Beyond Text Retrieval: A Survey of Multimodal RAG

Abstract

Retrieval-augmented generation (RAG) lets a language model answer from external documents instead of only from what it memorized during training. Multimodal RAG extends RAG to queries, evidence and outputs that include images, visually rich documents, video or audio. It inherits RAG's difficulties, such as retrieved items that are each relevant but together still lack a fact the answer needs. It also faces new ones that arise when content changes form. A page or recording is often searched through a caption, transcript or vector, and the conversion can drop the table header or sound event that a question depends on. When the output is itself an image or a sound, it must keep the details that retrieval supplied. We survey the field by following information from the stored source to the output, through four stages: representation, retrieval and selection, generation, and adaptive search and training. Across documents, video and audio, one design keeps detail at little cost: searching with a compact representation but handing the model the original page or clip. Because a system can fail at any step, evaluation should check separately whether the evidence was found, whether it was complete, and whether the output kept what it supplied. We close with open problems such as deciding which retrieved items add something new, verifying outputs, and trusting collections that change.

1 Introduction

Large language models can answer a wide range of questions, but their knowledge is fixed when training ends, and they can state false claims with confidence. *Retrieval-augmented generation* (RAG) addresses both problems by retrieving relevant documents from an external collection and generating the answer from them [1]. Because the collection can be updated or replaced, the same model can answer from current or specialized sources without retraining, and its answer can be checked against the documents it used [2]. Much of the information people ask about, however, is not stored as plain text. Scientific results sit in figures and tables, manuals in screenshots and diagrams, and meetings and lectures in recordings. Vision-language and audio-language models can now read such inputs directly, so these sources can serve as retrieved evidence too.

Retrieving such sources asks more of a system than retrieving text. Consider a photograph of a building. Asked "Which building is this?", a capable vision-language model may answer from the image alone. Asked "When was it built?", it needs a fact the photograph does not contain, so it must recognize the building and then find a page that gives the date [3]. Finding that page is only part of the work. If the date appears in a table, converting the page to text can keep the number but lose the header that says what the number is. If the evidence is a recording, a transcript can miss the sound or the moment the question asks about. Retrieved media can also help create something new instead of supporting a claim: a reference photograph can show an image generator what an uncommon object looks like before it draws one [4, 5].

As models have learned to read and generate images and audio, RAG systems have moved beyond text. Some take a photograph or a spoken question as the query and retrieve the facts it does not contain [6–8]. Others answer text questions from page images or video clips instead of from text extracted from them [9–11]. Still others retrieve reference images or sounds to guide what they generate [12, 13]. We refer to this family as *multimodal RAG*: the branch of RAG in which the query, the evidence or the output includes something other than text [14–16].

Building multimodal RAG systems is hard for four reasons. Two of them are inherited from text RAG, and multimodal sources make both harder. (i) *Relevant items do not add up to sufficient evidence.* With a photograph as the query, the pages that match best are the ones that show the building, so several can name it while none gives the date. Extra pages can also distract the generator or contradict one another [17, 18]. (ii) *Generators do not always use what they are shown.* A model can favor what it memorized over the retrieved evidence, or give a correct answer that the sources it inspected do not support [19, 20]. With media, the modality of the evidence also matters: the same content can be weighed differently against the model's own knowledge when it is shown as an image instead of as text [21].

The other two difficulties have no counterpart in text RAG, because they arise only when content changes modality. (iii) *Conversion loses information.* In text RAG, the passage that is matched is also the passage that is read (§2.1). Images, pages and recordings must instead be converted before they can be searched: descriptions, transcripts and compact vectors each keep some details and drop others. A dropped detail is lost for good unless the index still points to the original [22, 23]. (iv) *Generated media must keep what the evidence supplied.* A text answer can quote its source, but a generated image or sound must carry the retrieved detail through a transformation. It can look or sound plausible and still lack the detail that retrieval was meant to provide [5, 13]. The four difficulties arise at different points on the path from stored source to output, so we organize the survey along that path: how sources are represented (§3), how candidates are found and assembled into evidence (§4), how the generator uses that evidence (§5), and how search and training adapt these steps (§6). Because a system can fail at any of these points, we then ask how evaluation tells the failures apart (§7) and what deployed systems pay for and must withstand (§8).

Earlier surveys sort multimodal RAG systems by modality, by application, by where retrieval enters the generator, or by pipeline component [14–16, 24], and a survey of document understanding adds domain and retrieval granularity [25]. Under these categories, two systems that both search image captions fall into the same group, even if one shows the generator the original image and the other shows it only the caption. Yet what the generator can see is exactly where the two differ. We instead compare systems by what information each stage keeps and passes on. Figure 1 summarizes this organization. Our contributions are:

- A formulation that separates the stored source, the representation used for matching and the evidence the generator receives (§2). It shows when compression merely helps find a source and when it limits what the generator can recover.
- A comparison of methods by what information they keep or discard at each stage. The comparison uses the same terms for documents, video, audio and generated media, and separates the gains that come from search at inference time from those that come from training.
- An account of evaluation that separates access to evidence, completeness of the evidence set, support for the output and preservation of media attributes, with the controlled comparisons that isolate each (§7) and the open problems these distinctions reveal (§9).

2 Foundations

Multimodal RAG keeps the two parts of text RAG: a retriever that finds evidence and a generator that answers from it. But the form a source is stored in, the form that is searched and the form that is read no longer have to match (§2.1). How a system handles that split depends on two settings: the task, which determines what retrieval must add to the input (§2.2), and the scope of the search, which determines whether the source must be found or is already at hand (§2.3).

