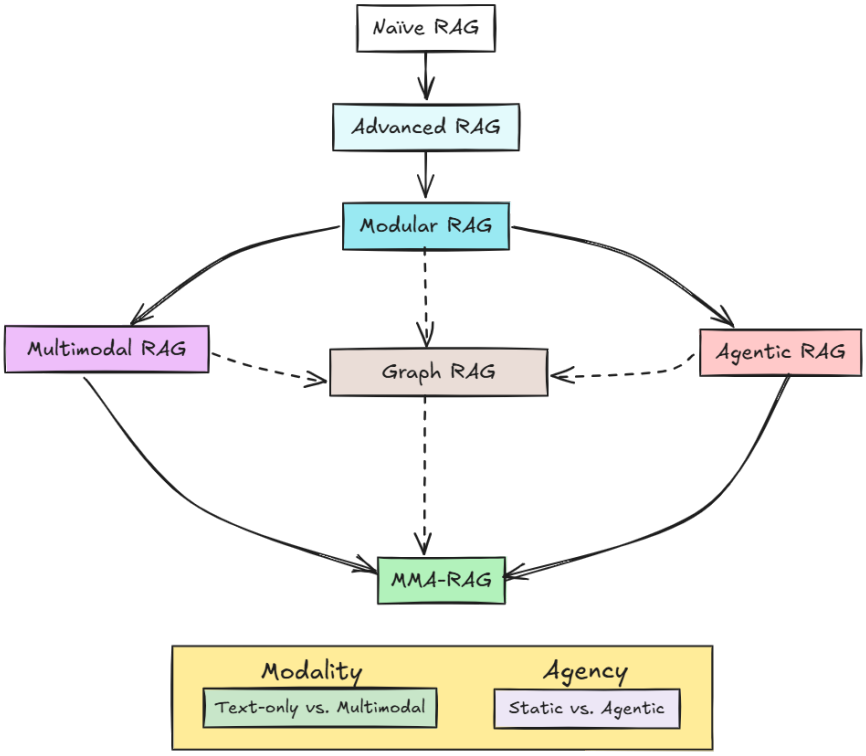

Figure 4: Architecture-centered taxonomy of RAG paradigms. This diagram illustrates the evolution and relationships among retrieval-augmented generation (RAG) paradigms, highlighting key branches: Naïve RAG, Advanced RAG, Modular RAG, Agentic RAG, Multimodal RAG, and their convergence in Multimodal Agentic RAG (MMA-RAG). Graph RAG is shown as an orthogonal enhancement that can be layered onto any paradigm to support relational and multi-hop reasoning. The taxonomy is organized along two complementary axes—modality (text-only vs. multimodal) and agency (static vs. agentic)—emphasizing architectural capabilities such as cross-modal retrieval, orchestration, and agentic planning that culminate in MMA-RAG.

- Relative to text-only RAG, Multimodal RAG adds cross-modal retrieval and fusion so the system can ground answers in images, tables, audio/video, and text.
- Relative to Multimodal RAG, agentic designs add planning and self-management of retrieval/generation workflows (Singh et al., 2025).
- MMA-RAG combines both: (a) specialized retrievers and agents per modality, (b) orchestration that breaks down the query and calls tools/APIs as needed, and (c) answers that surface citations (and optionally confidence signals).
- Graph-enhanced RAG is orthogonal and can be layered onto any of these to support relational and multi-hop reasoning (H. Han et al., 2025).

This survey focuses on MMA-RAG, providing a unified taxonomy and comprehensive analysis of this rapidly developing paradigm.

2.2 Core MMA-RAG Architectural Patterns

MMA-RAG frameworks are characterized by several key architectural and methodological innovations:

- **Hierarchical Multi-Agent Organization:** Systems are structured as a collection of specialized agents, each responsible for a particular modality (e.g., text, image, video) or function (e.g., query decomposition, retrieval, synthesis). These agents operate both independently and collaboratively, orchestrated by higher-level agents that plan and coordinate the overall workflow. For example, HM-RAG (Liu et al., 2025) and MDocAgent (S. Han et al., 2025) both use hierarchical agent structures to decompose and solve complex queries, while Yi et al. (2025) design a modular multi-agent workflow with five specialized agents for retrieval, drafting, refinement, visual analysis, and synthesis that mirrors the stepwise clinical reasoning process in radiology report generation.
- **Cross-Modal Retrieval and Fusion:** Modality-specific retrieval agents access heterogeneous data sources (vector DBs, graph KBs, web APIs) and produce intermediate representations that are aligned and fused by decision or synthesis agents. In HM-RAG (Liu et al., 2025), hierarchical fusion is achieved through multi-level attention and voting mechanisms, while MDocAgent (S. Han et al., 2025) uses schema-guided augmentation to ensure semantic consistency.
- **Agentic Capabilities:** Agents are endowed with planning, reflection, and tool-use abilities. They can decompose complex queries, iteratively refine their outputs, invoke external tools or APIs, and adapt their strategies based on intermediate results or feedback. CollEX (Schneider et al., 2025), for instance, leverages tool-use and reflection for interactive scientific exploration.
- **Dynamic Workflow Adaptation:** The agentic design allows the system to dynamically adjust retrieval and reasoning paths in response to evolving queries, ambiguous information, or user feedback, rather than following a static pipeline. GridMind (Chipka et al., 2025) demonstrates this by adapting to real-time sports data streams.
- **Iterative Refinement and Self-Correction:** A dominant pattern where the system is not single-pass but incorporates feedback to improve its output. This is often implemented as a loop of generation, evaluation, and revision. For example, CAL-RAG uses a *grader agent* to score a proposed layout and a *feedback agent* to provide corrective instructions (Forouzandehmehr et al., 2025), while other systems use a re-evaluation loop triggered by low trust scores to correct for agent overconfidence (Roumeliotis et al., 2025). The agentic RAG system in Thakrar et al. (2025) includes a self-reflection agent that triggers re-analysis when confidence is low, enabling autonomous error correction.
- **Trust-Aware Orchestration with Calibration and Re-evaluation.** Modular pipelines can separate perception (vision agents) from meta-reasoning (an orchestrator) and use calibrated trust signals to coordinate agents. The orchestrator in Roumeliotis et al. (2025) employs metrics such as Expected Calibration Error (ECE), Overconfidence Ratio (OCR), and Confidence-Correctness Correlation (CCC), and leverages CLIP-based (Radford et al., 2021) image retrieval with iterative re-evaluation to correct overconfidence, yielding substantial accuracy gains in zero-shot visual diagnosis.
- **Constraint- and Content-Aware Layout Generation.** CAL-RAG (Forouzandehmehr et al., 2025) retrieves relevant layout exemplars from a structured knowledge base, uses an LLM-based layout recommender, a vision–language grader agent, and a feedback agent to iteratively improve placements. Implemented with LangGraph, it achieves SOTA on PKU

PosterLayout across underlay effectiveness, element alignment, and overlap, outperforming LayoutPrompter (Lin et al., 2023).

