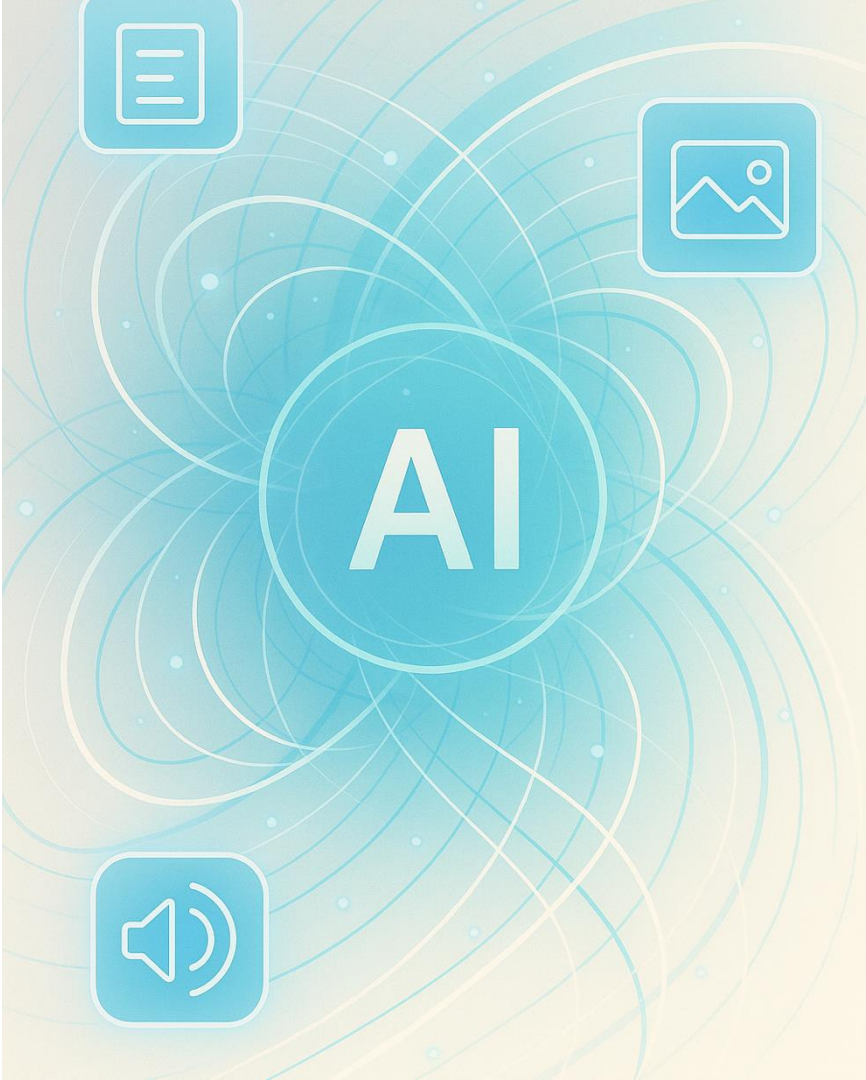
Multimodal Retrieval-Augmented Generation

From Text-Only RAG to Cross-Modal Intelligence

IJCAI 2025 Technical Survey

12 Aug 2025

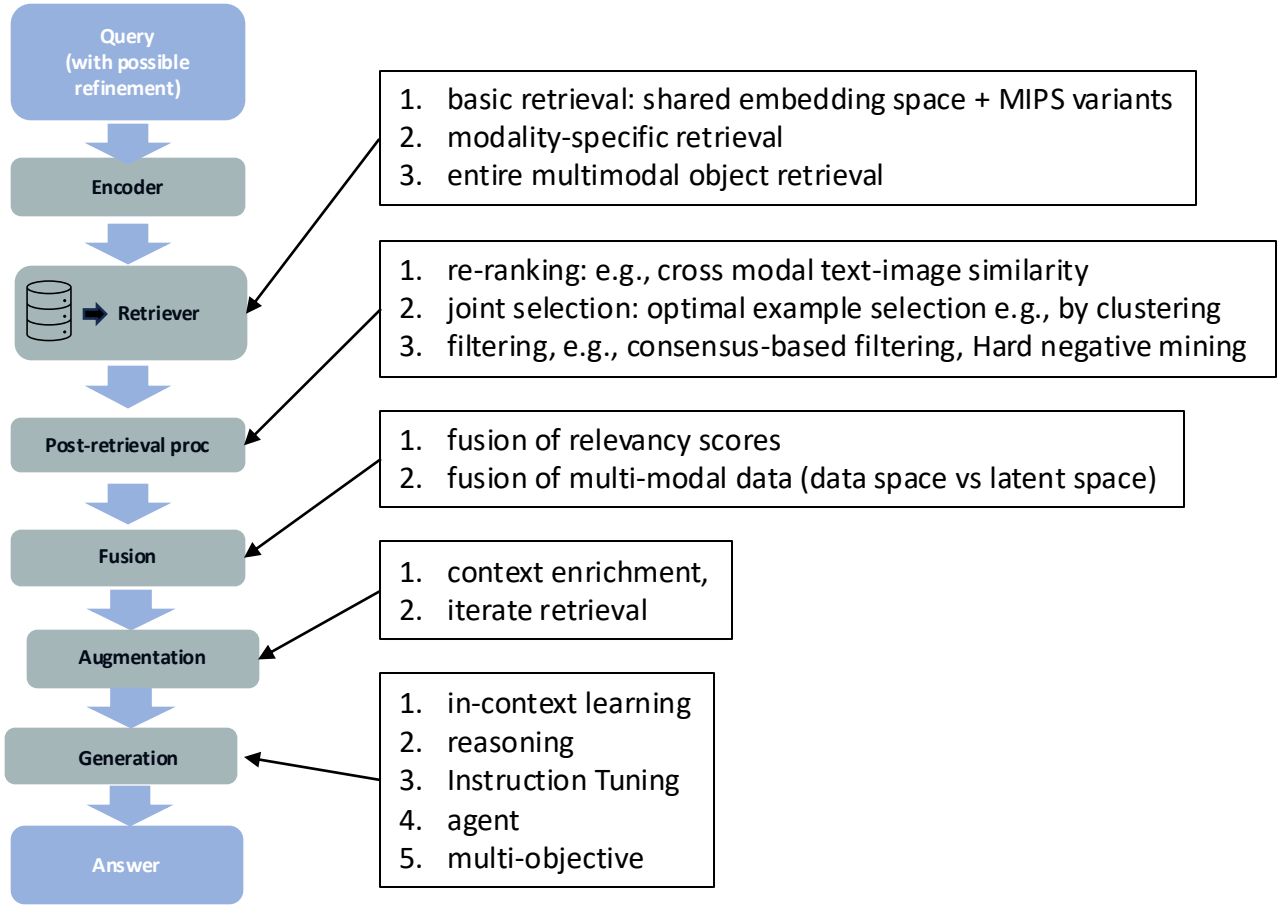
[1] [2]