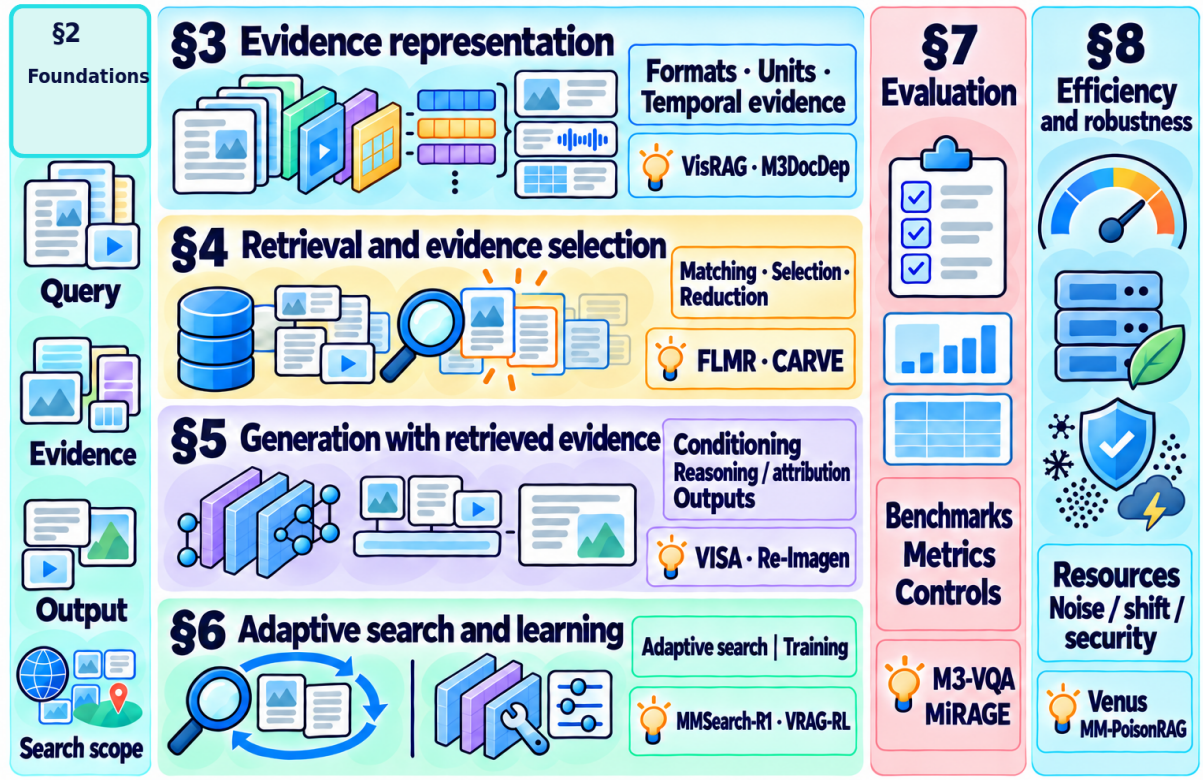


Figure 1: Organization of the survey. The foundations define the query, evidence and output and the scope of the search. Representation, selection, generation and adaptation determine how information reaches the output; evaluation, cost and robustness examine the resulting system.

2.1 From text RAG to multimodal RAG

Text RAG answers a query from an external collection in two steps. Given a query q and a collection $\mathcal{D}$, a retriever scores each document against the query with a matching function s and keeps the k best as candidates C (1a). A generator then produces the output y from the query and those candidates (1b):

$$C = \underset{d \in \mathcal{D}}{\text{top-}k}\, s(q, d), \tag{1a}$$

$$y \sim p_\theta(y \mid q, C). \tag{1b}$$

Both steps work on the same document d: the retriever matches it against the query, and the generator reads it.

With multimodal sources, what is compared and what is read often have to differ. A page image is too detailed to compare in full with every query, a recording is too long to read at once, and the meaning of a chart lies partly in the relations among its labels, axes and values. For each source $d \in \mathcal{D}$ we therefore distinguish its *stored payload* x_d, the original page, image or recording; its *matching representation* z_d, the caption, transcript or vector that is compared with the query; and the relations R_d that connect its parts in space, structure or time, or link it to other sources. The retriever now compares the query with matching representations (2a). A query-dependent mapping ϕ turns the selected candidates into the *evidence* E (2b), which the generator reads

(2c):

$$C = \underset{d \in \mathcal{D}}{\text{top-}k}\, s(q, z_d), \tag{2a}$$

$$E = \phi(q, \{(x_d, R_d) : d \in C\}), \tag{2b}$$

$$y \sim p_\theta(y \mid q, E). \tag{2c}$$

Two things change from Eq. (1): the retriever matches z_d instead of the source itself, and the generator reads E instead of the candidates as they were found. When the generator's job is to answer from that evidence, we also call it the *reader*. Evidence can consist of passages, page images, crops, frames, audio segments, structured records or learned features.

Because the matching representation and the evidence can differ, a system can search with something compact and still show the generator something detailed. A page vector can lead back to the original page, a caption to the image it describes, and a coarse video node to the original interval (Figure 3). The reverse is also possible: a compressor can hand the generator learned tokens with no path back to the pixels. Since ϕ decides what the generator can inspect, it is part of the model rather than a preprocessing step.

The running example shows two failures that this formulation tells apart. For "Which building is shown?", pages that match the building's appearance identify it, and the generator answers. For "When was it built?", the retriever returns much the same pages, since they still match the photograph best, but none gives the date: the candidate set C is incomplete, and only a history page or architect's biography would complete it. The second failure happens after the right page is found. If the date sits in a table, a mapping ϕ that converts the page to text may keep the number but drop the header that says it is the year of construction.

Query q	Candidate set C	Evidence E and output y
Which building is shown?	Pages matching the building's appearance	A page naming the building; correct answer
When was it built?	Much the same pages	No date; a history page must be added to C
When was it built? (date page found)	A page whose table gives the date	Text conversion keeps the number but drops its header

The field arrived at this formulation gradually (Figure 2). Early systems retrieved text or image–text pairs for tasks that models already handled without retrieval: knowledge passages for visual question answering, captions for image captioning, and reference images for text-to-image generation [4, 6, 26]. Work then moved along four lines. The first changed what is stored and shown to the generator, moving from text extracted from documents to page images and from whole videos to selected frames and linked intervals (§3) [9, 11, 23, 27]. The second changed how evidence is chosen, from reranking candidates by relevance to selecting them by what they contribute to the reader's answer and to the rest of the set (§4) [7, 28, 29]. The third extended the output beyond text, from retrieval-guided sound and images to answers that interleave prose with source figures (§5) [5, 12, 13, 30]. The fourth let the system decide during inference whether to search again, where to look and when to answer, and then trained those decisions with reinforcement learning (§6) [31–34].

2.2 Task formulations

The task determines what retrieval must add to the input. Retrieval can supply a fact the input lacks, recover context that is present but spread across a document, select the moments of a recording that matter, or provide material for a new output.

