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

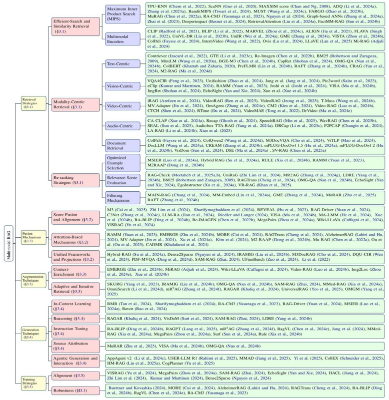
ACL 2025 VIENNA

Table of Contents

Key Contributions

ACL 2025 VIENNA

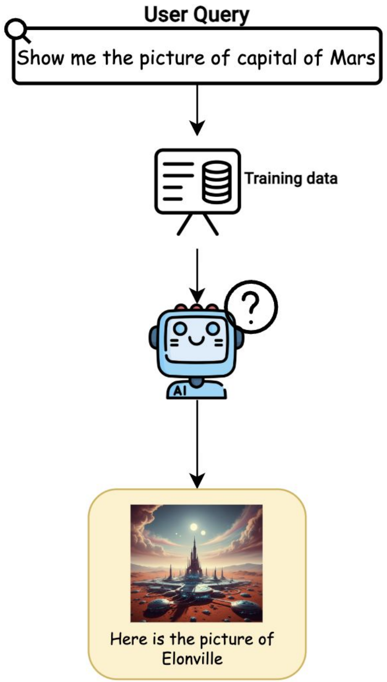
- First comprehensive survey on Multimodal RAG (100+ papers reviewed).
- Structured taxonomy covering different components and innovations.
- Open-access up to date resources (GitHub).
- Research gaps and actionable future directions.