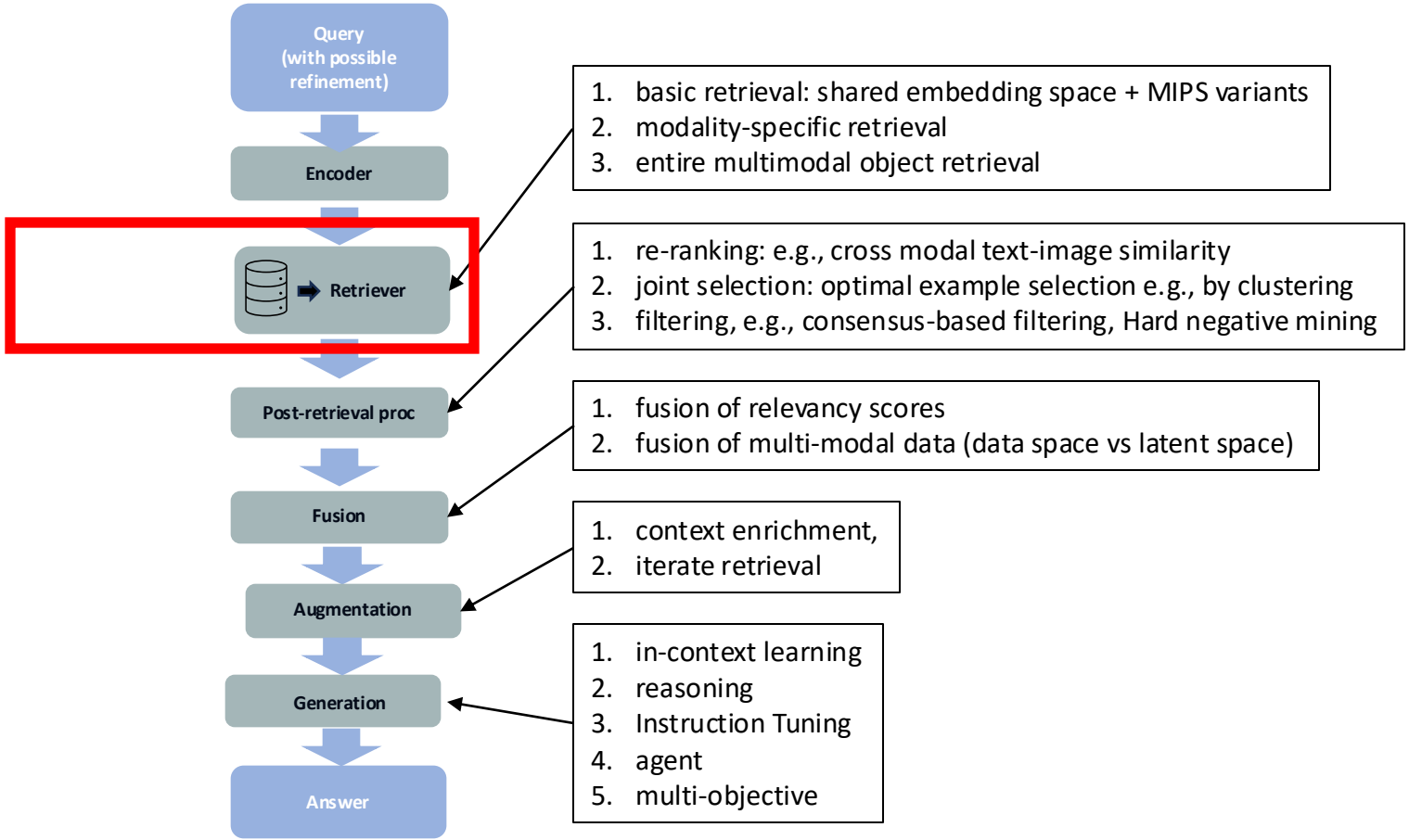


Our Outline

- Unstructured RAG
- Structured RAG
- Semi-structured RAG
- Multimodal RAG

Architecture of Multimodal RAG

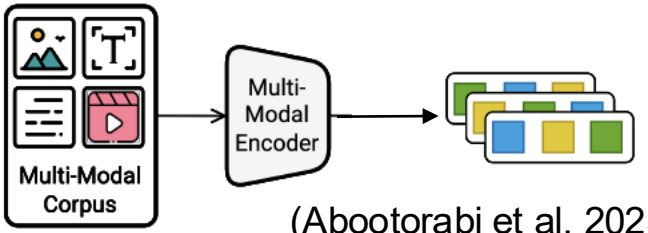



Query
(with possible
refinement)
Encoder
Retriever
Post-retrieval proc
Fusion
Augmentation
Generation
Answer
1. basic retrieval: shared embedding space + MIPS variants
2. modality-specific retrieval
3. entire multimodal object retrieval
1. re-ranking: e.g., cross modal text-image similarity
2. joint selection: optimal example selection e.g., by clustering
3. filtering, e.g., consensus-based filtering, Hard negative mining
1. fusion of relevancy scores
2. fusion of multi-modal data (data space vs latent space)
1. context enrichment,
2. iterate retrieval
1. in-context learning
2. reasoning
3. Instruction Tuning
4. agent
5. multi-objective

Basic retrieval: shared space + MIPS and variants

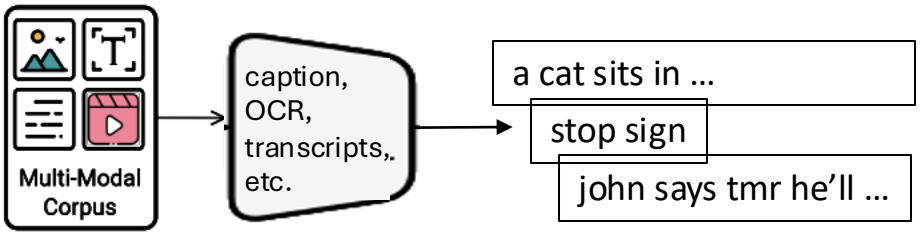
Step 1: Unifying

Type 1: shared embedding space



(Abootorabi et al. 2025)

Type 2: shared data space



Step 2: **Similarity**, sparse and dense, e.g., Max inner-product search (MIPS) and variants

Modality-specific retrieval

(i) Text-centric retrieval

- sparse: e.g., BM25 (Robertson and Zaragoza, 2009)
- dense: e.g., BGE-M3 (Chen et al., 2024b), RAFT (Zhang et al., 2024h) and CRAG (Yan et al., 2024), ColBERT (Khattab and Zaharia, 2020) and PreFLMR (Lin et al., 2024b)