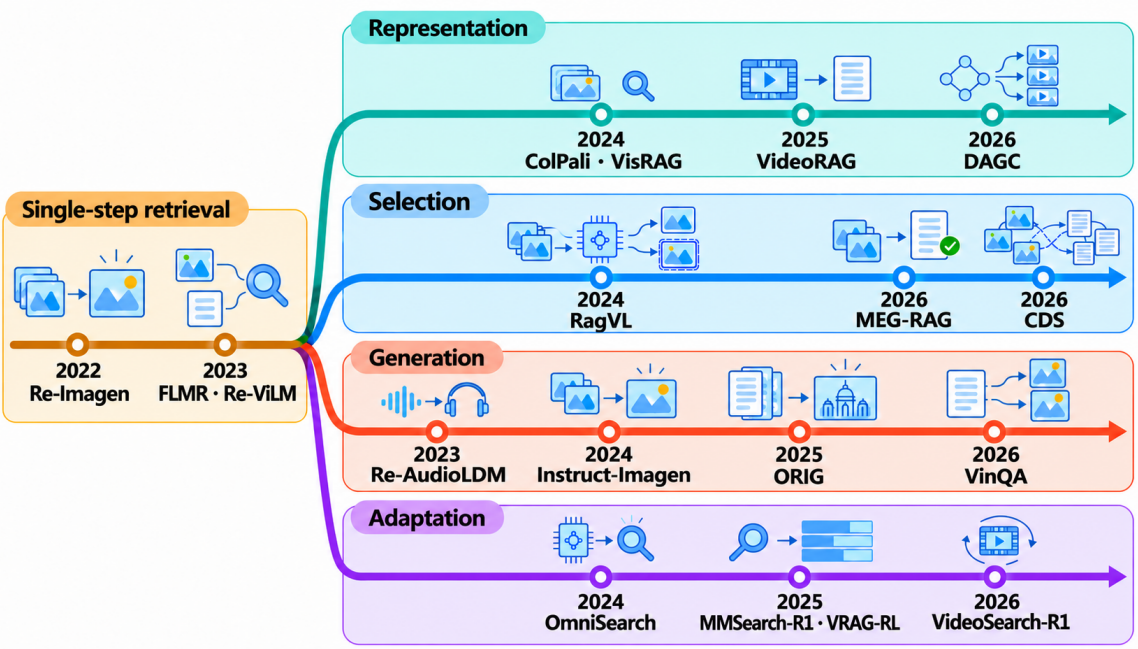


Figure 2: Multimodal RAG began with single-step retrieval and then split into four lines of work, each covered by one body section: what sources keep for matching and reading (Representation, §3), how evidence is selected (Selection, §4), what the output contains (Generation, §5), and how search adapts and is trained (Adaptation, §6) [4–7, 9–13, 23, 26–34].

Knowledge-intensive question answering When the input identifies the subject but not the answer, perception and retrieval divide the work: the model recognizes what the input shows, and retrieval supplies what it does not. The building question is of this kind, and benchmarks built on Wikipedia test the same link between recognizing an entity and finding detailed facts about it (InfoSeek [3]). Sound divides the work in the same way. A recording can mention a restaurant chain while the question asks where the chain was founded, so external text must supply the answer [35]. In these tasks a recognition error sends the search after the wrong entity, and no retriever can make up for it.

Visually rich document question answering In document question answering the answer is usually in the source, but its meaning depends on layout and on material elsewhere in the document. A plotted value needs its axis label, a table cell its row and column headers, and a figure the passage that explains it. Benchmarks test these dependencies across tables, charts and slides [36], across documents whose entities connect through charts and tables [37], and between scientific figures and the text that discusses them [38]. The domain decides which surrounding material counts: a scientific figure needs its experimental description, and a clinical decision-support system needs the guideline page that applies to the case [39]. Ranking the right page first is therefore not enough; the generator must still recover the relation the question asks about.

Temporal understanding and description Video and audio spread evidence over time, so the system must decide which moments to read as well as which source to search. A system can search a video corpus and then select frames and subtitles from the videos it finds [11], or start from a supplied video and select the frames that bear on the question instead of sampling at fixed intervals [40]. Retrieval can also supply wording instead of facts: captions of similar recordings help a model describe a new one [41]. Unlike the external facts in knowledge-intensive questions,

such guidance improves the description without adding information the recording lacks.

Nontext and interleaved generation When the output contains media, retrieved material either appears in it or shapes it. An answer about a document can insert the cited chart beside the sentence it supports, so the chosen chart, its placement and its explanation must agree [30]. A generator can instead use retrieved examples to make something new: text–audio pairs guide a generated sound [13], and retrieved facts with reference images guide a factual picture [5]. We call the first role *reuse* and the second *synthesis*. They call for different checks: reuse asks whether the source object accompanies the right claim, while synthesis asks whether the requested attributes survive the transformation.

2.3 Retrieval scope

The same task asks different things of retrieval depending on whether the source must be found or is already supplied. Searching a collection means discovering a source before choosing what in it to read; searching supplied media can go straight to the page, interval or region that matters.

External collections When the source is unknown, the system may first have to choose which collection to search. Knowledge questions search Wikipedia [3], document questions search collections of visually rich pages [37], and video questions search video corpora [11]. A system that serves all of these can route each query to the corpus whose modality and granularity suit it (UniversalRAG [42]). The choice of collection limits what can be found before ranking begins, so a routing error cannot be repaired by a better retriever.

Supplied media and combined search A supplied document or recording can be searched like a collection, but the target is a part of the source, not the source itself. Selecting the frames of one video that bear on a question is search inside a source [40]; recovering related moments across several long videos is search over a small collection [43]. Some tasks need both kinds: a question about a recording may require finding the relevant clip in a pool of audio and then looking up a fact about what it mentions [35]. These searches often share one name, video or audio retrieval, yet selecting an interval, finding the source that contains it and retrieving an external fact about its content are different searches, with different candidates and different ways to fail.

Scope is a separate question from when the system searches. A fixed pipeline chooses its evidence from the initial query; an interactive one can revise its search after reading what it found (§6), for instance by querying and navigating several parsed views of a PDF [44]. Scope also determines what an evaluation can show. A benchmark that scores only source ranking measures discovery [27], and one that supplies the paper and scores the answer measures understanding [38]; neither shows whether connecting search to generation helps, so a benchmark has to state whether retrieval is meant to discover a source, recover evidence within it, supply an absent fact or guide an output.