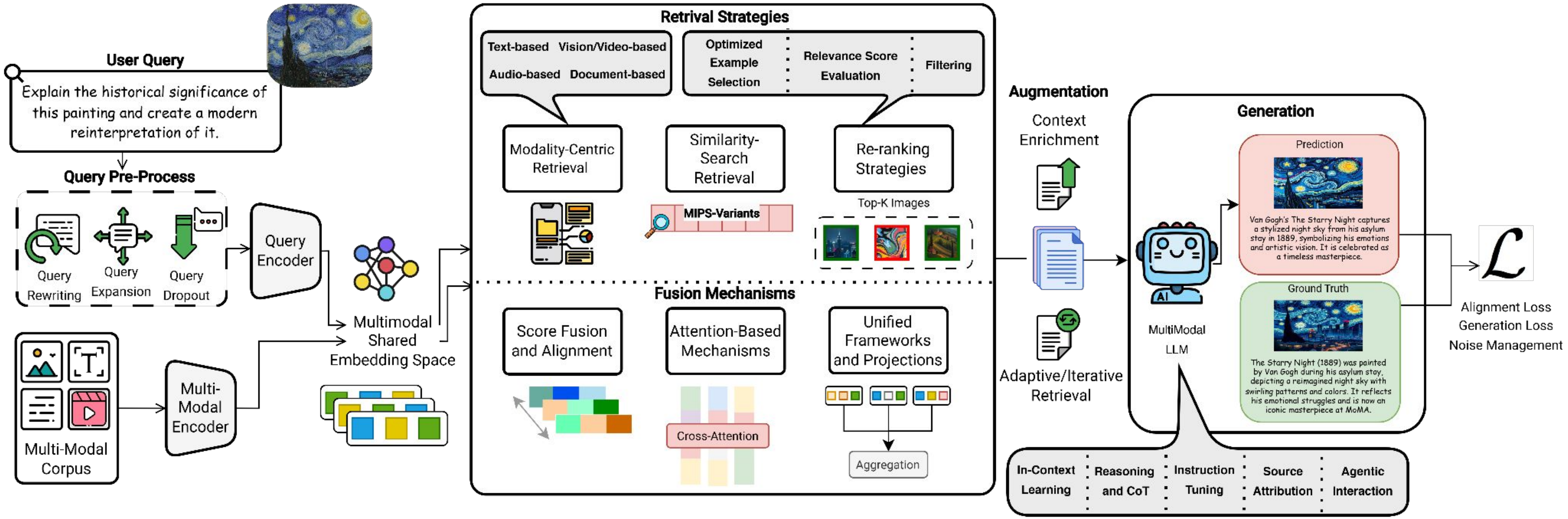
ACL 2025 VIENNA

Background & Motivation

- **Limitations of LLMs**
  - Hallucinations, outdated knowledge (static training data).
  - Poor performance in knowledge-intensive tasks.
- **RAG (Retrieval-Augmented Generation)**
  - Integrates dynamic external knowledge (Lewis et al., 2020).
  - Reduces hallucinations (Shuster et al., 2021).
- **Multimodal Learning**
  - CLIP (Radford et al., 2021) aligns vision-language.
  - Enables cross-modal reasoning (e.g., healthcare, robotics).
  - Multimodal LLMs.
- **Multimodal RAG**
  - Extends RAG to leverage multimodal data.



Introducing Multimodal RAG



Multimodal RAG Formulation

ACL 2025 VIENNA

1. The Multimodal Corpus: The Knowledge Source $D = \{d_1, d_2, ..., d_n\}$

- Each document d_i within the corpus can be of any modality (M_{d_i}), such as text, images, audio, or video, making it a rich source of information.
- Documents with multiple modalities are either broken down into single-modality parts or processed by a universal encoder.

2. Encoding into a Shared Space $z_i = Enc_{M_{d_i}}(d_i)$

- The goal is to project all modalities into a shared semantic space where they can be compared.
- collection of all encoded representations is denoted as $Z = \{z_1, z_2, ..., z_n\}$

Multimodal RAG Formulation (cont.)

ACL 2025 VIENNA

3. The Retrieval Step (R):

- A retrieval model computes a relevance score $s(e_q, z_i)$ between the encoded query (e_q) and each document embedding (z_i).
- A retrieved context (X) is created by selecting all documents that meet a relevance threshold (τ):

$$X = \{d_i \in D \mid s(e_q, z_i) \geq \tau_{M_{d_i}}\}$$

4. The Generation Step (G):

- A generative model (G) produces the final response (r), conditioned on both the original query (q) and the retrieved context (X).

$$r = G(q, X)$$

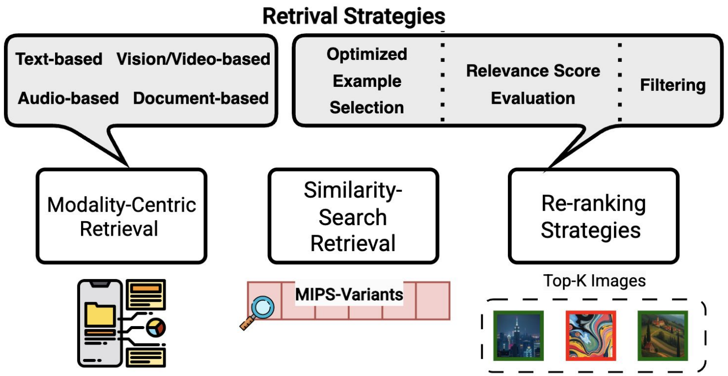
Taxonomy Overview

- Retrieval Strategies
- Fusion Mechanisms
- Augmentation Techniques
- Generation Methods
- Training Strategies & Robustness

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	

Retrieval Strategies

- **Efficient Search and Similarity Search**
  - *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 Mechanisms*



Efficient Search and Similarity Search

- **Multimodal Encoders**: create a **shared embedding space** to enable cross-modal retrieval.
    - *CLIP (Radford et al., 2021)*
    - *BLIP (Li et al., 2022)*
        - Cross-modal attention for richer image-text interaction
        - Unified encoder-decoder backbone handles retrieval and captioning
        - Bootstrapped data cleaning: synthetic captions filter noisy web pairs
    - *MARVEL (Zhou et al., 2024c)*
        - Visual Module Plugin
        - Two-Stage Adaption: Employs a specialized training strategy that first adapts the visual module, then freezes it to finetune the language model, effectively transferring its text-matching knowledge to the multimodal domain.
    - *UniIR (Wei et al., 2024a)*
        - Instruction-tuned universal retriever

Efficient Search and Similarity Search (cont.)