(ii) Vision-centric retrieval

- Basic: visual embedding (e.g., EchoSight (Yan and Xie, 2024), ImgRet (Shohan et al., 2024)
- Sub-dimensional: embedding per query, e.g., Cross-modal RAG (Zhu et al., 2025)
- Composed image retrieval (CIR): a reference image + modifying text (Feng et al., 2023; Zhao et al., 2024; Jang et al., 2024; Saito et al., 2023)

(iii) Video-centric retrieval

incorporating temporal dynamics and large video-language models

- Temporal understanding → e.g., TCA: Temporal Context Aggregation
- Multimodal fusion → e.g., MM Fusion Transformer
- Keyframe for efficiency → e.g., Visual-Subtitle Integration for Keyframe Retrieval
- Semantic-temporal alignment → e.g., TF-CoVR-Base

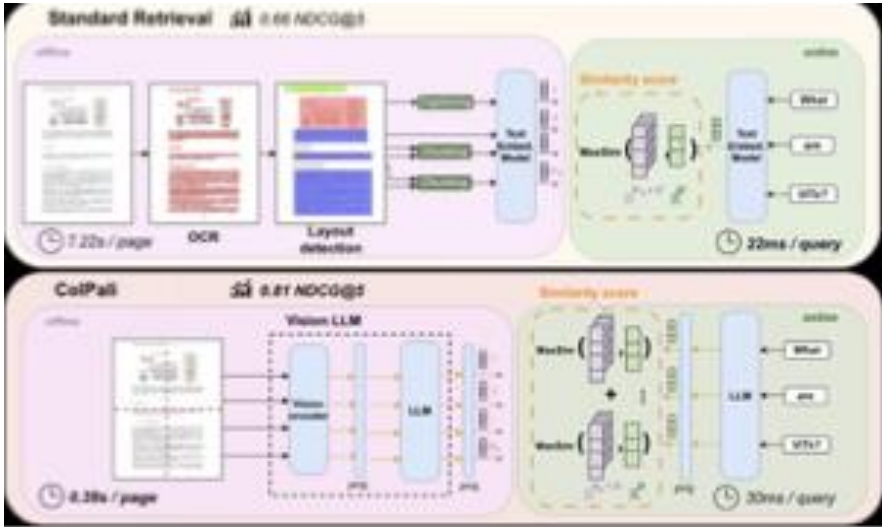
Modality-specific retrieval (cont.)

(iv) Audio-centric retrieval

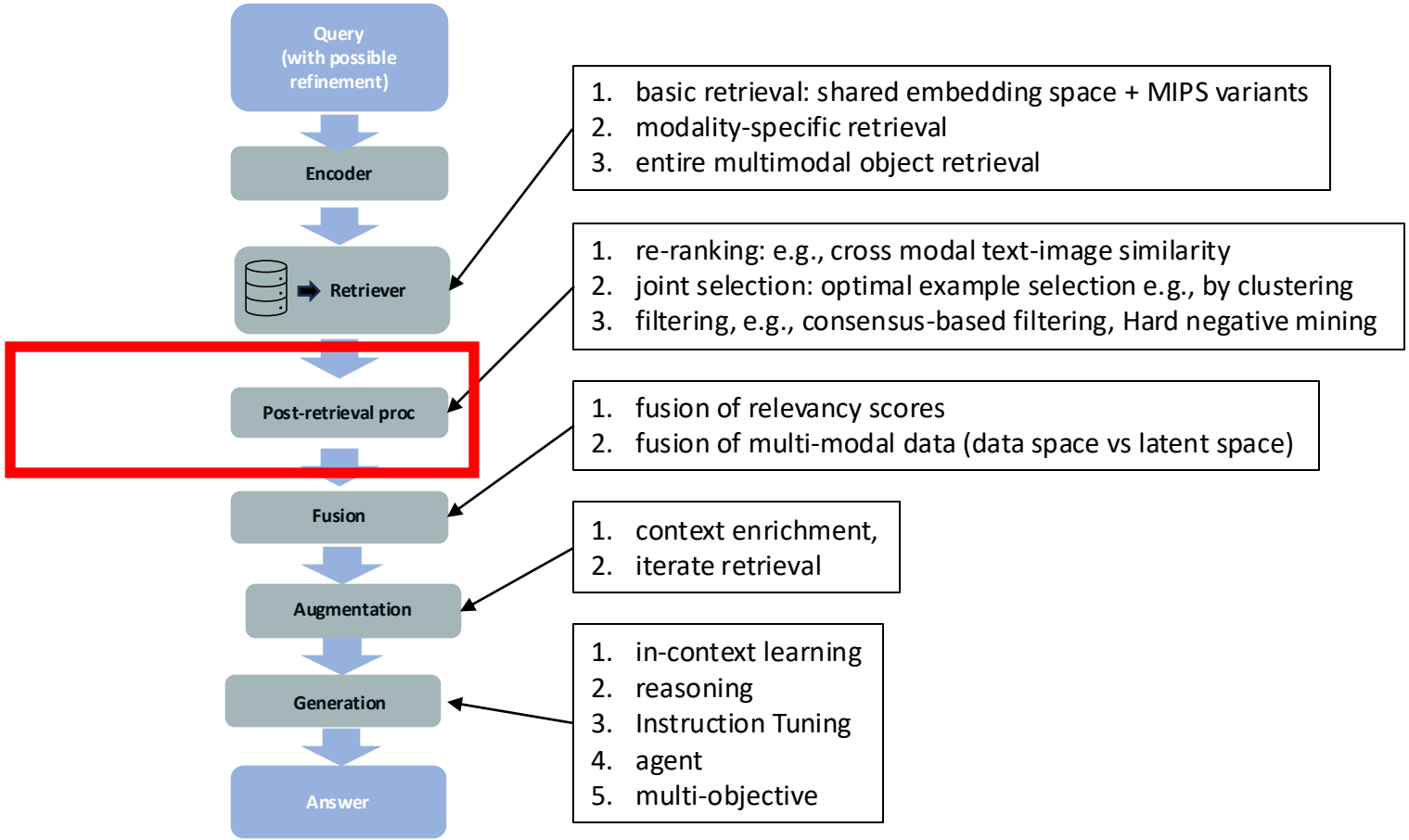
Characteristic	Technique Name	Example Method / Paper
Bypass ASR & unified embedding	WavRAG – end-to-end audio embedding	*WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models* (Chen et al., 2025) (arXiv)
Unified speech–text embedding	SEAL – shared latent space	*SEAL: Speech Embedding Alignment Learning for Speech Large Language Model with RAG* (Sun et al., 2025) (arXiv)
Prompt enrichment via retrieved audio	Audiobox TTA-RAG	(Yang et al., 2024a) as cited in survey (arXiv)
Audio-based captioning with domain adaptivity	DRCap – CLAP-bridged latent space	*DRCap* (Li et al., 2025c) as cited in survey (arXiv)
Dynamic caption queries	P2PCAP – regenerated captions	*P2PCAP* (Changin et al., 2024) as cited in survey (arXiv)
ASR error correction via speech retrieval	LA-RAG – speech-to-speech & forced alignment	*LA-RAG* (Li et al., 2024b) as cited in survey (arXiv)
Error correction in noisy environments	LLM-augmented hybrid system	*Xiao et al.* (2025) hybrid audio–text error correction with LLMs (arXiv)

Entire multimodal object retrieval

Recent research has moved beyond traditional unimodal retrieval, developing models that process entire documents by integrating textual, visual, and layout information.



