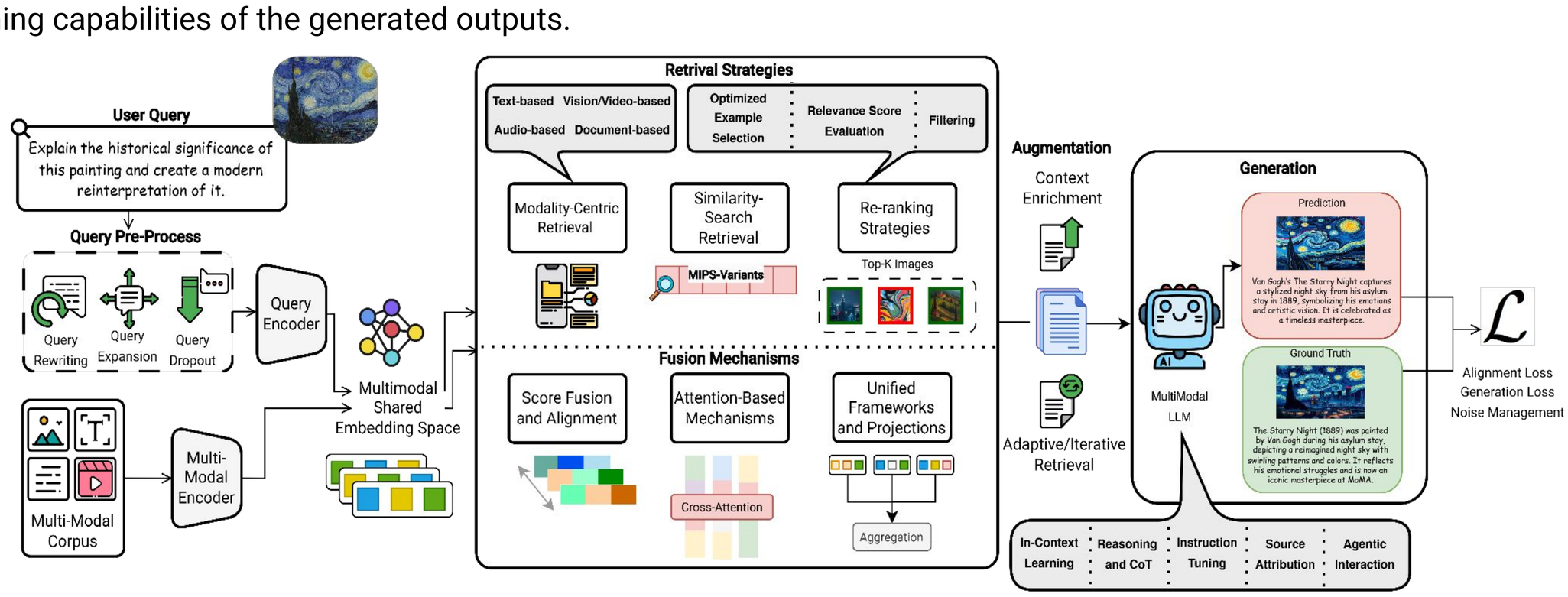
Ask in Any Modality
A Comprehensive Survey on Multimodal Retrieval-Augmented Generation

Mohammad Mahdi Abootorabi[‡◆†], Amirhosein Zobeiri[※], Mahdi Dehghani[¶], Mohammadali Mohammadkhani[§], Bardia Mohammadi[Δ], Omid Ghahroodi[†], Mahdieh Soleymani Baghshah[§,*], Ehsaneddin Asgari[†,*]

[†]Qatar Computing Research Institute, [‡]Saarland University, [◆]Zuse School ELIZA, [§]Sharif University of Technology, [※]University of Tehran, [Δ]Max Planck Institute for Software Systems, [¶]K.N. Toosi University of Technology

Multimodal RAG extends traditional Retrieval-Augmented Generation (RAG) frameworks by incorporating diverse data types, such as text, images, audio, and video, from external knowledge sources. This approach aims to address AI limitations like hallucinations and outdated knowledge through the dynamic integration of this retrieved information, thereby improving the factual grounding, accuracy, and reasoning capabilities of the generated outputs.