- **Non-Parametric Embodied Memory.** Embodied-RAG (Xie et al., 2025) augments embodied agents with a general non-parametric memory structured as a *semantic forest* of language descriptions at multiple granularities, enabling retrieval for navigation and explanation queries across kilometer-scale environments.
- **Domain-Specific Workflow Emulation:** Instead of relying on generic agent roles, some frameworks design their multi-agent collaborations to deliberately mirror the workflows of human experts. In medical VQA, for example, this can take the form of emulating peer consultation through multi-modal reasoning layers, as well as conducting reference checks using agentic RAG (Thakrar et al., 2025). Similarly, in radiology report generation, the stepwise process of clinical reasoning provides an implicit mechanism for aligning and integrating information in a way that remains clinically meaningful (Yi et al., 2025). Together, these architectural choices aim to reproduce the collaborative, evidence-driven practices at the core of medical diagnosis, positioning workflow emulation as a compelling alternative to extensive fine-tuning.

2.3 Dominant Agentic Design Patterns

Our analysis of the literature reveals that as MMA-RAG systems tackle more complex tasks, their agentic structures are not arbitrary. Instead, they converge around a few dominant design patterns. These patterns represent recurring architectural solutions to core challenges in task decomposition, quality control, and domain-specific reasoning. Across the MMA-RAG literature, several dominant agentic design patterns have emerged, each tailored to address specific types of problems, as illustrated in Figure 5. These patterns are not mutually exclusive but represent distinct architectural philosophies:

- **Hierarchical Orchestration.** In this pattern, a top-level agent (or orchestrator) decomposes a complex query and delegates sub-tasks to specialized, subordinate agents. This is effective for multi-faceted queries requiring diverse data sources. For example, `HM-RAG` employs a Decomposition Agent to split queries and a Decision Agent to synthesize results from vector, graph, and web retrieval agents (Liu et al., 2025). Similarly, the framework in Roumeliotis et al. (2025) uses a non-visual orchestrator for meta-reasoning, separating it from the perceptual tasks of its vision agents.
- **Iterative Refinement and Feedback Loops.** This pattern involves agents that generate, critique, and revise solutions in a loop. It is particularly powerful for tasks where the quality of the output is subjective or requires optimization against multiple constraints. `CAL-RAG`, for instance, uses a 'grader' and 'feedback' agent to iteratively improve creative layouts (Forouzandehmehr et al., 2025). Other systems use self-reflection agents triggered by low confidence scores to autonomously correct errors, emulating a process of reconsideration (Thakrar et al., 2025; Roumeliotis et al., 2025).
- **Domain-Specific Workflow Emulation.** Here, the multi-agent collaboration is explicitly designed to mimic the workflow of human experts in a specific domain. This provides a natural structure for reasoning and ensures the final output is contextually relevant and interpretable to professionals. The five-agent pipeline in Yi et al. (2025) mirrors the stepwise process of a radiologist generating a report, while the system in Thakrar et al. (2025) emulates clinical peer consultation and reference-checking. The 'Radiologist' and 'Medical Writer' agents in

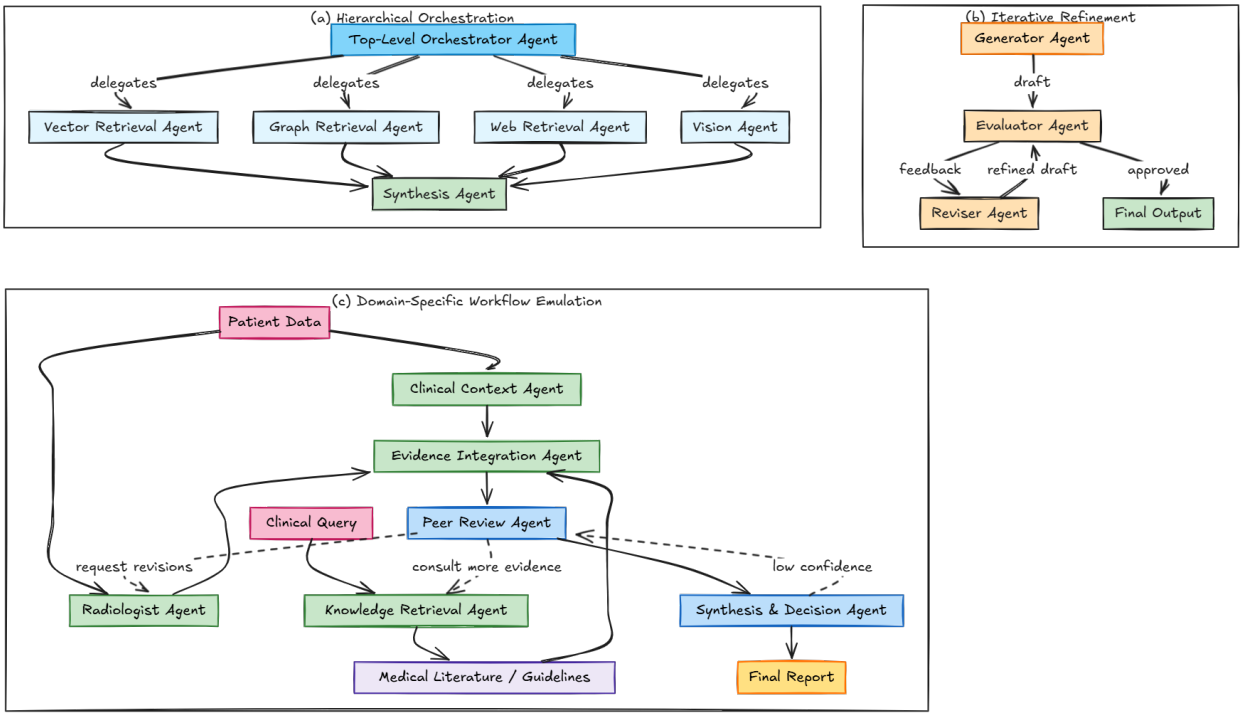

Figure 5: Dominant agentic design patterns in MMA-RAG frameworks. (a) **Hierarchical Orchestration**, where a top-level agent delegates tasks. (b) **Iterative Refinement**, where agents collaborate in a feedback loop. (c) **Domain-Specific Workflow Emulation**, where agent collaboration mimics human expert processes.

`IRR-with-CBM-RAG` are another clear example of this pattern (H. M. T. Alam et al., 2025).

2.4 Representative MMA-RAG Frameworks

Table 1 summarizes leading MMA-RAG frameworks, highlighting their application domains, key contributions, agentic features, challenges addressed, and notable results. These frameworks exemplify the integration of both multimodal and agentic capabilities.