Category	Focus	Representative Methods
OCR-Free, Vision-Only Retrieval	Direct image embeddings, no OCR or text parsing	**ColPali**, **VisRAG** (arxiv.org, Together AI)
Layout-Aware & Structured Comprehension	Integrating layout, multi-page context, generative alignment	**ViTLP**, **DocLLM**, **CREAM**, **mPLUG-DocOwl**, **SV-RAG** (github.com, arxiv.org)

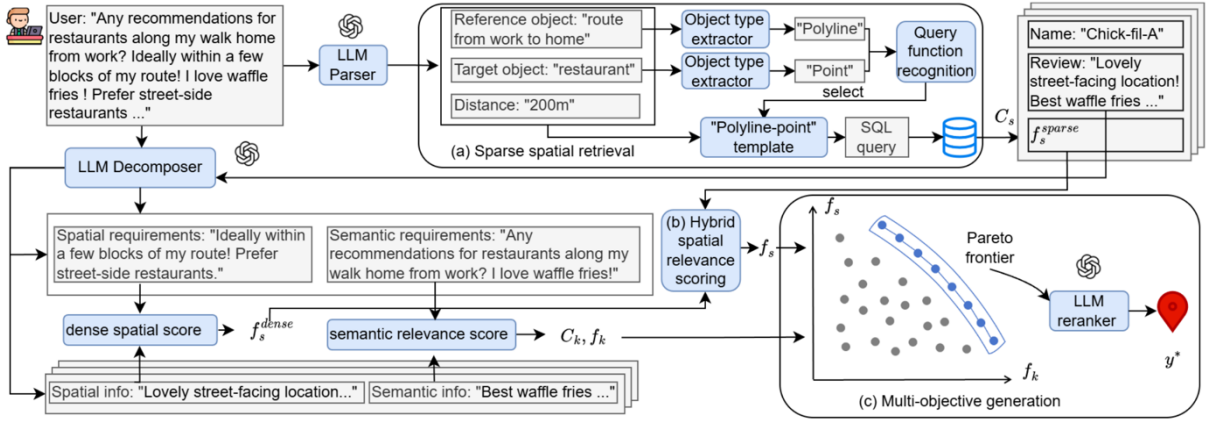

Query
(with possible
refinement)
Encoder
Retriever
Post-retrieval proc
Fusion
Augmentation
Generation
Answer
1. basic retrieval: shared embedding space + MIPS variants
2. modality-specific retrieval
3. entire multimodal object retrieval
1. re-ranking: e.g., cross modal text-image similarity
2. joint selection: optimal example selection e.g., by clustering
3. filtering, e.g., consensus-based filtering, Hard negative mining
1. fusion of relevancy scores
2. fusion of multi-modal data (data space vs latent space)
1. context enrichment,
2. iterate retrieval
1. in-context learning
2. reasoning
3. Instruction Tuning
4. agent
5. multi-objective

Re-ranking

Strategy	Description	Example Method or Technique
Averaging	Averages similarity scores from multiple sources	**VR-RAG**: Combines cross-modal (text–image) and intra-modal visual similarity via DINOv2, ensemble of CLIP/OpenCLIP/SigLIP.
Fusion	Learns a fused similarity function via tensor fusion layers	**MTFN-RR**: Multi-modal Tensor Fusion Network for image-text matching, enabling re-ranking.
Weighting	Combines multiple modality-specific scores with weights for final relevance	**mR2AG, EchoSight**
Consistency	items rank consistently high in cross-modal search will rank high	**LDRE** (Yang et al., 2024) employs semantic ensemble methods to adaptively weigh multiple features.

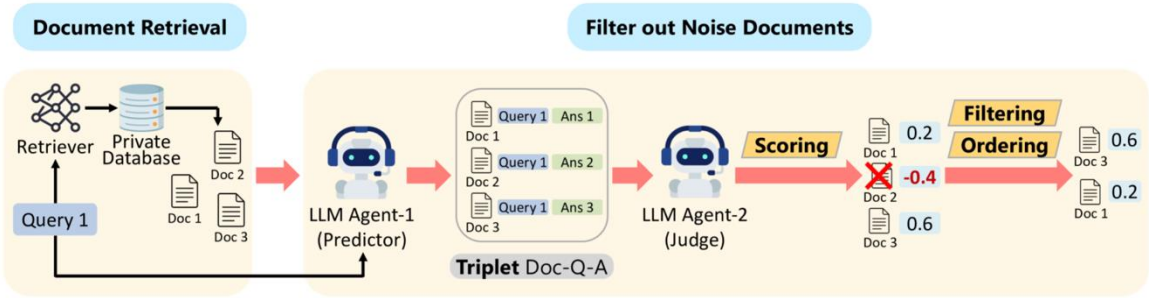
Joint Selection

Category	Description	Example Method(s)
Supervised Control Signals	Incorporates guided keywords or prompts to steer retrieval with modality-specific control	**Hybrid RAG: Su et al. (2024a)** Probabilistic control keywords
Statistical Calibration	Applies statistical corrections to filter unreliable or hallucinated context	**RULE (Xia et al., 2024b):** Uses Bonferroni correction
Unsupervised Diversity	Clusters modal units (e.g., video frames) to ensure diverse context coverage	**Dong et al. (2024b):** Key-frame selection
Hybrid Multi-Step Retrieval	Sequential retrieval combining supervised and unsupervised stages	**Luo et al. (2024a); Yuan et al. (2023):** Multi-step pipelines
Pareto Frontier	Considering all the possible weights of different modalities,	**Spatial-RAG (Yu et al, 2025)**