3 Evidence representation

Finding a source and reading it need different information. A short caption may be enough to find a chart but not to compare its values. A page vector cannot be read at all, but it can lead back to the original page. Representation therefore decides two things: what the matching representation z_d keeps for search, and what the evidence E keeps for the reader (§2.1). Table 1 shows how representative systems make these two choices, which often differ. We organize the section around three questions: what form a source takes for search and for reading (§3.1), how

Table 1: Stored payloads, matching features and generator inputs in representative systems.

System	Stored payload / unit	Matching features / unit	Generator-consumed content	Purpose / scope
VisRAG [9]	Document page images	Pooled VLM vector per page; cosine	Retrieved page images	QA / document collection
Beyond Text [22]	Extracted PDF images; text chunks	Text vectors of image summaries and chunks	Original extracted images and retrieved text	QA / industrial PDFs
M3DocRAG [10]	Document page images	ColPali token vectors per page; MaxSim	Retrieved page images	QA / supplied documents or corpus
LFRAG [45]	Layout-derived block images	Block vectors with page context and tags; MaxSim	Retrieved block images	QA / document collection
FLMR [6]	Knowledge-base text passages	Passage token vectors; multimodal query MaxSim	Retrieved passages plus image–question query	Knowledge VQA / external corpus
VideoRAG [11]	Videos and associated subtitles or ASR text	Video-level visual/text representations; similarity	Query-selected frames and associated text	QA / video corpus
DAGC [23]	Original sampled-frame video chunks	Coarse semantic graph; original-chunk mappings	Recovered, refined chunks and intermediate reasoning	QA / supplied long video
PlanRAG-Audio [46]	Time-aligned transcript, speaker, event and emotion records	Record fields; planned SQL filters and temporal joins	Retrieved stream records and output schema	QA / supplied recording
VoxRAG [47]	Podcast speech segments and aligned transcripts	CLAP vector per audio segment; cosine	Retrieved segment transcripts and transcribed query	Spoken QA / podcast collection
Re-Imagen [4]	Image–text pairs	Caption BM25 or CLIP similarity per pair	Encoded image–text neighbors via cross-attention	Image synthesis / external exemplars
Re-AudioLDM [13]	Text–audio pairs	CLAP similarity per candidate pair	T5 text and AudioMAE audio features via cross-attention	Audio synthesis / dataset exemplars

VLM: vision-language model; ASR: automatic speech recognition; CLAP: contrastive language–audio pretraining.

large an evidence unit is and which relations inside it survive (§3.2), and how recordings keep each event tied to its moment and to the other tracks (§3.3).

3.1 Representation formats

Systems choose among three forms: convert the source to text, keep the original media, or search a derived index that points back to the original. The choice fixes what information is available when the system searches and what is available when it reads.

Text conversion Converting media to text lets existing text retrievers and readers process it, but it decides in advance which parts of the source matter. *Optical character recognition* (OCR) keeps the visible words, captions describe the visual content, and transcription records the spoken words; none keeps the layout relations or nonverbal cues a question may turn on. Retrieval comparisons consistently favor keeping visual information. On visually rich documents, descriptions written by a vision-language model retrieve better than raw OCR text, but retrievers that read the page images do better still [48]. Page-image retrieval also outperforms extraction pipelines on the ViDoRe benchmark (ColPali [27]), and a comparison of single- and multi-vector

retrieval finds the largest gains when the vectors come from page images [49]. Conversion is a sound choice when the words carry the answer and a costly one when the layout does.

Retained visual content Returning the original page leaves it to the reader to decide what matters. The reader is better placed to decide than an indexing step, because it has seen the question. Page-image systems search with a pooled vector or compressed document tokens but hand the reader the page itself (e.g., VisRAG and VDocRAG [9, 37]), while multi-vector retrievers keep a separate feature for each image patch [27]. Reading the page instead of its extracted text improves answers with both open and proprietary generators, most of all on visually dense content [9]. The reader must still be able to perceive the page: on degraded scans unlike its training data, a text pipeline built on a much larger OCR model generalizes better than a visual one [50]. Keeping the original pays off when the layout carries the answer and the reader can see it clearly.

Derived indexes with original payloads A derived index searches with a compact surrogate, such as a summary of an image, and returns the original. It works only if each hit stays linked to the payload that answers the question. The surrogate and the payload need not share a modality, and the link can run either way: image features and captions can also serve as the query for finding text passages, which the reader then reads as text [6]. On industrial PDFs, one system stores image summaries next to text chunks and returns the image behind any summary that matches; returning images this way outperforms indexing them separately with CLIP [22]. On financial documents, by contrast, embedding pages directly with a multimodal model ranks better and keeps numbers more accurate than searching image summaries [51], whose descriptions can omit the very values a question asks about. Summaries suit sources a sentence can describe; dense numerical pages need a representation that keeps the numbers.

Audio representations Audio makes the split between matching and reading especially clear, because speech and other sound are not interchangeable. A retriever can match audio directly and keep acoustic information that a transcript drops [8], yet still hand the reader only text: one podcast system matches speech segments by their audio embeddings but gives a text reader the aligned transcripts (VoxRAG [47]). That reader can answer from what was said, but not from a sound nobody transcribed, even though the retriever matched the waveform that contains it. What the reader can hear is decided by the evidence mapping, not by the retriever.

3.2 Evidence structure

A detail is interpretable only together with its surroundings: a value needs its header, a figure its caption, a callout the object it points to. Choosing an evidence unit is therefore a trade-off, since a smaller crop spends more of the reader's input on the target but can cut away what explains it. We compare units within a page, the resolution spent on each, units that span pages, and tables and charts, whose relations are the most explicit.

Page and region units Whole pages keep charts, labels and prose together [10], but they spend much of the reader's input on material the question does not need. Blocks derived from the page layout focus that input on the relevant region, and they work best when they keep some page context. One method merges compatible layout regions, encodes each block together with its page, and returns the block image. It beats uniform partitioning and improves answers while using a fraction of the tokens that page-level methods need (LFRAG [45]). Smaller units help when they keep their surroundings; arbitrary crops do not.

Resolution and coverage Even with good units, a fixed image budget forces a choice between seeing a few pages in detail and many pages coarsely. Construction drawings make the trade-off concrete, because reading a detail callout means following a thin leader line to the object it labels. With the number of images held fixed, adding overlapping high-resolution tiles to a page overview improves decisions when the same projects are checked repeatedly, but makes no reliable difference on new projects [52]. Each tile takes the place of another page, so whether detail or coverage matters more depends on whether the question turns on one page or on several.