Table 1: Comparison of Multimodal Agentic Retrieval-Augmented Generation Frameworks

Framework	Application Domain	Key Contributions	Agentic Features	Challenges Addressed	Notable Results
HM-RAG (Liu et al., 2025)	Multimodal, multi-source question answering and knowledge synthesis	Hierarchical multi-agent system with query decomposition, modality-specific retrieval, and answer integration	Planning, multi-agent collaboration, reflection	Multimodal reasoning, cross-modal alignment, modularity	12.95% higher answer accuracy, 3.56% better question classification on ScienceQA and CrisisMMD; state-of-the-art in zero-shot settings
MDocAgent (S. Han et al., 2025)	Document Understanding	Role-based multi-agent system for text, image, and table processing with collaborative reasoning	Planning, tool use, agent specialization	Long-form document analysis, visual layout integration	12.1% average improvement on five benchmarks including MMLongBench, LongDocURL, PaperTab, PaperText, and FetaTab
GridMind (Chipka et al., 2025)	Sports Analytics	Multi-agent system for unified analysis of video, text, statistics, and commentary	Dynamic adaptation, multi-agent collaboration	Cross-modal correlation, real-time data integration	Effective NFL data analysis (58% accuracy in a 10-participant closed alpha, 17.5s avg. response)
CollEX (Schneider et al., 2025)	Scientific Exploration, Education	Multimodal agentic retrieval-augmented generation with vision-language agents for interactive exploration	Tool use, reflection, interactive planning	Intuitive access, interdisciplinary discovery, visual data integration	Proof-of-concept handles 64,000+ records across 32 collections (no formal evaluation)
IRR-with-CBM-RAG (H. M. T. Alam et al., 2025)	Healthcare (Radiology)	Multi-agent retrieval-augmented generation for interpretable radiology reports using concept bottleneck models	Multi-agent collaboration, concept-level explanation, intervention	Factual consistency, hallucination, cross-modal misalignment, clinical interpretability	Higher classification accuracy (81% vs. 47–78% baseline), high interpretability, more clinically realistic reports; multi-agent setup improves usefulness and correctness (up to 0.96)

Continued on next page

Framework	Application Domain	Key Contributions	Agentic Features	Challenges Addressed	Notable Results
Multimodal Multi-Agent Framework for RRG (Yi et al., 2025)	Healthcare (Radiology)	Modular multi-agent system for radiology report generation, with five specialized agents emulating clinical workflow	Agent specialization, sequential collaboration, modular workflow, retrieval grounding	Factual accuracy, clinical relevance, multimodal integration, modularity	Substantial improvements over single-agent baseline on IU X-ray: BLEU 0.0036→0.0466, ROUGE-1 0.2398→0.3652, ROUGE-2 0.0278→0.1292, ROUGE-L 0.1537→0.2471, METEOR 0.1437→0.3618, BERTScore 0.8617→0.8819; better diagnostic accuracy, style, conciseness, and key finding coverage
CAL-RAG (Forouzandehmehr et al., 2025)	Content-aware layout/UI design	Retrieval-augmented, multi-agent layout generation with LLM recommender, VLM grader, and feedback agent (LangGraph)	Planning, collaborative refinement, tool use	Content–layout coupling, structural validity, visual coherence	SOTA on PKU PosterLayout (Hsu et al., 2023) (underlay effectiveness, alignment, overlap); outperforms LayoutPrompter (Lin et al., 2023)
Orchestrator-Agent Trust (Roumeliotis et al., 2025)	Visual classification (apple leaf disease)	Modular framework separating perception agents from a non-visual orchestrator with RAG-based reasoning and CLIP retrieval (Radford et al., 2021)	Calibration-aware orchestration, iterative re-evaluation	Reliability in zero-shot, overconfidence correction, interpretability	Achieved 85.63% accuracy, a 77.94% relative improvement over the zero-shot baseline, demonstrating that trust-aware orchestration can approach the performance of fine-tuned models without retraining.

Continued on next page

Framework	Application Domain	Key Contributions	Agentic Features	Challenges Addressed	Notable Results
Clinical Collaboration for MM-VQA (Thakrar et al., 2025)	Healthcare (dermatology telemedicine)	Clinical-inspired multi-model/agentic collaboration via peer-consultation-style reasoning and RAG with medical literature	Collaborative reasoning, retrieval grounding	Generalization, explainability, literature-grounded answers	Achieved up to 70% accuracy, maintaining performance on unseen data where fine-tuning catastrophically failed (degrading 4 of 7 models by an average of 30%).
Embodied-RAG (Xie et al., 2025)	Embodied AI & robotics	A hierarchical non-parametric memory (*semantic forest*) that automatically structures multi-modal embodied experiences for efficient, multi-level retrieval	Hierarchical retrieval over the semantic forest, guided by LLM-based selection	Scalable semantic memory, multimodal data correlation, and hierarchical abstraction for embodied agents	Successfully handles $>$ 250 navigation/explanation queries across kilometer-scale environments

2.5 Fusion and Alignment Techniques in MMA-RAG

A defining challenge for MMA-RAG systems is the alignment and fusion of evidence from heterogeneous modalities. State-of-the-art frameworks employ:

- **Cross-Modal Embedding:** Mapping visual, textual, and structured data into a shared latent space (e.g., via BLIP-2 (Li et al., 2023) or similar models) to enable joint retrieval and reasoning. For example, CollEX (Schneider et al., 2025) uses vision-language models for embedding and retrieval.
- **Hierarchical Fusion:** Employing multi-level attention or voting mechanisms to integrate evidence at both the agent and system level. HM-RAG (Liu et al., 2025) implements hierarchical fusion to combine outputs from modality-specific agents.
- **Critical Information Extraction**: Employing a specialized agent to identify and extract the most salient textual and visual information from a retrieved context. MDocAgent (S. Han et al., 2025) leverages a critical agent to focus downstream analysis on the most relevant evidence.
- **Hybrid Similarity Metrics for Clustering:** For embodied data, which is inherently spatial and multimodal, fusion can begin at the memory construction phase. Embodied-RAG builds its semantic forest by hierarchically clustering nodes using a hybrid distance metric that combines spatial proximity (Haversine distance) and semantic similarity (cosine similarity of text embeddings), effectively fusing location and perception (Xie et al., 2025).
- **Retrieval-Grounded Re-evaluation:** To align agent confidence with visual evidence, an orchestrator can trigger a re-evaluation loop. This process uses CLIP-based (Radford et al., 2021) image retrieval to fetch visually similar precedents, providing external grounding that helps an agent correct an initial, overconfident prediction (Roumeliotis et al., 2025).
- **Constraint-Aware Iterative Grading.** For layout generation, agents must respect structural constraints (alignment, non-overlap) while staying content-aware. CAL-RAG integrates a vision–language grader and feedback agent to iteratively optimize these constraints (Forouzandehmehr et al., 2025).
- **Hierarchical Episodic Indices.** Embodied-RAG indexes multimodal episodes as a semantic forest of language descriptions, enabling cross-granularity retrieval for both navigation and explanation (Xie et al., 2025).

These methodologies are critical for ensuring that agentic workflows can effectively synthesize and reason over heterogeneous evidence, a hallmark of advanced MMA-RAG systems.