ACL 2025 VIENNA

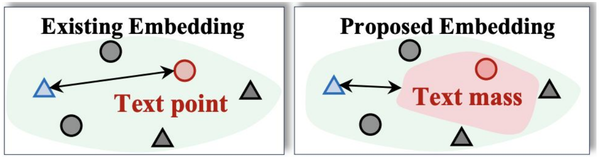
- **Maximum Inner Product Search:** Crucial for speeding up the search process in large-scale multimodal RAG by **approximating** the top-k most relevant items.
  - *ScaNN (Scalable Nearest Neighbors) (Guo et al., 2020)*
    - Introduces Anisotropic Vector Quantization: Moves beyond traditional methods by no longer minimizing simple reconstruction error.
    - Heavily penalizes error parallel to a vector, which is most disruptive to the inner product score, rather than treating all errors equally.
  - *TPU-KNN (Chern et al., 2022)*
  - *BanditMIPS (Tiwari et al., 2023)*
    - Instead of full comparisons, it estimates inner products by sampling coordinates. It adaptively focuses computation on the most promising vectors while quickly eliminating poor candidates.
  - *MUST (Wang et al., 2023)*

Modality-Centric Retrieval

ACL 2025 VIENNA

- **Text-Centric Retrieval:** Interaction mechanisms that preserve nuanced textual details to improve precision for multimodal queries.
  - *ColBERT (Khattab and Zaharia, 2020)*
  - *PreFLMR (Lin et al., 2024b)*
  - *BGE-M3 (Chen et al., 2024b)*
- **Vision-Centric Retrieval:** Retrieve based on visual similarity or compositional image features.
  - *EchoSight (Yan and Xie, 2024)*
  - *eClip (Kumar and Marttinen, 2024)*
- **Audio-Centric Retrieval:** bypass traditional ASR pipelines; enable audio-based retrieval.
  - *WavRAG (Chen et al., 2025b)*
  - *SEAL (Sun et al., 2025)*

Modality-Centric (cont.)

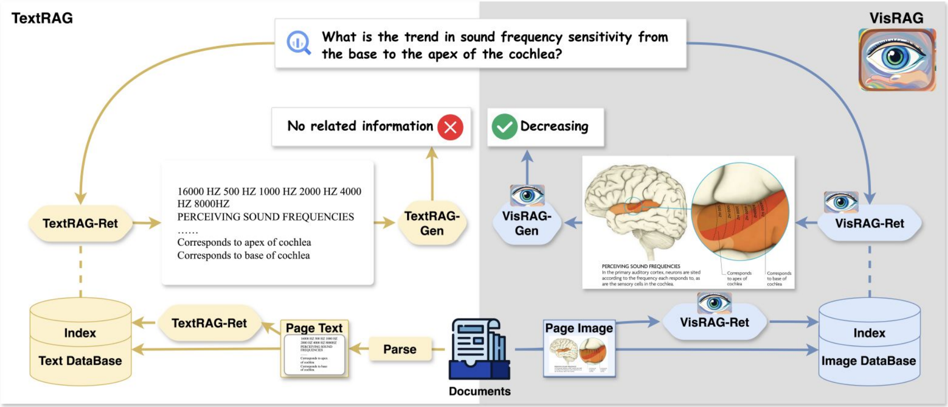


T-MASS (Wang et al., 2024)