Dependency-aware chunks Page boundaries often fall across the relation a question needs: a caption on the next page, a table that continues, an image referred to from elsewhere. One response is to build chunks that keep these links. M3DocDep [53] reconstructs a dependency tree for the whole document and chunks along its hierarchy, page spans and text–image links; given the same blocks, its chunks improve both ranking and answers over those of other chunkers. The other response is to make the links searchable and follow them only when a query needs them. Graphs can encode containment, reading order and layout [54], connect segments through the questions they answer [55], link text to images [56], or link concepts to the media that support them [57]. Assembled chunks pay the context cost on every query; a graph pays it only when a relation is followed, at the price of a search over the graph.

Tables and charts In tables and charts a value means nothing without the structure around it. A cell needs its row and column headers and often a unit; a plotted value depends on axes, legends and the relations among the entities shown. Making that structure explicit helps both reading and search. In the running example, it is what keeps the year on the history page attached to the header that says it is the year of construction. Given the right table, a reader answers more accurately when values and headers are encoded as nested JSON than when they are flattened to Markdown [58]. Attaching headers to cell vectors improves ranking over linearized tables [59], and representing chart entities and their relations as a graph lets questions chain evidence across charts [60]. A table inside a page image holds the same structure, but the reader must recover it unaided; making the structure explicit removes that step.

3.3 Temporal evidence

Recordings add time, a relation no single frame or transcript line expresses. A sentence, frame or sound event is useful only if its moment, its neighbors and its alignment with the other tracks stay available to the retriever and the reader. Systems keep this information by converting recordings to timestamped text, by keeping visual and textual tracks side by side, by indexing context and nonverbal events explicitly, or by searching a coarse index that leads back to the original footage.

Temporal conversion Converting a recording to timestamped text makes it searchable with text tools, but only for what the text records. A transcript keeps an utterance without its delivery, and a single frame shows an object without the sequence that makes it an action. Within those limits conversion is cheap and effective. Combining visual captions with transcripts and timestamps finds relevant videos within a shorter candidate list than cross-modal embeddings do [61]. Replacing frames with descriptions in the same system cuts estimated token use by more than an order of magnitude. Video libraries use a similar interface, built on speech and visual metadata tied to particular moments [62]. Conversion pays off when the events a question asks about are ones a description would mention.

Visual sequences and synchronized tracks Keeping visual and textual tracks side by side avoids committing to one description of every event. Across a video corpus, combining visual and text representations retrieves better than either alone, and reading selected frames improves answers over text-only baselines (VideoRAG [11]). Within one video, adding OCR text, speech transcripts and detected objects to the frames gives small but consistent gains across readers [63]. The best form can even change from one segment to the next, so some systems store several views of each segment and choose among them at query time (§4.2). Relations among clips in space and time [64] and separate frame and audio memories [65] keep further structure for later search. A relation that no track records, such as who was speaking when an object appeared, is beyond the reach of later selection.

Context and acoustic events Time also carries reference. In a meeting recording, "Yes, let's go with that" records an utterance but not which proposal was accepted; the preceding discussion does. Resolving pronouns and ellipsis before indexing timestamped actions, utterances and conversations makes that context searchable [66]. Other information is never spoken. Storing detected sound events with their intervals lets a system answer questions about them, but only about the event classes the detector recognized during ingestion [67]. Records can combine these streams: with time-aligned speech, speaker, emotion and sound-event records, a question about when the music starts can be answered from an event record although no transcript mentions music (PlanRAG-Audio [46]). Compared with giving the model the whole recording, these records greatly improve speaker counting and event ordering for a strong multimodal reader, though they help a text-only reader count speakers only slightly. Structured records thus expose temporal facts that a transcript hides.

Coarse indexes with original chunks A coarse index can make a long recording cheap to search without making it coarse to read. Merging visually similar neighboring segments into the nodes of a semantic graph shrinks the search, and a mapping from each node back to its original frames lets the system recover and refine those frames for the reader (DAGC [23]); reading only the coarse nodes gives worse answers. Figure 3 shows the same pattern in documents and video: an image summary or a graph node finds the content, and the original image or chunk is what the reader sees.

4 Retrieval and evidence selection

Retrieval decides which sources the reader will see, and a source can be relevant without being useful. In the running example, pages about the building are relevant to "When was it built?" but useless if none gives the date, and a second page that repeats the building's name adds nothing to the first. An item's *relevance* is how closely it relates to the query; its *utility* is how much it helps the reader produce a correct, supported output. Retrieval methods differ in which of the two they pursue and in whether they judge items one at a time or as a set. We organize this section around three steps: matching the query against sources in other modalities to find candidates (§4.1), choosing which candidates together form the evidence (§4.2), and shrinking that evidence to what the reader can afford (§4.3).

4.1 Cross-modal matching

Matching decides which sources enter consideration before any reader inspects them. Methods differ in three choices: whether all modalities share one score space, whether local correspondences survive when a source is pooled into a vector, and whether exact words or document structure can recover facts that semantic similarity misses.