- **Multimodal Encoding:** User queries and corpus documents (text, image, video, etc.) are mapped into a shared semantic space using modality-specific or unified multimodal encoders (e.g., CLIP, BLIP, UniIR). This foundational process creates aligned vector embeddings, enabling direct cross-modal comparison and retrieval.
- **Retrieval Strategy:** Relevant documents are retrieved using efficient search algorithms (e.g., MIPS) combined with modality-specific strategies (e.g., text, vision, audio, video, or document layouts). Re-ranking and filtering strategies then refine candidates by optimizing relevance scores and diversity.
- **Fusion Mechanisms:** Retrieved multimodal data is integrated through score fusion for representation alignment, attention-based mechanisms that dynamically weigh cross-modal interactions, or unified frameworks that consolidate diverse inputs into a coherent representation for generation.
- **Augmentation Techniques:** Retrieved context is refined through methods like context enrichment (e.g., entity expansion), adaptive retrieval (dynamically selecting sources based on query needs), and iterative retrieval (multi-step refinement using feedback).
- **Generation:** A Multimodal Large Language Model (MLLM) generates responses conditioned on the query and augmented context, leveraging in-context learning, chain-of-thought reasoning, source attribution, and agentic generation via interactive or self-refining feedback loops.
- **Training Strategies:** Models are trained in a multi-stage process, typically involving pre-training on large paired datasets to learn cross-modal relationships, followed by task-specific fine-tuning. This often relies on contrastive losses (e.g., InfoNCE) for alignment and cross-entropy for generation tasks.
- **Robustness Management:** Model resilience is enhanced via noise-injected training to handle irrelevant data, knowledge distillation to reduce noise while preserving key signals, and regularization techniques like Query Dropout to mitigate the impact of noisy or biased inputs.

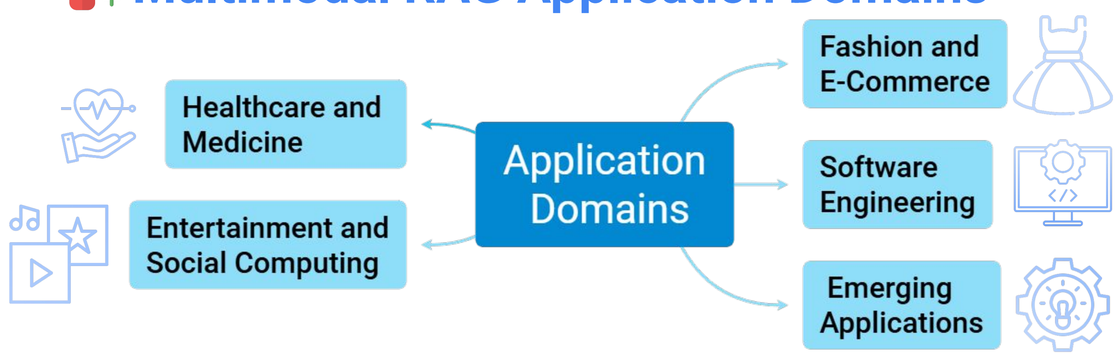
Taxonomy of Recent Advances

Multimodal RAG	Retrieval	Efficient Search and Similarity Retrieval	Maximum Inner Product Search
			Multimodal Encoders
		Modality-Centric Retrieval	Text-Centric Retrieval
			Vision-Centric Retrieval
			Video-Centric Retrieval
			Audio-Centric Retrieval
			Document Retrieval and Layout Understanding
		Re-ranking Strategies	Optimized Example Selection
			Relevance Score Evaluation
			Filtering Mechansims
	Fusion Mechanisms	Score Fusion and Alignment	
		Attention-Based Mechanisms	
		Unified Frameworks and Projections	
	Augmentation Techniques	Context Enrichment	
		Adaptive and Iterative Retrieval	
	Generation Techniques	In-Context Learning	
		Reasoning	
		Instruction Tuning	
		Source Attribution	
		Agentic Generation and Interaction	
	Training Strategies	Alignment	
		Robustness and Noise Management	

Open Problems and Future Directions

- **Robustness & Explainability:** Improve domain adaptation and adversarial robustness while mitigating modality biases and ensuring precise source attribution.
- **Advanced Reasoning & Retrieval**: Enhance compositional reasoning and entity-aware retrieval by developing unified, bias-resistant multimodal embedding spaces.
- **Agentic Frameworks & Self-Guidance:** Develop agents that use RL and other feedback mechanisms for self-assessment, dynamic modality selection, and iterative refinement.
- **Embodied & Real-World Grounding:** Integrate real-world sensor data to ground reasoning and enable context-aware applications in robotics and embodied AI.
- **Efficiency & Scalability:** Overcome long-context bottlenecks for efficient processing and scalable deployment, particularly on edge devices.
- **Personalization & Evaluation**: Advance privacy-preserving personalization while establishing more rigorous benchmarks to evaluate complex reasoning and robustness.

Multimodal RAG Application Domains



Website | Repository | Paper