- **Video-Centric Retrieval:** Handle temporal dynamics and long-context video retrieval.
  - *OmAgent (Zhang et al., 2024e)*
    - addresses the challenge of complex video understanding with a divide and conquer framework.
  - *VideoRAG (Ren et al., 2025)*
    - Graph-grounded clip index: builds a knowledge graph that links transcripts and visual captions across temporally segmented clips, letting the retriever hop between related moments in different videos.
    - Dual-channel retrieval & generation: combines text-based graph hops with frame-level visual similarity so the model can fetch and fuse evidence from both modalities before answering.
  - *T-MASS (Wang et al., 2024)*
    - Short, concise text query is often not descriptive enough to capture all the rich, redundant information in a video.
    - Introduce Stochastic Text Embedding (Text Mass): models text not as a single point, but as a probabilistic "mass" in the embedding space.

Modality-Centric (cont.)

- **Document Retrieval and Layout Understanding:** Process entire documents by integrating textual, visual, and spatial layout signals.
  - *VisRAG (Yu et al., 2025)*
    - Treating entire document pages as single images for both the VLM-based retriever and generator, completely eliminating the error-prone text parsing/OCR stage.



VisRAG (Yu et al., 2025)

Modality-Centric (cont.)

- **Document Retrieval and Layout Understanding:** Process entire documents by integrating textual, visual, and spatial layout signals.
  - *SV-RAG (Chen et al., 2025)*
    - Relying on the MLLM's inherent ability for holistic layout understanding.
    - Uses dual LoRA adapters to efficiently specialize a single, shared MLLM for the separate tasks of retrieval and question-answering.
    - For retrieval, it adopts a sophisticated late-interaction mechanism, similar to ColBERT.
  - *ColPali (Faysse et al., 2025)*
    - Bypasses brittle OCR and layout parsers by directly creating multi-vector embeddings from the image of the document page.
    - Employs a late-interaction mechanism (inspired by ColBERT) to compute fine-grained similarity between query text and the document's visual patch embeddings, enabling layout-aware retrieval.

Re-ranking Strategies

Effective retrieval in multimodal RAG systems requires not only identifying relevant information but also prioritizing retrieved candidates.

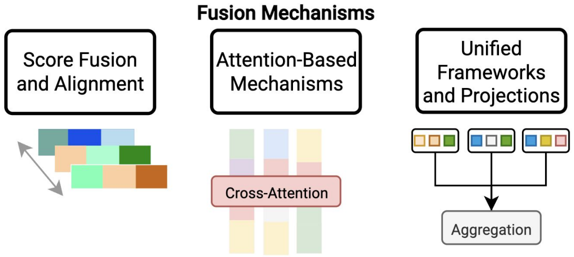
- **Optimized Example Selection:** Select the best context candidates using statistical, semantic, or visual signals.
- **Relevance Score Evaluation:** Measure and refine the semantic similarity between query and candidate contexts.
    - *RAG-Check (Mortaheb et al., 2025)*
        - *Introduces a Relevance Score (RS) model to explicitly evaluate the selection performance of a multimodal RAG system, directly addressing "selection-hallucination" by using the power of multi-head cross-attention.*

Re-ranking Strategies (cont.)

- **Filtering Mechanisms:** Eliminate irrelevant, noisy, or biased results before generation.
  - *MuRAR (Zhu et al., 2025)*
    - Generates an initial text-only answer and uses snippets of it as queries for multimodal retrieval.
    - Refines the initial response by prompting an LLM to integrate the retrieved multimodal evidence, implicitly filtering or ignoring less relevant information during the final generation.
  - *GME (Zhang et al., 2024i)*
    - Clusters training data into modality-specific groups to learn fine-grained, within-modality correlations.
    - Employs intra-group hard negative mining to force the model to distinguish between highly similar items.