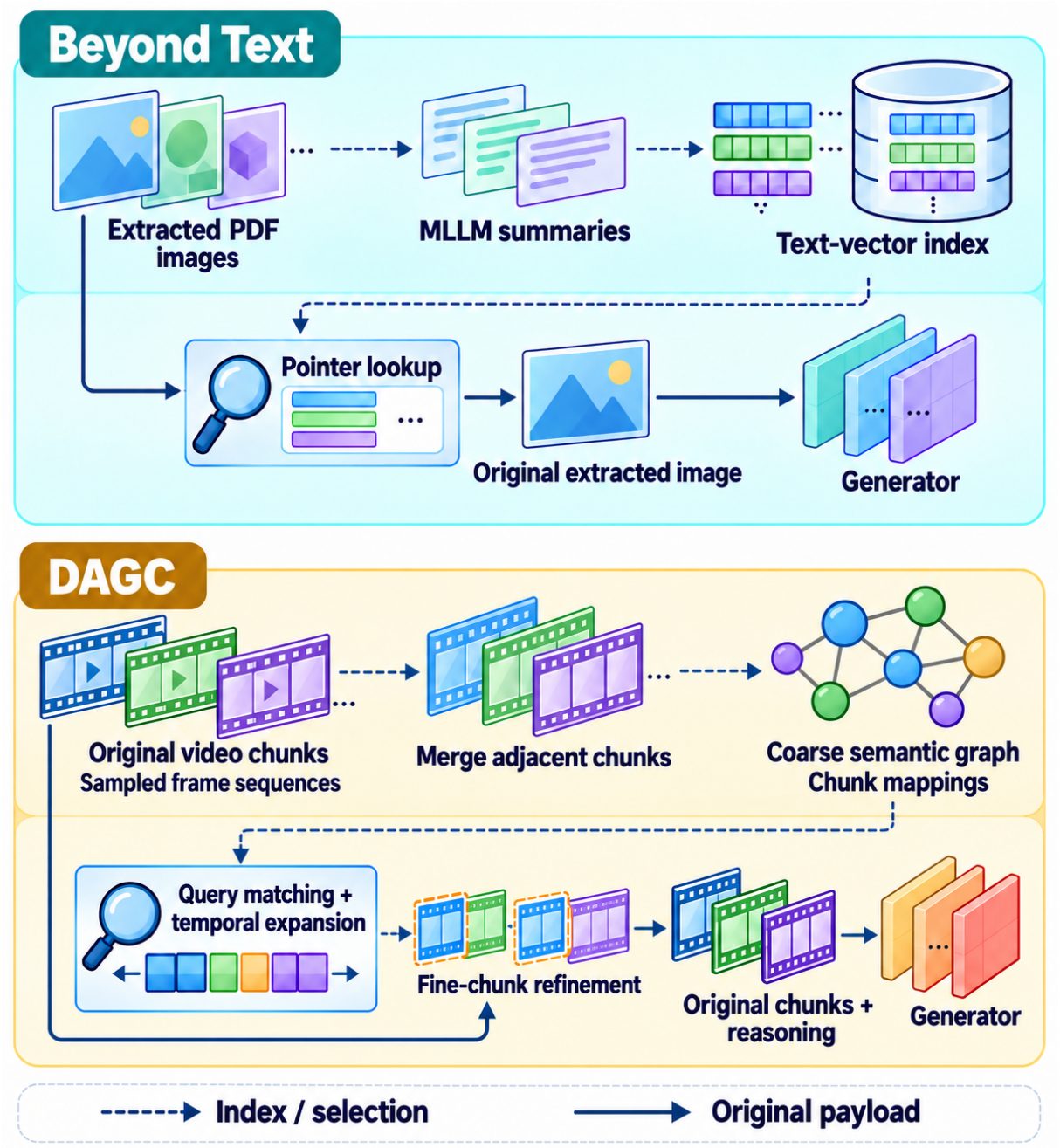


Figure 3: Compact search representations linked to original media. Indexes of image summaries and coarse video graphs locate content and link back to the original images and fine-grained chunks for reading [22, 23].

Global embeddings and modality alignment Encoding queries and sources independently into one vector space makes search fast, but a text item and an image can receive scores that are not comparable. A unified embedding can favor text even when the answer needs an image; routing each query to a separate index for the right modality and granularity avoids that bias (UniversalRAG [42]). A separate problem is representing items that mix images and text. A retriever can learn such representations, especially with feedback on the quality of downstream answers [68]. A match can also cross modalities: matching a query image against image–text pairs ranks the right item first more often than matching it against images alone [69]. Routing decides where to search; mixed-modal representation decides what an item means when it is matched.

Fine-grained late interaction A single pooled vector compresses every correspondence between a query and a source into one number. *Late interaction* instead keeps one vector per token or image patch and sums the best local matches when scoring. In knowledge-based visual question answering, one retriever maps global and region-level image features into a text retriever's space, adds them to the question's tokens, and scores each query token against its best-matching passage token (FLMR [6]). For the building question, image tokens can then match passages about the building while the question's tokens match a passage about its construction. Region features improve this multi-vector retriever but hurt a single-vector one, and the multi-vector retriever with region features also improves answers from the same reader. Local features help only when the scoring keeps their correspondences separate.

Other late-interaction settings The same design extends to queries of any modality [70], to models that offer single- and multi-vector retrieval in one aligned space [71], and to video, where token matching uses features with spatial and temporal context [72]. These retrievers report ranking gains; the knowledge-VQA retriever above also shows that such gains carry through to answers. Late interaction buys finer matching with more storage and more comparisons per query, a cost that pooled candidate search followed by multi-vector reranking can reduce (§8.1).

Lexical and hybrid search Semantic similarity can miss the exact name or metric that distinguishes one number from another. On financial documents that mix text and tables, BM25 outperforms the dense retriever tested, and fusing sparse and dense results before a neural reranker retrieves best [73]; expanding queries adds little for precise numerical questions. When similar reports differ mainly in their numbers, exact word matches remain the most reliable signal.

Hierarchical routing Document structure lets retrieval reach material whose own similarity to the query is weak, such as the table under a matching section heading. One approach first chooses a document and then ranks its sections by their best-matching unit. Within the reader's budget, it assembles each chosen section together with the headings above it and the figures and tables beside it (HiKEY [74]). The table enters the context because of its place in the hierarchy, not its own score.

Graph retrieval A graph lets a system follow links beyond the hierarchy of one document. The system retrieves the entities that match the query and then follows relations to associated evidence. It can prune a graph of entities and text–image links to a small subgraph connecting them [56], route through the graph's global structure before matching locally [75], or expand and prune page and element nodes into a query-specific subgraph [54]. In each case, evidence linked to a matching item can be retrieved even when it does not match the query itself.

Audio-origin queries When the query begins as sound, even forming it requires perception. In web search that starts from audio clues and leads to text, image and video sources, recognizing the entity in the audio and turning it into a search query are the main bottlenecks [76]. Both bottlenecks lie before any retriever receives a useful query, so a better retriever cannot remove them.

4.2 Evidence-set selection

Ranking candidates one at a time does not ensure that the selected set supports the answer: the top pages may all name the building and none give its date. Selection methods differ in what they judge (relevance to the query, utility to a particular reader, or what an item adds to the

others) and in whether they judge items alone or together. We start with methods that judge relevance, move to methods that judge utility and then whole sets, and end with deciding when to leave an item out.