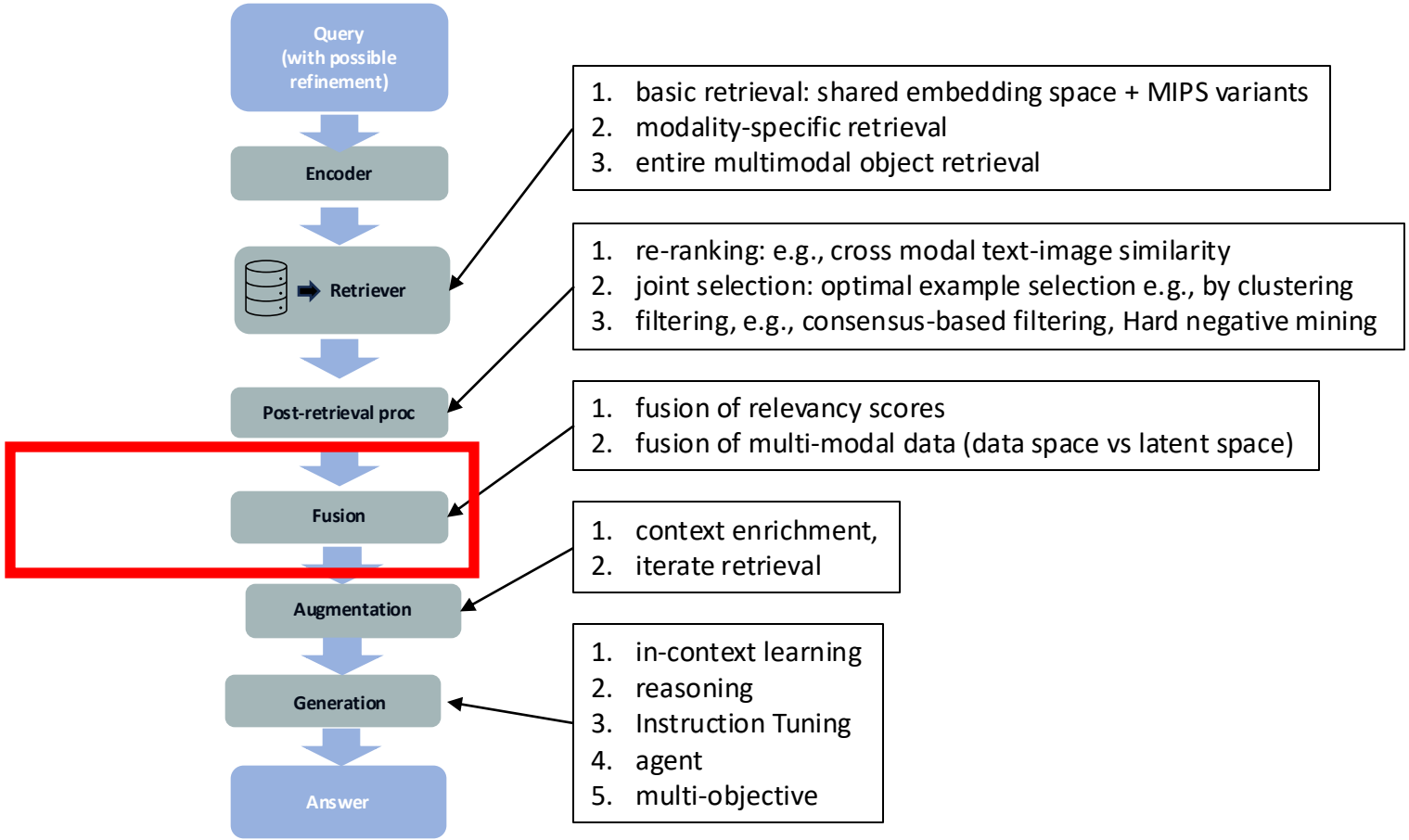


Spatial-RAG, Yu et al., 2025

Filtering



Category	Description	Representative Methods
Hard Negative Mining	Actively selects hard negatives specific to each modality to counter bias in training	**RAFT (**Zhang et al., 2024**)** **GME** (Zhang et al., 2024i), **MM-Embed** (Lin et al., 2024a)
Agent-Based Filtering	Uses agents to predict and score candidates	**MAIN-RAG (Chang et al., 2024)**
Learn to filter	Learns to disregard confusing or low-quality modalities at retrieval time	**RAFT** (Zhang et al., 2024h), **MAIN-RAG** (Chang et al., 2024) (OpenReview, ACL Anthology)

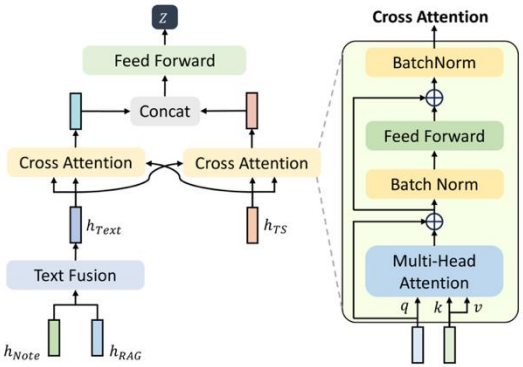

Query
(with possible
refinement)
Encoder
Retriever
Post-retrieval proc
Fusion
Augmentation
Generation
Answer
1. basic retrieval: shared embedding space + MIPS variants
2. modality-specific retrieval
3. entire multimodal object retrieval
1. re-ranking: e.g., cross modal text-image similarity
2. joint selection: optimal example selection e.g., by clustering
3. filtering, e.g., consensus-based filtering, Hard negative mining
1. fusion of relevancy scores
2. fusion of multi-modal data (data space vs latent space)
1. context enrichment,
2. iterate retrieval
1. in-context learning
2. reasoning
3. Instruction Tuning
4. agent
5. multi-objective

Fusion

1. fusion in data space

Interleave or concatenate raw modality data before any embedding occurs, allowing early joint processing.