Fusion Mechanisms

- **Score Fusion and Alignment**
  - Aligning modalities by embedding them in a shared space.
  - Fusing relevance scores from different retrieval models.
  - *RA-BLIP (Ding et al., 2024b)*
    - 3-layer BERT-based fusion of vision and language embeddings
    - provides richer interaction compared to simpler contrastive alignment or late-interaction methods.
  - *MUST (Wang et al., 2024c)*
    - *Introduces a model to automatically learn the importance (weights) of each modality for creating a joint similarity score.*



Fusion Mechanisms (cont.)

- **Attention-Based Mechanisms**
  - Using cross-attention to dynamically integrate features from different modalities.
  - *REVEAL (Hu et al., 2023)*
    - Introduces a novel attentive fusion layer that injects retrieval scores directly into the generator's attention mechanism.
    - This makes the retriever differentiable, allowing the entire system (retriever and generator) to be jointly trained to optimize the final answer generation.
  - *EMERGE (Zhu et al., 2024b)*
    - Uses a bidirectional cross-attention network to fuse embeddings from clinical time-series data and RAG-enhanced textual notes.

Fusion Mechanisms (cont.)

ACL 2025 VIENNA

- **Unified Frameworks and Projections**
  - Consolidating diverse inputs into a single, coherent representation.
  - Converting modalities (e.g., image-to-caption) to unify the input format.
  - *SAM-RAG (Zhai, 2024)*
    - Converts image inputs to captions to unify modality into text.
    - Simplifies downstream generation using unimodal LLMs.
  - *DQU-CIR (Wen et al., 2024)*
    - Converts images into text captions for complex queries and overlaying text onto images for simple ones.
    - Fuses visual-text features using learned MLP weights

Augmentation

ACL 2025 VIENNA

- **Context Enrichment**
    - Integrating extra data like entity relationships to enrich context.
    - *EMERGE (Zhuet al., 2024b)*
    - *Img2Loc (Zhou et al., 2024e)*
        - *including both similar and dissimilar points in prompts, helping rule out implausible locations.*
    - Reformulating user queries to pull in more multimodal information.
    - *Video-RAG (Luo et al., 2024b)*
        - *reformulates user queries into structured retrieval requests to extract auxiliary multimodal context*

Augmentation (cont.)

- **Adaptive Retrieval**
  - Dynamically choosing the best data source and granularity for a query.
  - *UniversalRAG (Yeo et al., 2025)*
    - Introduces an LLM-based retrieval router that dynamically selects the most appropriate knowledge source for a query.
    - Routes queries based on the required modality (text, image, video) and granularity (e.g., paragraph vs. document, clip vs. full video).
  - *OmniSearch (Li et al., 2024d)*
    - Decomposes multimodal queries into structured sub-questions, planning retrieval actions in real time.

Augmentation (cont.)

- **Iterative Retrieval**
  - Refining search results over multiple steps using feedback from each iteration.
  - *UniversalRAG (Yeo et al., 2025)*
    - Orchestrates a multi-step, coarse-to-fine retrieval process for knowledge-based VQA.
    - Begins with a broad entity search and progressively refines results using multimodal reranking and textual filtering to pinpoint evidence.
  - *RAGAR (Khaliq et al., 2024)*