3 The Evolution of RAG Paradigms

RAG has advanced significantly to meet the needs of more complex real-world applications requiring scalability, multi-step reasoning, and multimodal grounding. Surveys by Y. Huang and J. Huang (2024), Gupta et al. (2024), and Singh et al. (2025) provide comprehensive perspectives on this trajectory, which we summarize in seven paradigms:

1. **Naïve RAG:** The foundational implementation, coupling simple keyword-based retrieval (e.g., TF-IDF, BM25) with a generator. Easy to implement and effective for fact-based queries, but limited by lack of semantic awareness, fragmented outputs, and poor scalability.
2. **Advanced RAG:** Introduced dense retrieval (e.g., DPR), neural re-ranking, and multi-hop retrieval. Improved semantic alignment and contextual reasoning, though at higher computational cost.
3. **Modular RAG:** Decomposed the retrieval–generation pipeline into configurable components. Enabled hybrid sparse+dense retrieval, API/tool integration, and domain-specific workflows—making RAG adaptable and scalable across domains such as finance, law, and biomedicine.
4. **Graph RAG:** Augmented retrieval with graph-based structures, capturing entity–relation networks for richer relational reasoning (H. Han et al., 2025). Particularly suited for structured domains like healthcare or legal research, though dependent on high-quality graph data.
5. **Agentic RAG:** Introduced autonomy into the pipeline. Rather than static retrieval, agentic systems dynamically plan retrieval, manage workflows, and iteratively refine results (Singh et al., 2025). Effective for multi-step reasoning and real-time adaptation, but computationally costly and complex to orchestrate.
6. **Multimodal RAG:** Extended retrieval and grounding beyond text to images, tables, audio, and video (Abootorabi et al., 2025; Mei et al., 2025). Opened new application domains (e.g., visual question answering, multimedia knowledge access), but introduced challenges in cross-modal alignment and fusion.
7. **Multimodal Agentic RAG (MMA-RAG):** The emerging frontier, unifying multimodal grounding with agentic decision-making. Systems like HM-RAG (Liu et al., 2025), MDocAgent (S. Han et al., 2025), GridMind (Chipka et al., 2025), IRR-with-CBM-RAG (H. M. T. Alam et al., 2025), and CollEX (Schneider et al., 2025) demonstrate dynamic workflows where specialized agents coordinate retrieval across modalities and synthesize evidence collaboratively. More recent additions reveal a variety of agentic patterns, including hierarchical orchestrators that separate perception from meta-reasoning and use trust metrics to coordinate agents (Roumeliotis et al., 2025), collaborative frameworks that emulate clinical peer consultation (Thakrar et al., 2025), iterative feedback loops for creative tasks like layout design (Forouzandehmehr et al., 2025), and systems with non-parametric memory for navigation in embodied AI (Xie et al., 2025). While highly promising, MMA-RAG remains in its infancy, lacking standard benchmarks and taxonomies but marking the most powerful RAG evolution to date.

These paradigms are not strictly sequential or mutually exclusive; rather, they represent an expanding set of capabilities, with advanced systems often incorporating features from multiple preceding stages. A comparative overview of these paradigms is shown in Table 2, adapted and extended from Singh et al. (2025).

Building on this evolution, we distill the seven paradigms into a coarser taxonomy along two orthogonal axes—modality (text-only vs. multimodal) and agency (static vs. agentic):

- **RAG:** Naïve and advanced, text-only, non-agentic.
- **MRAG:** Multimodal RAG, non-agentic.
- **Agentic RAG:** Agent-driven autonomy, text-only.
- **MMA-RAG:** Unified multimodal+agentic.

Specialized forms such as Modular and Graph RAG can be understood as refinements within these broader categories.

Recent Advances. RAG has advanced on two complementary axes: multimodal integration and agentic

Table 2: Comparative analysis of Retrieval-Augmented Generation (RAG) paradigms. Table adapted from Singh et al. (2025), with extensions to include Multimodal RAG and Multimodal Agentic RAG (MMA-RAG).

Paradigm	**Key Features**	**Strengths**
Naïve RAG	• Keyword-based retrieval (e.g., TF-IDF, BM25) • Simple, easy to implement • Suitable for fact-based queries	• Fast and lightweight • Low infrastructure cost
Advanced RAG	• Dense retrieval (e.g., DPR) • Neural ranking and re-ranking • Multi-hop retrieval	• High retrieval precision • Improved contextual relevance
Modular RAG	• Hybrid retrieval (sparse + dense) • API/tool integration • Composable domain-specific pipelines	• Highly flexible and customizable • Suitable for diverse applications • Scalable
Graph RAG	• Graph-based structure integration • Multi-hop relational reasoning • Contextual enrichment with nodes	• Strong relational reasoning • Mitigates hallucinations • Ideal for structured data
Agentic RAG	• Autonomous retrieval-augmented agents • Dynamic decision-making & task decomposition • Iterative workflow refinement	• Adaptive in real time • High accuracy • Multi-domain scalability
Multimodal RAG	• Retrieval across text, images, video, audio, tables • Cross-modal alignment strategies • Fusion of multimodal signals	• Richer context beyond text-only • Captures semantic diversity • Useful for multimedia access
MMA-RAG	• Combines multimodality with agentic reasoning • Autonomous planning for multi-modal retrieval • Multi-agent collaboration (e.g., text/image/audio sub-agents)	• Most powerful paradigm so far • Solves complex, cross-modal tasks • Highly adaptive and collaborative

reasoning. Surveys of multimodal RAG (Abootorabi et al., 2025; Mei et al., 2025) highlight challenges in alignment and evaluation, while studies on agentic RAG (Singh et al., 2025) emphasize planning, tool use, and adaptability. Recent frameworks suggest the potential of unifying these directions in MMA-RAG, but a systematic taxonomy and consolidated analysis remain missing.

4 Applications

MMA-RAG systems are being applied across a diverse range of domains where integrating heterogeneous data is critical for robust reasoning and decision-making.

4.1 Document Understanding and Processing

Document understanding represents a critical application domain for MMA-RAG systems, as real-world documents often contain complex combinations of text, images, tables, and structured data. MDocAgent exemplifies this application by providing a sophisticated framework for document question answering that integrates multiple specialized agents (S. Han et al., 2025). Its five-agent architecture enables comprehensive document analysis, with each agent handling a specific modality or function, coordinated by a general agent.

This multi-agent approach is particularly effective for long-form documents and complex visual layouts, achieving significant performance improvements on challenging benchmarks. The framework's ability to jointly reason over textual and visual content makes it valuable for applications such as technical documentation analysis, legal document review, and academic paper comprehension.

Furthermore, CAL-RAG (Forouzandehmehr et al., 2025) demonstrates that retrieval-augmented, multi-agent iteration (LLM layout recommender, VLM grader, feedback agent) can synthesize layouts that are both content-aware and structurally valid, achieving SOTA on PKU PosterLayout across underlay effectiveness, element alignment, and overlap.

4.2 Sports Analytics and Cross-Modal Data Integration

Sports analytics is an emerging application domain where MMA-RAG systems demonstrate significant value through cross-modal data integration. GridMind exemplifies this as a multi-agent NLP framework designed for unified analysis of NFL data (Chipka et al., 2025). The system processes heterogeneous sports data including video footage, commentary transcripts, statistical records, and structured game data to provide comprehensive insights.