Zhi Lim et al. (2024): Converts text, tables, and images into a unified textual format, enabling cross-encoder evaluation across fused data.
Sharifymoghaddam et al. (2024): Interleaves few-shot image–text pairs in the input (e.g., stacking images vertically alongside text prompts), then applies CLIP or BLIP for alignment. ICLR Proceedings+4ACL



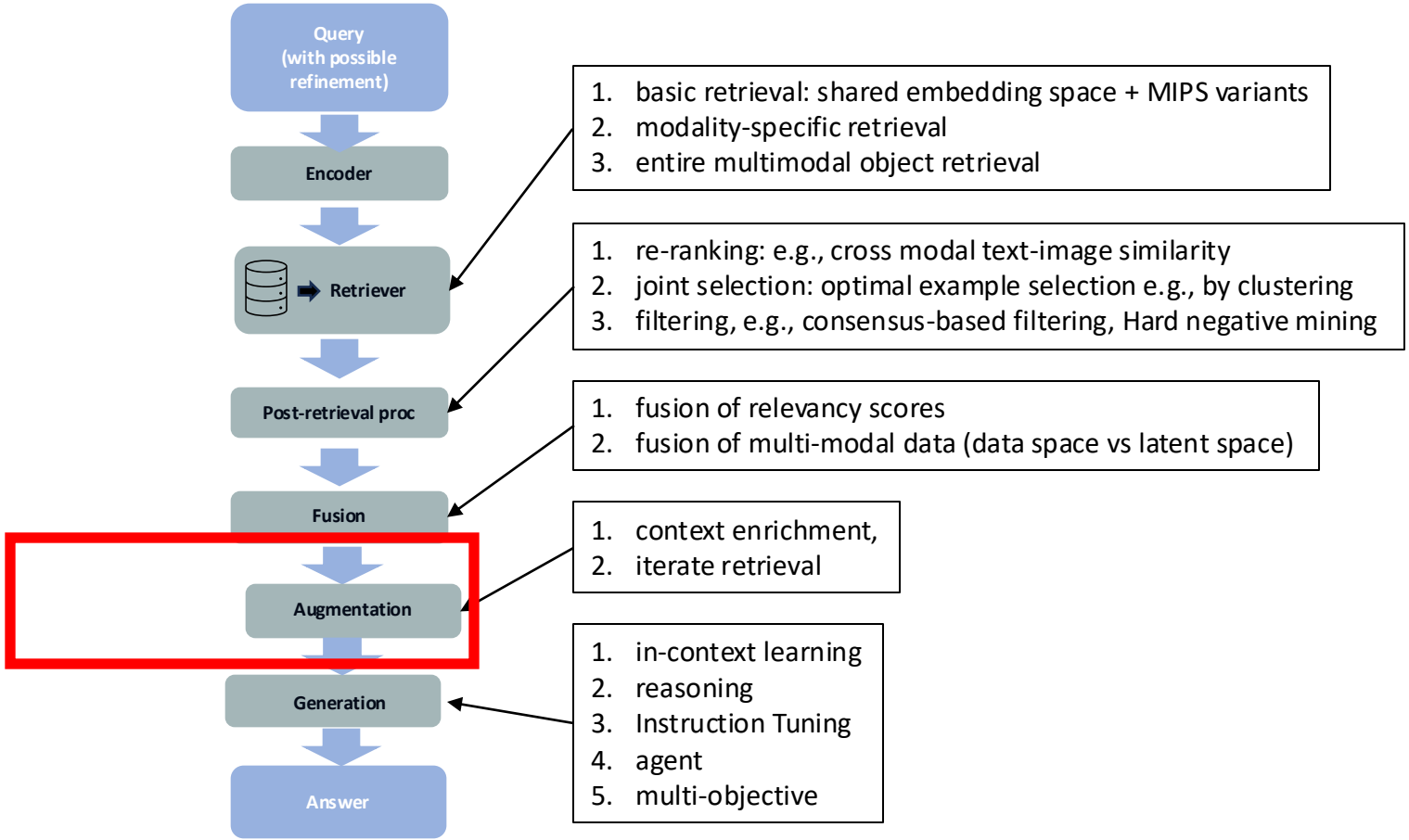
EMERGE (Zhu et al., 2024b)

2. fusion in latent space

Embed each modality separately and then fuse their vector representations—often via attention, projection, or multi-modal alignment.

EMERGE (Zhu et al., 2024b): After encoding each modality separately, it uses adaptive multimodal fusion with cross-attention to integrate the embeddings. arXiv+6arXiv+6ACL Anthology+6

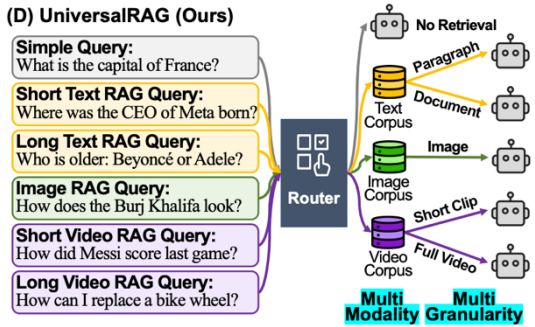
RAMM (Yuan et al., 2023): Employs a dual-stream co-attention transformer, combining self-attention and cross-attention over separate embeddings of retrieved biomedical images and text. ACL Anthology+4arXiv+4ACL Anthology+4


Query
(with possible
refinement)
Encoder
Retriever
Post-retrieval proc
Fusion
Augmentation
Generation
Answer
1. basic retrieval: shared embedding space + MIPS variants
2. modality-specific retrieval
3. entire multimodal object retrieval
1. re-ranking: e.g., cross modal text-image similarity
2. joint selection: optimal example selection e.g., by clustering
3. filtering, e.g., consensus-based filtering, Hard negative mining
1. fusion of relevancy scores
2. fusion of multi-modal data (data space vs latent space)
1. context enrichment,
2. iterate retrieval
1. in-context learning
2. reasoning
3. Instruction Tuning
4. agent
5. multi-objective