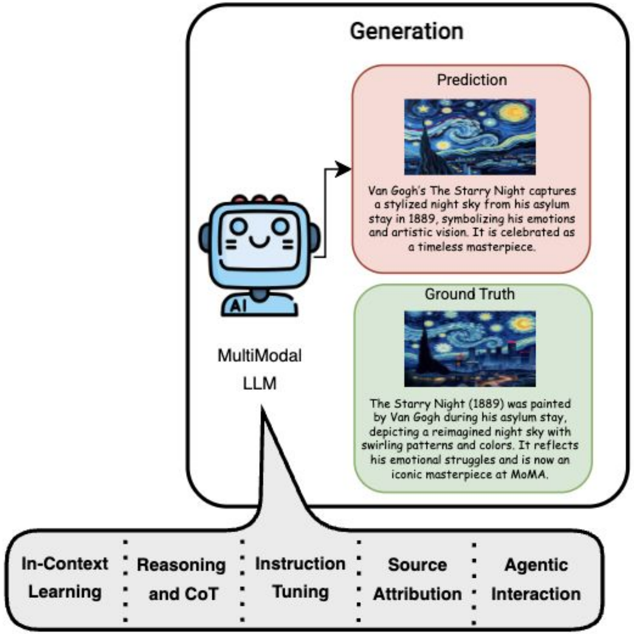
Augmentation

Context Enrichment

Adaptive/Iterative Retrieval

Generation

- **In-Context Learning:** Using retrieved content as few-shot examples to guide the model without retraining.
- **Reasoning:** Decomposing complex problems into sequential steps to improve coherence and robustness.
- **Instruction Tuning:** Fine-tuning the generation model to better handle specific tasks and user instructions.
- **Source Attribution:** Prompting the model to explicitly cite evidence from retrieved sources in its response.
- **Agentic Generation:** Using autonomous agents that can perform complex reasoning, interact with users, and coordinate tasks.



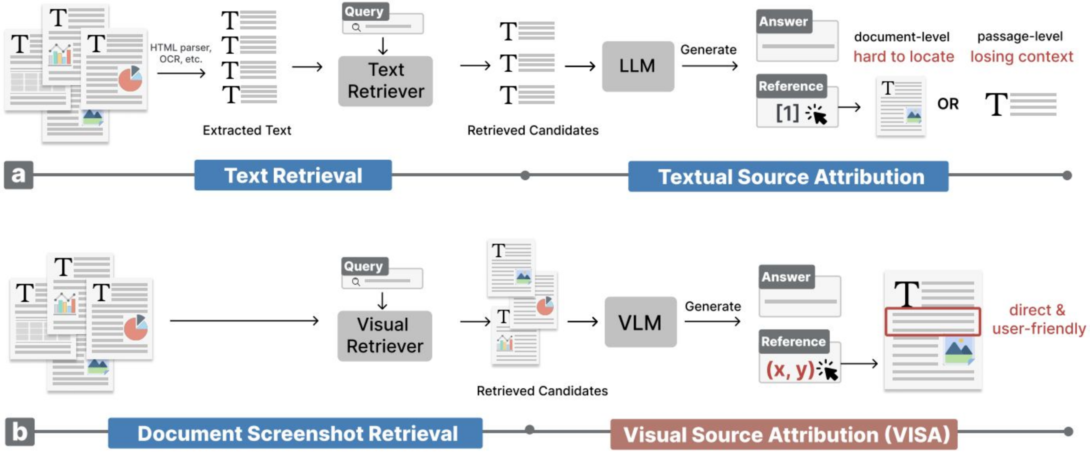
Generation (cont.)

- In-Context Learning
  - *RA-CM3 (Yasunaga et al., 2023), RMR (Tan et al., 2024), MSIER (Luo et al., 2024a)*
- Reasoning
  - *RAGAR (Khaliq et al., 2024)*
    - *Chain of RAG (CoRAG): This method sequentially asks a follow-up question based on the answer to the previous one, creating a step-by-step chain of evidence to verify a claim.*
    - *Tree of RAG (ToRAG): This technique generates multiple branches of questions at each step, evaluates them, and selects the single best question-answer path to pursue for fact-checking.*
  - *VisDoMRAG (Suri et al., 2025)*
    - *Introduces a consistency-constrained fusion where the reasoning chains from parallel visual and textual pipelines are aligned to produce a coherent final answer.*

Generation (cont.)

- **Instruction Tuning**
  - *RagVL (Chen et al., 2024e), MMed-RAG (Xia et al., 2024a)*
  - *Rule (Xia et al., 2024b)*
    - Refines a medical large vision language model through direct preference optimization (DPO) to mitigate overreliance on retrieved contexts.
- **Source Attribution and Evidence Transparency**
  - *OMG-QA (Nan et al., 2024b)*
    - Prompts LLMs for explicit evidence citation in generated responses.
  - *VISA (Ma et al., 2024b)*
    - Traditional systems typically cite the entire source document, which forces users to search through dense text to find the supporting evidence.
    - VISA solves this by attributing the answer to a specific content area—such as a passage, table, or image—within the document screenshot.