Multimodal relevance reranking The simplest step beyond first-stage matching is to rerank a small candidate pool with a model that sees more of each candidate. Cross-attention between query and candidate improves knowledge-intensive visual question answering [77], a multimodal language model trained as a reranker separates coarse from fine mismatches (e.g., RagVL [7]), and a zero-shot listwise reranker ranks the right item first more often [69]. A reranker helps only when it can see the content that matters. Reranking raw text snippets with a text-only model makes a page-image retriever's ranking worse, whereas adding visual captions, section paths and entity annotations to each candidate nearly doubles the accuracy of evidence selection over retrieval alone [78]. The reranker needs its own view of the evidence, even when the reader will later see the page image.

Cross-modal score calibration Reranking across modalities also needs comparable scores. When speech and text passages are retrieved together without normalizing their scores, text results crowd out speech results. Normalizing restores speech retrieval, an audio–language reranker further improves answers over the mixed sources, and scoring shorter audio windows locates the answer within long speech passages [79]. Retrieval over mixed modalities therefore needs explicit calibration before a reranker can compare candidates fairly.

Presentation-aware selection Selection can also decide how an interval is shown. An object may be clearest in a single frame, while an action needs the whole clip. CARVE [80] stores frame- and clip-level views of both the visual and the textual track, since no single view is best everywhere. It reranks each pair of interval and view, so every selected interval reaches the reader in the view that made it match. Passing each interval's best view to the reader improves answers over choosing one view for the whole query, because the choice of form is carried into generation instead of being fixed in advance.

Reader-oriented utility A source can satisfy a reranker and still add nothing for the reader that will use it. Two methods measure utility with the deployed reader itself. One labels each document by whether a frozen vision-language model answers correctly when given it, and trains the reranker on the resulting preferences, which outperform preferences mined from ranks or answer likelihood [81]. The other measures how much adding a document raises the probability of the informative tokens in the reference answer, ignoring common words, and trains a reranker on that signal (MEG-RAG [28]); it improves question answering, captioning and fact verification over semantic reranking. Both move selection from relevance toward utility. Both still score documents one at a time, however, so neither can tell that a second page naming the building adds nothing once the first is in the set.

Complementary sets Set selection judges each candidate against the items already chosen. A set can fail because its items repeat one another or because it holds fragments that cannot be connected, and the two failures call for different objectives. One method favors items that are relevant to the query but dissimilar to one another, which improves document question answering over taking the top items by similarity [29]. The same choice appears in synthesis, where references are chosen to cover different requested aspects of an image [82]. Another method instead keeps fragments together: it routes a fixed budget through connected text, image and table nodes and penalizes incoherent jumps, which helps most on questions over split tables

and slides [83]. Diversity suits questions that need several independent facts; connectivity suits questions whose answer is spread across the pieces of one object.

Selective admission More evidence can lower answer quality once the reader cannot separate support from conflict. When a diffusion-based vision-language reader is given more pages, the answer page is more likely to be included, but the answers get worse as conflicting pages enter its parallel decoding. Adding a second image only when an entropy test says it helps beats the best fixed context size [84]. Fact-checking shows the same pattern: adding images indiscriminately hurts accuracy relative to text-only evidence, while deciding for each claim whether visual evidence is needed recovers it [85]. Admission should weigh what an item adds against whether this reader can use it alongside what it already has.

4.3 Context reduction

Even a well-chosen evidence set can be too large or too detailed for the reader. Reduction methods shrink it in three ways: extracting the relevant part of each source, cutting the number of sources, and compressing what remains. Each saves reader input at the risk of discarding the context that explains a detail. Narrowing the query before search, by cropping the query image [86] or masking objects in it [87], changes what is sought; the methods below change what the reader sees.

Selective extraction Parsing at query time can recover the relevant part of a source without transcribing every page in advance. A parser can locate, crop, rotate and read regions of retrieved pages, returning both the recognized text and the image crops [88]. Document metadata can give direct access to pages, sections, tables and figures [89], and graphs that link chunks to the questions they answer can guide which chunks to combine [55]. Finer methods score individual sentences and image regions by their contribution to generation [90]. The risk is losing the surroundings: retrieving image patches helps with tiny targets but can do worse than viewing the whole image for larger objects, because the patches split the object's spatial context [91]. Extraction works when the unit it returns still contains what the detail depends on.

Page-set reduction and consolidation The number of pages a question needs varies, so a fixed k either wastes the reader's input or cuts off support. Comparing the retrieved pages' similarities to the query can set a cutoff for each question, reducing the pages used while keeping or improving answers [92]. Long searches raise a related problem, because the record of what was found grows with every step. Keeping compact notes with pointers to their sources during the search, and reloading the original images only for the final answer, saves context without giving up visual detail where it matters [93].

Compression and token allocation Text RAG offers several ways to drop unneeded context: selecting supporting sentences [94], learning extractive or abstractive summaries [95], or keeping question-relevant tokens and reordering them [96]. Multimodal readers allow another option: changing how the text is encoded. One method renders retrieved passages as images and gives the relevant ones more resolution; in question answering, it uses roughly an eighth of the baseline's text tokens while improving accuracy [97]. Compression saves the most when it spends resolution where the question needs it rather than shrinking every passage alike.

5 Generation with retrieved evidence

Retrieval helps only if the generator can interpret what it is given and carry what matters into its output. Evidence plays different roles: a passage can supply a missing fact, a reference case can guide a medical comparison, and a reference image or sound can shape a synthesized artifact. Each role needs its own interface and its own check. We organize this section around three questions: how evidence enters the model (§5.1), how the model connects its statements to their support (§5.2), and what must survive when the output itself contains media (§5.3).

5.1 Evidence conditioning

Evidence can enter the generator as text in the prompt, as demonstrations of the task, as media in a multimodal context, or as learned features and latent tokens inside the model. In that order, the evidence moves further from its original form, taking up less context but leaving less for the reader to inspect directly.

Text context and demonstrations Putting retrieved text into the prompt reuses a language model's existing reading ability. A multimodal model can keep the query image beside the retrieved passages [98], while a text-only model can receive images as detailed descriptions, together with formatted tables and worked examples of the task [99]. In the second setting, two choices matter. The first is how detailed the image descriptions are: richer descriptions improve answers to image questions over ordinary captions, because they change what the reader knows. The second is how the reader is prompted, which changes what it does with that knowledge. Step-by-step reasoning helps questions that combine tables and passages but hurts simple image or text questions, so choosing the prompt by question type beats either fixed strategy. A question that must identify entities in an image and a passage before comparing table values needs the reasoning, while a recognition question does not.