Augmentation

Type 1: forward retrieval -> context enrichment

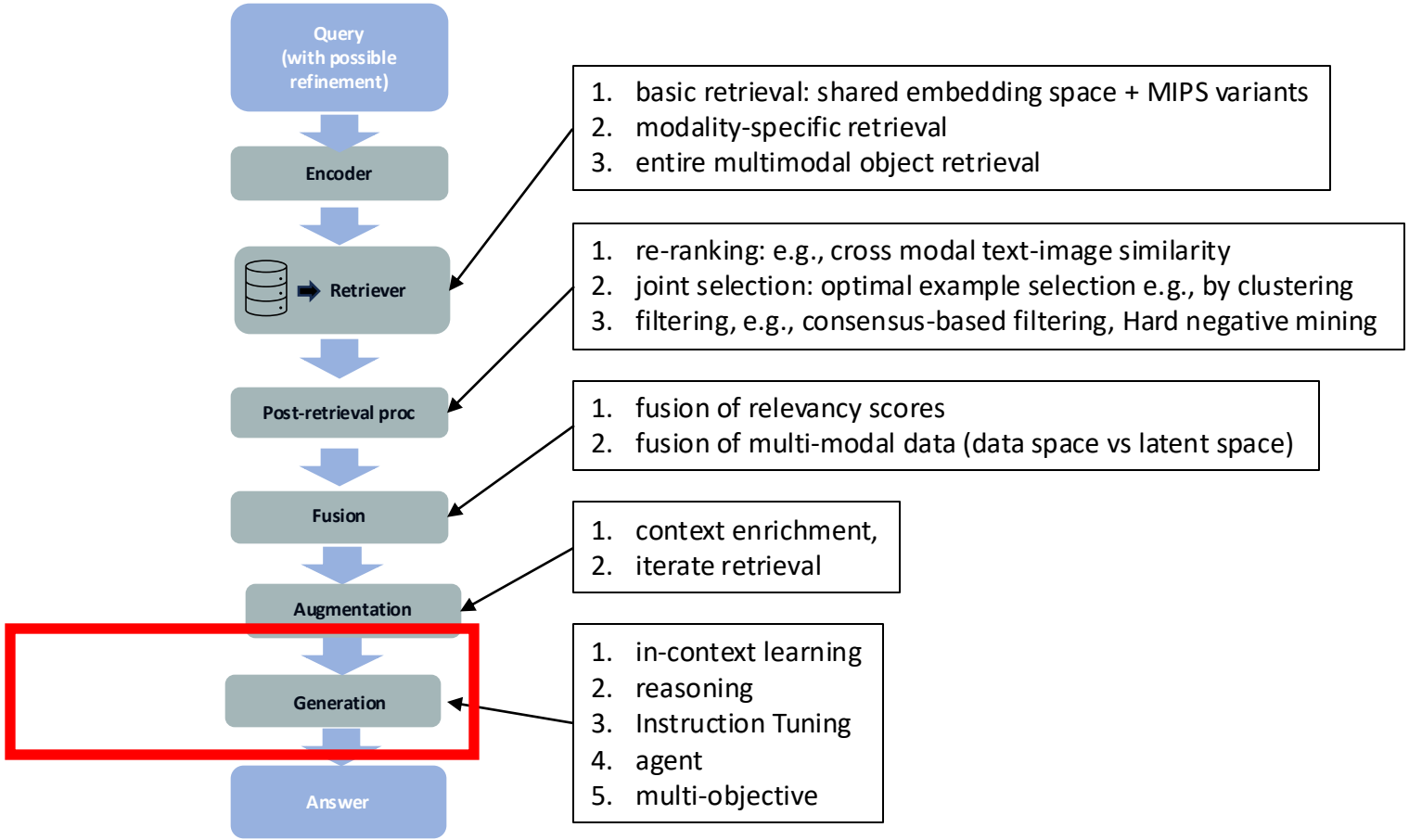
1. Entity-Centric Enrichment, e.g., EMERGE (Zhu et al., 2024b), MiRAG (Adjali et al., 2024)



2. Contrastive Context Inclusion, e.g., Img2Loc (Zhou et al., 2024e) UniversalRAG: prompt, finetune

Type 2: dynamic retrieval -> quantity and quality improvement of retrieved results

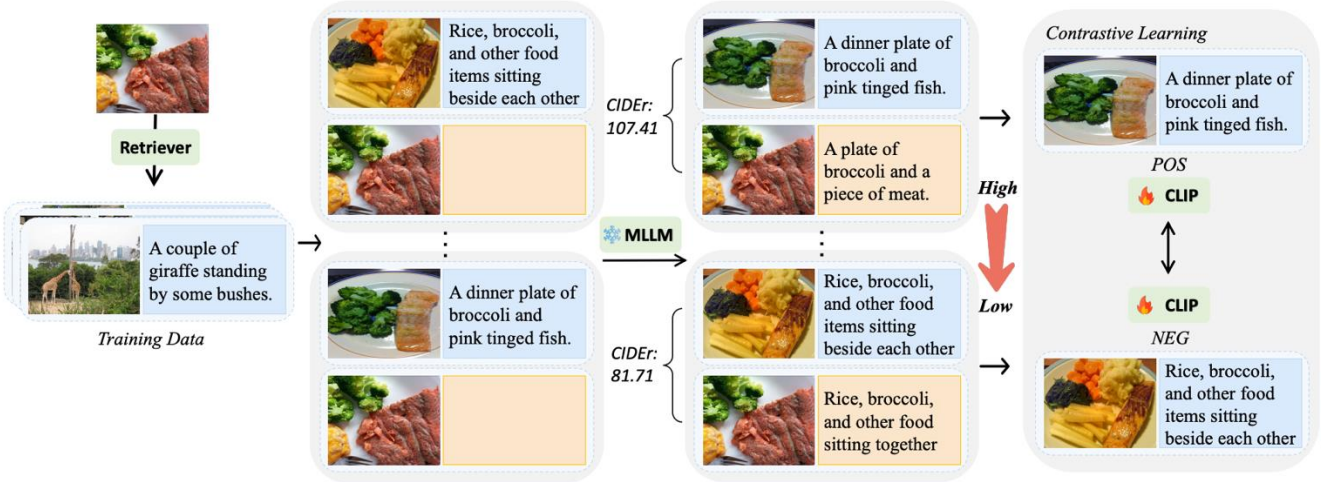
Category	Description	Example Methods
Adaptive Retrieval	Dynamically routes queries, chooses granularity, or filters results based on modality and task relevance.	UniversalRAG, SKURG, SAM-RAG, mR2AG, MMed-RAG, OmniSearch
Iterative Retrieval	Multi-step refinement using feedback, reranking, and memory to improve multimodal retrieval over time.	OMGM, IRAMIG, OMG-QA, RAGAR


Query
(with possible
refinement)
Encoder
Retriever
Post-retrieval proc
Fusion
Augmentation
Generation
Answer
1. basic retrieval: shared embedding space + MIPS variants
2. modality-specific retrieval
3. entire multimodal object retrieval
1. re-ranking: e.g., cross modal text-image similarity
2. joint selection: optimal example selection e.g., by clustering
3. filtering, e.g., consensus-based filtering, Hard negative mining
1. fusion of relevancy scores
2. fusion of multi-modal data (data space vs latent space)
1. context enrichment,
2. iterate retrieval
1. in-context learning
2. reasoning
3. Instruction Tuning
4. agent
5. multi-objective

In-context learning

ICL with retrieval augmentation enhances reasoning in multimodal RAGs by leveraging retrieved content as few-shot examples without requiring retraining.

e.g., RA-CM3 (Yasunaga et al., 2023), RAG-Driver (Yuan et al., 2024), MSIER Method (Luo et al. 2024)



Reasoning

1. Linear CoT with Multimodal Evidence

   Sequential reasoning combining modalities with curated evidence

   VisDoMRAG—curate evidence from text and visuals, then reasons via CoT alignment

2. Branching / Tree-Structured CoT

   Explores multiple reasoning paths before selecting answer

   RAGAR with Chain of RAG (CoRAG) and Tree of RAG (ToRAG)