Generation (cont.)

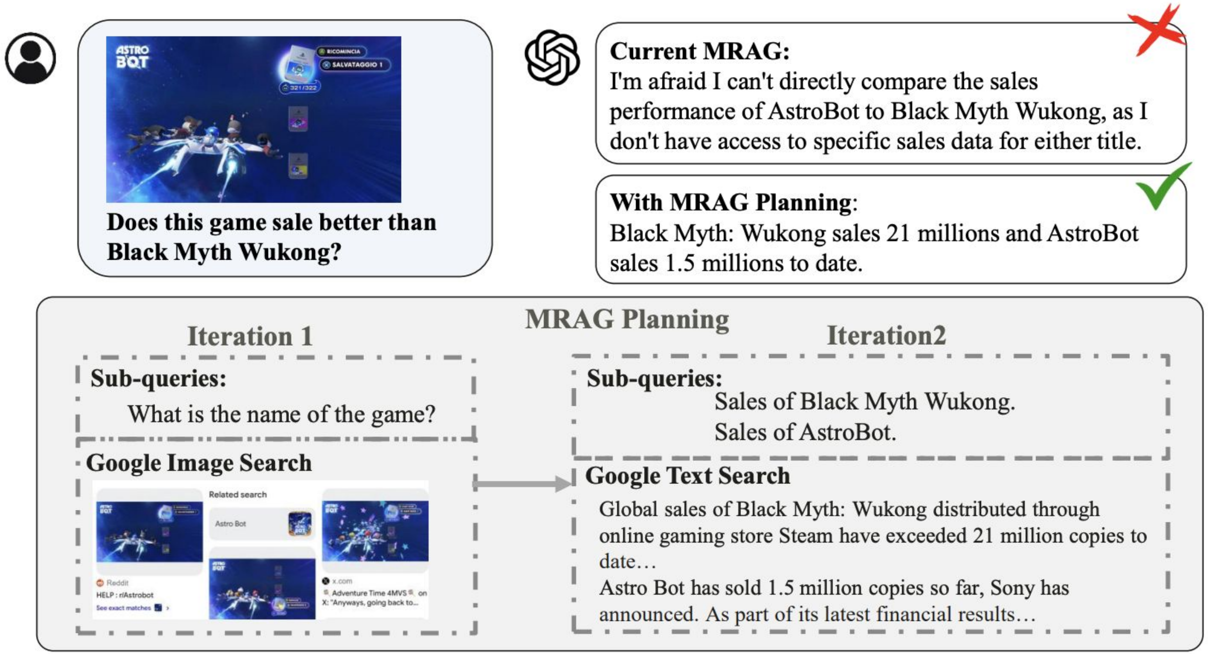


VISA (Ma et al., 2024b)

Generation (cont.)

- **Agentic Generation and Interaction**
  - *HM-RAG (Liu et al., 2025a)*
    - Introduces a three-tiered agent architecture to deconstruct, retrieve, and synthesize information, moving beyond single-agent limitations.
    - Coordinates multiple retrievers for different modalities.
  - *CogPlanner (Yu et al., 2025)*
    - Introduces a "Planning Expert" Agent: This agent dynamically creates a multi-step plan for each query, deciding if, what (text or image), and how to search, moving beyond the rigid, single-step pipelines of previous systems.
    - The agent mimics human cognitive processes by iteratively reformulating complex queries and adapting its retrieval strategy at each step.

Generation (cont.)