GridMind's multi-agent architecture enables collaborative analysis where specialized agents handle different data modalities and reasoning tasks. This approach is valuable for correlating visual events with statistical outcomes, textual commentary, and historical performance data, supporting advanced applications such as predictive modeling and strategic decision-making.

4.3 Healthcare and Medical Applications

Healthcare applications of MMA-RAG systems have shown particular promise in medical imaging and clinical decision support. In radiology, frameworks for report generation use multi-agent RAG with concept bottlenecks (H. M. T. Alam et al., 2025) or modular systems that emulate clinical workflows (Yi et al., 2025).

H. M. T. Alam et al. (2025) introduce a multi-agentic framework for radiology report generation that leverages concept bottleneck models to enhance interpretability and factual consistency. Similarly, Yi et al. (2025) propose a multimodal multi-agent system for radiology report generation, where five specialized agents (retrieval, draft generation, visual analysis, refinement, and synthesis) align with the stepwise clinical reasoning process. Both approaches address challenges such as factual inconsistency, hallucination, and cross-modal misalignment, and demonstrate improved accuracy and clinical relevance on radiology benchmarks.

These multi-agentic frameworks involve specialized agents for image analysis, clinical correlation, and report synthesis. By integrating medical images with patient histories, clinical guidelines, and domain

knowledge, these systems can generate detailed and accurate diagnostic reports while maintaining transparency in their reasoning process.

Beyond radiology, MMA-RAG is proving effective in telemedicine, where clinical context is limited. For dermatological visual question answering, (Thakrar et al., 2025) architected a system that mimics the clinical collaboration process. By using reasoning layers for "peer consultation" and agentic RAG for "reference-checking" against medical literature, the system achieved high accuracy and robustness on unseen data, in stark contrast to standard fine-tuning methods which suffered significant performance degradation. This highlights the value of MMA-RAG in creating explainable, literature-grounded outputs essential for clinical trust.

The principles of MMA-RAG also extend to general scientific and biological diagnostics. A framework developed for apple leaf disease diagnosis uses a modular design that separates perception agents from a non-visual reasoning orchestrator (Roumeliotis et al., 2025). This orchestrator uses trust calibration metrics (e.g., ECE, OCR) and triggers a RAG-based re-evaluation loop to correct agent overconfidence. The approach significantly improved zero-shot accuracy, demonstrating how trust-aware, agentic systems can deliver reliable and interpretable reasoning in high-stakes diagnostic domains beyond human medicine.

4.4 Scientific Exploration

Scientific research often involves navigating vast, complex collections of multimodal data, including textual documents, images, and experimental records. CollEX (Schneider et al., 2025) is a state-of-the-art MMA-RAG system designed to facilitate interactive, curiosity-driven exploration of large scientific collections. By leveraging large vision-language models (LVLMs) as multimodal agents accessible through an intuitive chat interface, CollEX enables learners, educators, and researchers to independently explore diverse scientific records, fostering scientific excitement and discovery.

4.5 Vision Classification with Trust-Aware Orchestration

The orchestrator-agent trust framework (Roumeliotis et al., 2025) separates perception agents from a calibration-aware orchestrator that uses Expected Calibration Error (ECE), Overconfidence Ratio (OCR), and Confidence-Correctness Correlation (CCC), plus CLIP-based (Radford et al., 2021) image retrieval with iterative re-evaluation. On apple leaf disease diagnosis it reaches 85.63% accuracy and yields a +77.94% improvement in zero-shot via trust-aware orchestration and RAG, correcting overconfidence and supporting interpretable decisions.

4.6 Embodied AI and Robotics

Embodied-RAG (Xie et al., 2025) introduces a non-parametric, hierarchical memory (semantic forest) enabling agents to retrieve prior multimodal experience for navigation and explanation across kilometer-scale environments, successfully handling over 250 queries and illustrating MMA-RAG in interactive, partially observable settings.

5 Evaluation and Benchmarks

5.1 Benchmarks Used in MMA-RAG Frameworks

To assess the capabilities of Multimodal Agentic RAG (MMA-RAG) systems, recent research has adopted a range of benchmark datasets and evaluation metrics. In this survey, we present and analyze the primary benchmarks employed by the leading MMA-RAG frameworks.

- **ScienceQA** (Lu et al., 2022): A multimodal scientific question-answering benchmark with over 21,000 examples requiring joint reasoning over text and images. *Used by **HM-RAG** (Liu et al., 2025)* to evaluate multimodal retrieval, reasoning, and answer accuracy.

- **CrisisMMD** (F. Alam et al., 2018; Ofli et al., 2020): Comprising approximately 35,000 social media posts from crisis events, annotated for disaster categories. *Used by **HM-RAG** (Liu et al., 2025)* to test cross-modal alignment and classification in real-world, high-stakes scenarios.
- **Document Understanding Benchmarks**: A suite of benchmarks for long-form multimodal document understanding, requiring joint processing of extended textual, visual, and tabular content. Key examples include **MMLongBench** (Z. Wang et al., 2025), **LongDocURL** (Deng et al., 2024), **PaperTab**(Hui et al., 2024), **PaperText**(Hui et al., 2024), and **FetaTab** (Hui et al., 2024). *Used by **MDocAgent** (S. Han et al., 2025)* to evaluate document analysis and reasoning.
- **COVID-QU** (Chowdhury et al., 2020): A large-scale chest X-ray dataset for COVID-19, pneumonia, and normal cases. *Used by **IRR-with-CBM-RAG** (H. M. T. Alam et al., 2025)* to evaluate both classification and report generation, with a focus on interpretability and clinical usefulness.
- **IU X-ray** (Demner-Fushman et al., 2015): A widely used dataset for radiology report generation, containing chest X-ray images paired with corresponding diagnostic reports. *Used by the **Multimodal Multi-Agent Framework for RRG** (Yi et al., 2025)* to benchmark improvements in factual accuracy, clinical relevance, and report structure.
- **PKU PosterLayout** (Hsu et al., 2023): A benchmark for visual-textual presentation layout, comprising 9,974 poster-layout pairs with annotations for text, logos, and underlays. *Used by **CAL-RAG** (Forouzandehmehr et al., 2025)* to evaluate content-aware, constraint-satisfying layout generation.
- **DermaVQA-DAS**: A dataset of 300 unique patient encounters for dermatology, including patient context, multiple images per case, and structured diagnostic questions. It reflects telemedicine complexity with variable image quality and informal language. *Used by the **Clinical Collaboration framework for MM-VQA** (Thakrar et al., 2025)* to evaluate answer accuracy and literature-grounded explainability.
- **Embodied-Experiences**: A collection of topological graphs from 14 simulated and 5 real environments, including indoor, outdoor, and large-scale street-view scenes. *Used by **Embodied-RAG** (Xie et al., 2025)* to evaluate navigation and explanation queries using its non-parametric, semantic-forest memory.