3. Compositional Reasoning (e.g., for Composed Image Retrieval (CIR))

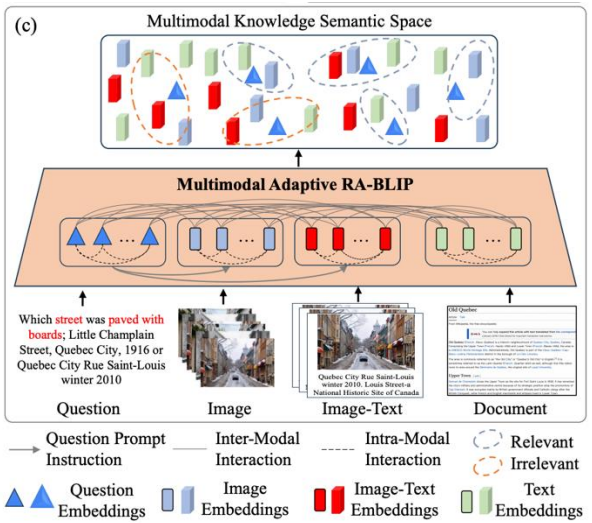
   Iteratively improves captions and combines visual-text reasoning

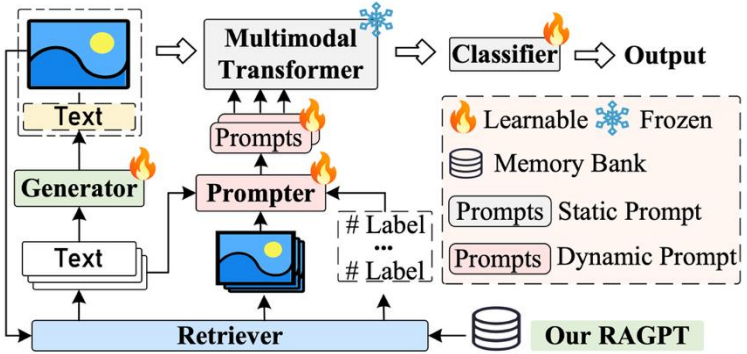
   LDRE—refines captions via dense descriptions for zero-shot composed image retrieval (multimodal CoT)

Instruction Tuning

Instruction-Aware Visual Tuning
Models adapt visual encoding based on task-specific instructions in tandem with text, enabling responsive, grounded feature extraction.
Example: **RA-BLIP** uses InstructBLIP to selectively extract visual features and fuse them via a tiny retrieval-adaptive module.

Retrieval-Augmented Prompt Tuning
Models dynamically generate prompts or decide when to retrieve using context-aware mechanisms—often integrating retrieved instances directly into the generation process.
•**Example**: **RAGPT** uses a context-aware prompter that, based on retrieved multimodal instances, produces dynamic prompts tailored to incomplete or varying modality inputs

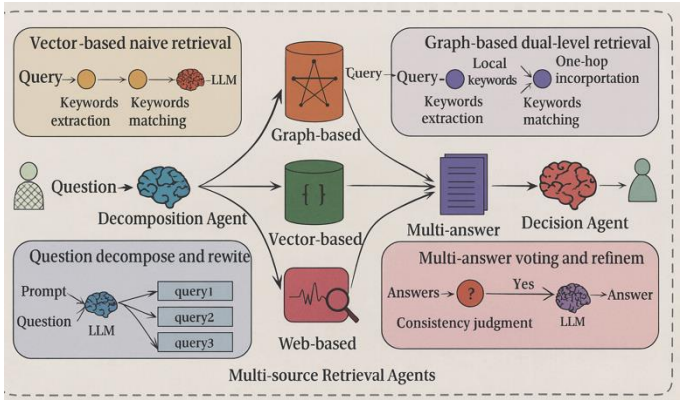




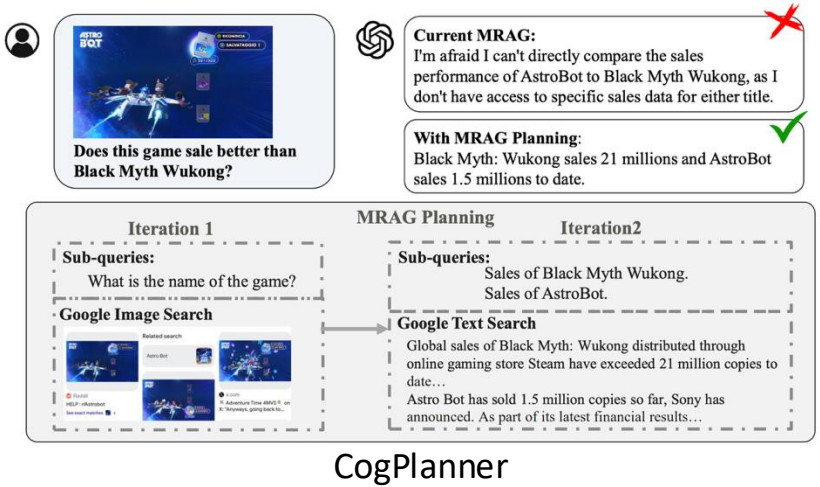
Agentic

Actively drives retrieval, tool calls, and reasoning like an agent.

Category	Description	Representative Examples
Domain-Specific Agents	Autonomous or semi-autonomous systems designed for particular contexts like mobile GUI navigation, personalized conversational agents, or anomaly detection.	**AppAgent v2, USER-LLM R1, MMAD, Yi et al., CollEX**
Hierarchical Multi-Agent Systems	Multi-tier architectures that split queries, retrieve across modalities and sources, and integrate results with consistency refinement.	**HM-RAG**
Cognitive Planning Frameworks	Frameworks that emulate human reasoning—iteratively refining queries and adaptively selecting retrieval modules to optimize performance.	**CogPlanner**



HM-RAG



CogPlanner

Challenges and Outlook of RAG on complex data

- Unified Representation
  - arbitrary format data (structured, unstructured, semi-structured, multimodal)
  - preserve rich semantics and consistency
- Heterogeneous Retrieval
  - integrate, dynamic update, cross-modal search of multi-source structured knowledge
  - efficient search for large-scale sources and complex queries
- Grounded Generation
  - errors in the sources
  - hallucinations: retrieved information with the generation process
  - inconsistencies between answers and sources
- Reasoning & Planning
  - decompose complex queries and iteratively retrieve and integrate information
  - keep coherence and avoiding error propagation
- Trust & Alignment
  - no standard framework: still unverifiable citations, hallucinations, bias, and privacy concerns
  - unverifiable citations, hallucinations, bias, and privacy concerns
- Evaluation Paradigms
  - existing benchmarks largely focus on single-modality or static scenarios
  - standardized evaluation frameworks: interpretability, real-time retrieval performance, or multi-modal reasoning, accuracy, efficiency, grounding fidelity, and alignment.