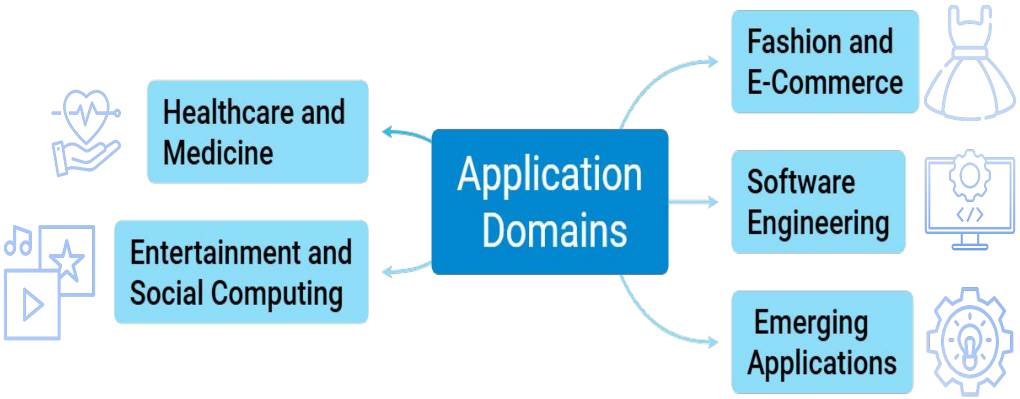


CogPlanner (Yu et al., 2025)

Training Strategies

- **Alignment**
  - Contrastive learning
  - Hard-negative mining
    - *HACL (Jianget al., 2024)*
      - Mitigates hallucinations by incorporating adversarial captions as distractors.
- **Robustness Management**
  - Training with noisy inputs and irrelevant results.
  - Enhancing focus through progressive knowledge distillation.
  - Using regularization, like randomly dropping query tokens.
    - *RA-CM3 (Yasunaga et al., 2023)*
      - Query Dropout: The model learns to handle imperfect retrieval and still generate correct outputs.
- **Loss Functions**

Application Domains



Application Domains

- Multimodal RAG Application Domains
  - Healthcare and Medicine (§E)
    - MMED-RAG (Xia et al., 2024a), RULE (Xia et al., 2024b), AsthmaBot (Bahaj and Ghogho, 2024), Realm (Zhu et al., 2024c), Su et al. (2024a), FactMM-RAG (Sun et al., 2024b), RA-RRG (Choi et al., 2025)
  - Software Engineering (§E)
    - DocPrompting (Zhou et al., 2023), RACE (Shi et al., 2022), CEDAR (Nashid et al., 2023), RED-CODER (Parvez et al., 2021)
  - Fashion and E-Commerce (§E)
    - Unifashion (Zhao et al., 2024), Dang (2024), Fashion-RAG (Sanguigni et al., 2025), LLM4DESIGN (Chen et al., 2024d)
  - Entertainment and Social Computing (§E)
    - SoccerRAG (Strand et al., 2024), MMRA (Zhong et al., 2024)
  - Emerging Applications (§E)
    - RAG-Driver (Yuan et al., 2024), ENWAR (Nazar et al., 2024), Riedler and Langer (2024), Img2Loc (Zhou et al., 2024e)

Open Problems & Future Directions

- **Robustness & Explainability**
  - Modality biases (text over-reliance), provide precise source attribution.
- **Advanced Reasoning & Retrieval**
  - Compositional reasoning gaps, unified embedding spaces.
- **Agentic Frameworks & Self-Guidance**
  - Self-guided retrieval with reinforcement learning and interactive feedback.
- **Efficiency & Scalability:**
  - Long-context bottlenecks (videos/documents), edge deployment.
- **Personalization & Evaluation**
  - Advance privacy-preserving personalization, create more robust evaluation benchmarks
- **Embodied & Real-World Grounding**
  - Integrate real-world sensor data

Thank You!

Full paper: https://aclanthology.org/2025.findings-acl.861/

GitHub: https://github.com/llm-lab-org/Multimodal-RAG-Survey

Website: https://multimodalrag.github.io/

Any Question?

Website

Repository

Paper