Frameworks like **CollEX** (Schneider et al., 2025) and **GridMind** (Chipka et al., 2025) do not use standard public benchmarks. **CollEX** is a proof-of-concept that currently lacks formal evaluation. **GridMind** is evaluated using internal benchmarks and a closed alpha user study. The **Orchestrator–Agent Trust** framework (Roumeliotis et al., 2025) uses a curated dataset of 800 apple leaf disease images for its visual classification task. The dataset is included on the project's GitHub page[2], but no further descriptions are provided.

5.2 Evaluation Metrics

Commonly reported metrics include accuracy, precision, recall, F1-score, BLEU, ROUGE, and BERTScore. For more nuanced assessments, many frameworks rely on an "LLM-as-a-Judge" approach to evaluate semantic accuracy and clinical usefulness. Additional domain-specific metrics include:

- **Trust and Calibration**: Expected Calibration Error (ECE), Overconfidence Ratio (OCR), and Confidence-Correctness Correlation (CCC) are used to measure an orchestrator's ability to manage agent reliability (Roumeliotis et al., 2025).
- **Layout Quality**: Metrics such as underlay effectiveness, element alignment, and overlap are used for design generation tasks on benchmarks like PKU PosterLayout (Forouzandehmehr et al., 2025).
- **Embodied Navigation**: Query-level success for navigation and explanation tasks, along with metrics like Success Weighted by Path Length (SPL), are used to evaluate the efficiency of paths generated by embodied agents (Xie et al., 2025).

[2] https://github.com/Applied-AI-Research-Lab/Orchestrator-Agent-Trust

5.3 Performance Highlights and Limitations

Across the surveyed frameworks, authors generally report significant improvements over unimodal or non-agentic baselines.

- **HM-RAG** reports a **12.95%** improvement in answer accuracy on ScienceQA and establishes a state-of-the-art zero-shot result on both ScienceQA and CrisisMMD (Liu et al., 2025).
- **MDocAgent** achieves an average improvement of **12.1%** over state-of-the-art methods on five document understanding benchmarks (S. Han et al., 2025).
- The **Orchestrator–Agent Trust** framework demonstrates that agentic reasoning can approach the performance of fine-tuned models without any retraining, achieving **85.63%** accuracy in a zero-shot setting—a **77.94%** relative improvement over its baseline (Roumeliotis et al., 2025).
- The **Clinical Collaboration for MM-VQA** framework achieves up to **70%** accuracy and maintains performance on unseen data, a scenario where fine-tuning catastrophically failed and degraded model performance by an average of **30%** (Thakrar et al., 2025).
- **CAL-RAG** reports state-of-the-art results on the PKU PosterLayout benchmark, outperforming strong baselines like LayoutPrompter in underlay effectiveness, alignment, and overlap (Forouzandehmehr et al., 2025).
- In radiology, **IRR-with-CBM-RAG** achieves **81%** classification accuracy, outperforming baselines (H. M. T. Alam et al., 2025), while the **Multimodal Multi-Agent Framework for RRG** significantly improves report quality metrics like BLEU (from 0.0036 to 0.0466) and ROUGE-L (from 0.1537 to 0.2471) (Yi et al., 2025).
- **Embodied-RAG** successfully handles over 250 navigation and explanation queries across kilometer-scale environments (Xie et al., 2025).

Despite these promising results, most evaluations are domain-specific, and there remains a lack of standardized, large-scale benchmarks that assess end-to-end agentic behavior—spanning planning, tool use, and grounded reasoning—under consistent protocols. To advance the field, we propose the development of unified, large-scale benchmarks that assess end-to-end agentic behavior, including planning, tool use, and grounded reasoning across modalities.

6 Challenges

6.1 Cross-Modal Alignment and Reasoning

A fundamental challenge in MMA-RAG systems lies in effectively aligning and reasoning across heterogeneous modalities such as text, images, and structured data. Achieving fine-grained semantic alignment remains difficult. For example, in clinical applications, a system may retrieve a relevant decision tree or image but struggle to associate specific nodes or visual cues with the corresponding textual context or patient information, leading to misinterpretations or incomplete reasoning (Yi et al., 2025; Abootorabi et al., 2025).

Moreover, multimodal reasoning requires the system to integrate complementary information from different modalities to perform compositional inference. Current approaches often rely on separate retrievers for each modality, resulting in fragmented retrieval pipelines that complicate downstream synthesis and can produce inconsistent or incoherent outputs. Hallucination and factual inconsistency remain critical concerns, as models may misinterpret visual data or generate unsupported assertions, undermining trustworthiness especially in sensitive domains like healthcare. Addressing these issues demands advances in unified embedding spaces, cross-modal attention mechanisms, and entity-aware retrieval strategies to enhance alignment and reasoning fidelity.

6.2 Scalability and Efficiency

Implementing and scaling agentic multimodal RAG systems pose significant computational and architectural challenges. The complexity of managing multiple specialized agents—each responsible for modality-specific retrieval, reasoning, and synthesis—introduces overhead in coordination, latency, and resource consumption.

Concrete bottlenecks include GPU/TPU memory constraints, inference latency, and the cost of multi-agent orchestration. Maintaining modularity while ensuring seamless integration and communication among agents requires sophisticated orchestration frameworks.

Response times can be adversely affected by iterative retrieval and multi-step reasoning workflows inherent to agentic designs, which may limit real-time applicability in latency-sensitive scenarios. Efficient indexing, caching strategies, and parallelization are critical but nontrivial to implement given the heterogeneity of data sources and modalities. Training and fine-tuning multimodal models with diverse objectives—such as cross-modal alignment, hallucination mitigation, and domain adaptation—demand substantial computational resources and careful optimization.

Multi-agent architectures like those employed in MDocAgent (S. Han et al., 2025) and GridMind (Chipka et al., 2025), while effective, introduce additional complexity in terms of inter-agent communication and coordination, potentially affecting system efficiency and requiring careful optimization of agent interaction protocols.

6.3 Evaluation Fragmentation and Lack of Standardization

A significant challenge surfaced by our review is the fragmentation of evaluation protocols. Unlike established NLP tasks, the performance of MMA-RAG systems is often measured with custom or domain-specific methodologies, making direct comparison difficult. While some frameworks are validated on established benchmarks like ScienceQA (Liu et al., 2025), many others rely on an "LLM-as-a-Judge" paradigm for qualitative assessment (Yi et al., 2025; H. M. T. Alam et al., 2025). Furthermore, some pioneering works present proof-of-concept systems with no formal evaluation (Schneider et al., 2025) or report findings from small-scale, internal user studies (Chipka et al., 2025). This methodological diversity underscores an urgent need for comprehensive, standardized benchmarks that can assess end-to-end agentic behavior—including planning, tool use, and grounded reasoning—across multiple modalities.

7 Future Directions

7.1 Robustness, Adaptability, and Interactivity

Ongoing research in MMA-RAG focuses on enhancing system robustness, adaptability, and generalizability across diverse domains. Current models face challenges such as hallucinations and sensitivity to noisy or incomplete inputs, which limit their reliability in real-world applications. Recent advances propose training strategies that improve cross-modal consistency and incorporate dynamic retrieval mechanisms that adapt to evolving user queries and data distributions.

Reinforcement learning approaches show promise for optimizing collaboration among retrieval agents and generators in dynamic environments. Chen et al. (2025) present a multi-agent reinforcement learning framework that improves response quality by learning optimal cooperation strategies during retrieval and answer synthesis.

Domain adaptation techniques and continual learning frameworks are being explored to enable models to generalize better across tasks without extensive retraining. In healthcare, adaptive MMA-RAG systems can personalize outputs by learning from user interactions and feedback, improving explainability and trustworthiness.

A promising direction is the development of interactive MMA-RAG systems, where human-in-the-loop feedback, active learning, and explainable agentic decisions are integrated to further improve reliability and transparency.

Key open questions include: How can agentic multimodal systems be evaluated holistically? What are the best practices for cross-modal fusion and alignment? How can human-in-the-loop feedback and reinforcement learning be integrated to improve reliability and transparency?

7.2 Integration of New Modalities

The modularity of current MMA-RAG architectures, such as hierarchical multi-agent frameworks, facilitates the seamless integration of new data modalities beyond traditional text and images. Future work should move beyond static datasets to incorporate dynamic, real-time data streams. This includes integrating agents capable of processing spatio-temporal data from video feeds, continuous sensor readings from IoT devices, or interactive environmental data, as hinted at by frameworks like 'Embodied-RAG' (Xie et al., 2025). Such advancements are critical for applications in autonomous robotics, environmental monitoring, and real-time decision support. This extensibility is essential for expanding system capabilities to incorporate audio, video, sensor data, and structured knowledge graphs, enabling richer contextual understanding and reasoning.

Future frameworks aim to support plug-and-play modality-specific agents that can be dynamically added or updated without disrupting existing pipelines. Advances in cross-modal embedding and fusion techniques will further enable smooth alignment and synthesis of heterogeneous data types, enhancing the system's flexibility and scalability.

This modular approach paves the way for applications in domains like autonomous driving, telemedicine, and IoT, where integrating diverse real-time data streams is crucial.

8 Limitations

While this survey aims to provide a comprehensive overview of Multimodal Agentic Retrieval-Augmented Generation (MMA-RAG), several limitations should be acknowledged regarding both our review process and the broader state of the field.

Generalizability Across Domains

Most MMA-RAG systems discussed here have been evaluated in specific domains like healthcare, scientific research, or sports analytics. It remains to be seen how well these approaches will transfer to other areas with different data types and reasoning requirements.

Scope of Literature Coverage

The MMA-RAG landscape is evolving rapidly, with new models and applications emerging frequently. Given the recency and fast pace of this field, much of the foundational and state-of-the-art work is currently available as preprints or in early-access venues, with relatively few peer-reviewed publications as of mid-2025. In this survey, we have prioritized the most recent and influential works—including preprints, arXiv submissions, and early conference papers—up to mid-2025. As a result, some industrial systems, less-publicized research, or works published after our review period may not be included.

Selection Bias and Taxonomy

The frameworks and methodologies highlighted here were selected based on their prominence and relevance to major trends. However, the field is highly diverse, and our taxonomy may not capture every possible architectural variation or hybrid approach.

Evaluation and Benchmarking

Direct comparison between MMA-RAG systems is challenging due to differences in datasets, evaluation metrics, and experimental setups. Many benchmarks are domain-specific, which may limit the generalizability of reported results. The lack of standardized, large-scale multimodal benchmarks—especially those evaluating agentic capabilities—remains a significant gap.

Rapidly Evolving Techniques

The techniques and frameworks described in this survey reflect the state of the art as of mid-2025, but the field is advancing quickly. New methods for cross-modal alignment, agentic planning, and tool use are emerging regularly, and best practices are likely to evolve.

Despite these limitations, we hope this survey provides a useful foundation for understanding current trends and open challenges in MMA-RAG, and encourages further research in this rapidly developing area.

9 Conclusion

This paper has defined and systematically analyzed the emergence of Multimodal Agentic Retrieval-Augmented Generation (MMA-RAG) as a distinct and powerful paradigm in artificial intelligence. By tracing the evolution from text-only RAG to systems that unify multimodal integration with agentic reasoning, we provide the first comprehensive taxonomy of this new frontier. Our analysis of dominant architectural patterns—hierarchical orchestration, iterative refinement, and domain-specific workflow emulation—reveals a clear and accelerating trend toward modular, collaborative, and interpretable AI.

Across applications in healthcare, scientific discovery, and embodied AI, MMA-RAG frameworks consistently outperform non-agentic or unimodal baselines. They fulfill the promise of RAG by grounding responses in verifiable, multi-source evidence, directly addressing the core LLM limitations of hallucination and static knowledge. However, as our review highlights, significant challenges in evaluation, scalability, and cross-modal alignment are critical hurdles that the community must now address to unlock the full potential of this paradigm.

Ultimately, MMA-RAG represents a fundamental shift from static, single-pass generation to dynamic, evidence-driven synthesis. The principles of agentic collaboration and multimodal grounding analyzed in this survey are not merely incremental improvements; they are foundational components for the next generation of reliable, transparent, and context-aware AI.

References

Abootorabi, M. M., A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. Soleymani Baghshah, and E. Asgari (2025). "Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation". In: arXiv: 2502.08826.

Alam, F., F. Ofli, and M. Imran (June 2018). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*.

Alam, H. M. T., D. Srivastav, M. A. Kadir, and D. Sonntag (2025). "Towards Interpretable Radiology Report Generation via Concept Bottlenecks Using a Multi-agentic RAG". In: *Advances in Information Retrieval*, pp. 201–209. DOI: 10.1007/978-3-031-88714-7_18.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, I. Sutskever, D. Amodei, et al. (2020). "Language Models are Few-Shot Learners". In: arXiv: 2005.14165.

Byun, C., P. Vasicek, and K. Seppi (June 2024). "This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance". In: *Proc. 3rd Workshop Bridging Human–Computer Interaction and Natural Language Processing*. Mexico City, Mexico: Association for Computational Linguistics, pp. 28–39. DOI: 10.18653/v1/2024.hcinlp-1.3.

Cao, B., D. Cai, and W. Lam (2025). "InfiniteICL: Breaking the Limit of Context Window Size via Long Short-term Memory Transformation". In: arXiv: 2504.01707.

Chen, Y., L. Yan, W. Sun, X. Ma, Y. Zhang, S. Wang, D. Yin, Y. Yang, and J. Mao (2025). "Improving Retrieval-Augmented Generation through Multi-Agent Reinforcement Learning". In: arXiv: 2501.15228.

Cheng, J., M. Marone, O. Weller, D. Lawrie, D. Khashabi, and B. Van Durme (2024). "Dated Data: Tracing Knowledge Cutoffs in Large Language Models". In: arXiv: 2403.12958.

Chipka, J., C. Moyer, C. Troyer, T. Fuelling, and J. Hochstedler (2025). "GridMind: A Multi-Agent NLP Framework for Unified, Cross-Modal NFL Data Insights". In: arXiv: 2504.08747.

Chowdhury, M. E. H., T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam (2020). "Can AI Help in Screening Viral and COVID-19 Pneumonia?" In: *IEEE Access* 8, pp. 132665–132676. DOI: 10.1109/ACCESS.2020.3010287.

Demner-Fushman, D., M. Kohli, M. Rosenman, S. Shooshan, L. Rodriguez, S. Antani, G. Thoma, and C. McDonald (2015). "Preparing a Collection of Radiology Examinations for Distribution and Retrieval". In: *J. Amer. Medical Informatics Assoc.* 23. DOI: 10.1093/jamia/ocv080.

Deng, C., J. Yuan, P. Bu, P. Wang, Z. Li, J. Xu, X. Li, Y. Gao, J. Song, B. Zheng, and C. Liu (2024). "LongDocURL: A Comprehensive Multimodal Long Document Benchmark Integrating Understanding, Reasoning, and Locating". In: arXiv: 2412.18424.

Forouzandehmehr, N., R. Yousefi Maragheh, S. Kollipara, K. Zhao, T. Biswas, E. Korpeoglu, and K. Achan (2025). "CAL-RAG: Retrieval-Augmented Multi-Agent Generation for Content-Aware Layout Design". In: arXiv: 2506.21934.

Gupta, S., R. Ranjan, and S. N. Singh (2024). "A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions". In: arXiv: 2410.12837.

Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang (2020). "REALM: Retrieval-Augmented Language Model Pre-Training". In: arXiv: 2002.08909.

Han, H., Y. Wang, H. Shomer, K. Guo, J. Ding, Y. Lei, M. Halappanavar, R. A. Rossi, S. Mukherjee, X. Tang, Q. He, Z. Hua, B. Long, T. Zhao, N. Shah, A. Javari, Y. Xia, and J. Tang (2025). "Retrieval-Augmented Generation with Graphs (GraphRAG)". In: arXiv: 2501.00309.

Han, S., P. Xia, R. Zhang, T. Sun, Y. Li, H. Zhu, and H. Yao (2025). "MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding". In: arXiv: 2503.13964.

Hsu, H., X. He, Y. Peng, H. Kong, and Q. Zhang (2023). "PosterLayout: A New Benchmark and Approach for Content-aware Visual-Textual Presentation Layout". In: arXiv: 2303.15937.

Huang, J. and K. C. Chang (2024). "Citation: A Key to Building Responsible and Accountable Large Language Models". In: arXiv: 2307.02185.

Huang, L., W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu (Jan. 2025). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2, pp. 1–55. DOI: 10.1145/3703155.

Huang, Y. and J. Huang (2024). "A Survey on Retrieval-Augmented Text Generation for Large Language Models". In: arXiv: 2404.10981.

Hui, Y., Y. Lu, and H. Zhang (2024). "UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis". In: arXiv: 2406.15187.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela (2021). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: arXiv: 2005.11401.

Li, J., D. Li, S. Savarese, and S. Hoi (2023). “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”. In: arXiv: 2301.12597.

Lin, J., J. Guo, S. Sun, Z. J. Yang, J. Lou, and D. Zhang (2023). “LayoutPrompter: Awaken the Design Ability of Large Language Models”. In: arXiv: 2311.06495.

Liu, P., X. Liu, R. Yao, J. Liu, S. Meng, D. Wang, and J. Ma (2025). “HM-RAG: Hierarchical Multi-Agent Multimodal Retrieval Augmented Generation”. In: arXiv: 2504.12330.

Lu, P., S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan (2022). “Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering”. In: *Proc. 36th Conf. Neural Inf. Process. Syst. (NeurIPS)*.

Mei, L., S. Mo, Z. Yang, and C. Chen (2025). “A Survey of Multimodal Retrieval-Augmented Generation”. In: arXiv: 2504.08748.

Minaee, S., T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao (2025). “Large Language Models: A Survey”. In: arXiv: 2402.06196.

Ofli, F., F. Alam, and M. Imran (May 2020). “Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response”. In: *Proc. 17th Int. Conf. Inf. Syst. Crisis Response Manag. (ISCRAM)*.

Park, B. and J. Choi (2024). “Identifying the Source of Generation for Large Language Models”. In: arXiv: 2407.12846.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: arXiv: 2103.00020.

Roumeliotis, K. I., R. Sapkota, M. Karkee, and N. D. Tselikas (2025). “Agentic AI with Orchestrator-Agent Trust: A Modular Visual Classification Framework with Trust-Aware Orchestration and RAG-Based Reasoning”. In: arXiv: 2507.10571.

Schneider, F., N. Baba Ahmadi, I. Vogel, M. Semmann, and C. Biemann (2025). “CollEX: A Multimodal Agentic RAG System Enabling Interactive Exploration of Scientific Collections”. In: arXiv: 2504.07643.

Schreieder, T., T. Schopf, and M. Färber (2025). “Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models”. In: arXiv: 2508.15396.

Shuster, K., S. Poff, M. Chen, D. Kiela, and J. Weston (2021). “Retrieval Augmentation Reduces Hallucination in Conversation”. In: arXiv: 2104.07567.

Singh, A., A. Ehtesham, S. Kumar, and T. Talaei Khoei (2025). “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG”. In: arXiv: 2501.09136.

Thakrar, K., S. Basavatia, and A. Daftardar (2025). “Architecting Clinical Collaboration: Multi-Agent Reasoning Systems for Multimodal Medical VQA”. In: arXiv: 2507.05520.

Wang, X., M. Salmani, P. Omidi, X. Ren, M. Rezagholizadeh, and A. Eshaghi (2024). “Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models”. In: arXiv: 2402.02244.

Wang, Z., W. Yu, X. Ren, J. Zhang, Y. Zhao, R. Saxena, L. Cheng, G. Wong, S. See, P. Minervini, Y. Song, and M. Steedman (2025). “MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly”. In: arXiv: 2505.10610.

Wu, K., E. Wu, A. Cassasola, A. Zhang, K. Wei, T. Nguyen, S. Riantawan, P. Shi Riantawan, D. E. Ho, and J. Zou (2024). “How well do LLMs cite relevant medical references? An evaluation framework and analyses”. In: arXiv: 2402.02008.

Xie, Q., S. Y. Min, P. Ji, Y. Yang, T. Zhang, K. Xu, A. Bajaj, R. Salakhutdinov, M. Johnson-Roberson, and Y. Bisk (2025). “Embodied-RAG: General Non-parametric Embodied Memory for Retrieval and Generation”. In: arXiv: 2409.18313.

Xu, Z., S. Jain, and M. Kankanhalli (2025). “Hallucination is Inevitable: An Innate Limitation of Large Language Models”. In: arXiv: 2401.11817.

Yi, Z., T. Xiao, and M. V. Albert (2025). “A Multimodal Multi-Agent Framework for Radiology Report Generation”. In: arXiv: 2505.09787.

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Chen, Y. Yang, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen (2025). “A Survey of Large Language Models”. In: arXiv: 2303.18223.