Retrieved demonstrations Retrieved demonstrations adapt a frozen model to a task without training. In medical imaging, examples matched to the query improve classification and visual question answering over no examples, while fixed or random examples can do worse than none [100]. Retrieved reference reports need the same care. A report retrieved because its radiograph looks similar may assert a finding the query image lacks. Aligning retrieval on clinical entities and their relations instead improves the clinical accuracy of generated reports over both the same model without retrieval and other multimodal retrievers [101]. A useful reference must match the clinical facts and not only the phrasing, or the reader may copy the wrong finding.

Multimodal context Returning the reference image along with its report lets the reader compare the two patients directly. Training the reader to associate images with text and to attend to the target image in this format improves report accuracy over text-only references [102]. The reader must then tell which parts of the reference report transfer to the new image and which describe differences between the two images, a multi-image skill that training on single images does not provide.

Attention, features and latent tokens Learned conditioning moves the evidence inside the model. Query features can attend to retrieved biomedical image–text cases [103], and question-guided visual queries can feed a decoder trained to pick out the useful retrieved knowledge [104]. Captioning shows most clearly the choice between keeping retrieved language as text and turning it into a learned visual signal. One captioner cross-attends separately to encoded retrieved captions and to the query image, outperforming a larger model that does not retrieve (Re-ViLM

[26]). Another keeps retrieved captions as text prompts and also maps their meaning into visual prompts aligned with image patches, improving captions with fewer trainable parameters [105]. Text that stays text remains open to inspection; a learned visual signal is not.

Latent conditioning Compressing each retrieved item into a few continuous tokens hands the decision about what to keep from the reader to a compressor. Projecting each item's compressed token into a frozen model's embedding space improves answers to image-grounded web questions while using far fewer context tokens than a text-embedding baseline [106]. The summary indexes of Section 3.1 keep a link to the original image; here nothing leads back to the pixels, so the reader sees only what the compressor kept. Training the compressor to reconstruct each item's semantic embedding helps it keep content, and removing that objective hurts both retrieval and generation. Latent conditioning is only as good as the compressor's training.

5.2 Evidence grounding

Once evidence is in the context, the reader must decide how far to trust it, combine sources that each hold part of the answer, and show where each claim comes from. A final check can then ask whether the answer is supported.

Reconciling memory and sources Retrieved sources can disagree with the query image, with each other or with what the model memorized. Benchmarks that separate conflicts between context and memory from conflicts among sources find that the evaluated multimodal models tend to prefer their internal knowledge [19]. Making the disagreement explicit helps. One method has the model answer once from memory and once from the evidence, then reconciles the two answers. Without any training this improves slightly over a single pass, and fine-tuning on inconsistent cases improves it further [107]. Asking for both answers exposes conflicts that a single pass would resolve silently in favor of memory.

Composing evidence Consistent sources may still need an explicit operation to connect them. Questions whose evidence chains across pages, or that fan out to several sources, expose failures in comparison and table alignment [108]. One remedy decomposes the question, checks each retrieved page and keeps intermediate findings before assembling the answer [109]; another fuses textual and visual evidence before explaining a fact-checking verdict [110]. Both make composition an explicit step instead of leaving the reader to connect the pages itself.

Source references *Attribution* links a generated statement to the source that supports it. A model can generate the answer together with a document identifier and a box around the supporting region (VISA [111]), or attach page indexes and boxes to each reasoning step [112]. It can also quote the supporting language and resolve the quotation to a region by parsing and retrieval. Quoting greatly improves evidence recall and roughly halves attribution errors across open models while leaving answer quality largely unchanged [113]. Since the answers barely change, the models mostly knew them already. Quoting fixes their ability to say where each answer came from. A box makes the source inspectable, but whether the boxed region supports the statement is still a semantic question.

Verification Verification asks whether a cited source supports the statement. What a check can establish depends on what it sees. One check compares the answers produced with and without the retrieved context and breaks the result into verification questions, all with the model that answered; it tests consistency, not support [114]. An independent inspector that receives the candidate video evidence and the question, but not the planner's exploratory reasoning, can judge

whether the evidence suffices and write the final answer (VideoSEAL [20]). A policy trained with reinforcement learning can instead judge each snippet's usefulness within its own reasoning [115]. These checks differ in their access to the original evidence and in their independence from the reasoning they check. Only a check that compares the claim with the source can establish support; agreement among answers cannot.

5.3 Multimodal outputs

Putting media into the output adds obligations that scores for text answers miss. Reusing a source figure keeps its identity but leaves the generator to place and explain it; synthesizing a new image or sound transforms the source and must carry the requested attributes through the transformation.

Interleaved source reuse A generator that interleaves source visuals with prose must choose both the element and its place in the explanation. Document identifiers let a model insert cited tables, charts and figures into long-form answers [30], and corpus-level tasks require choosing and ordering source images [116]. The form in which the source reaches the model affects how easily it learns to choose and place visuals. Before fine-tuning, models do better with OCR text plus separately cropped visuals than with full pages; after training on mixed text-and-image answers, reading full pages nearly matches reading parsed elements [30]. Both inputs keep the elements citable, and with training the model can read the surrounding prose from pixels about as well as from extracted text.

Placement and source context The context a visual came from still matters after it is selected. Supplying each image's heading path, position and nearby text improves how well images match the surrounding text over generation that ignores the document's hierarchy [117]. A photograph taken after the building's renovation can show the right building and still mislead beside a sentence about its original design. Long reports extend the problem across sections. Checklists can organize the report and the placement of its citations [118], marking which images were retrieved and which are charts generated from the analysis clarifies each visual's role [119], and linking source figures to the statements and sections they support keeps originals beside generated plots [120]. In each case the visual must stay attached to the claim it helps explain.

Image synthesis Reference images can supply the appearance of an uncommon entity while a text prompt asks for a new composition. A generator can retrieve image–text pairs and, during sampling, balance agreement with the prompt against agreement with the references (Re-Imagen [4]). A factual image can also need attributes and relations that no single reference shows, such as the relative size of two real objects never photographed together. One system supplies them by searching the open web for text and images and extracting detailed descriptions as well as visual references before generation [5]. The two kinds of evidence constrain different parts of the scene: reference images fix appearance, and retrieved text fixes the facts the image must depict.

Identity and style References can also specify a subject, a style or a structural control, and these goals can compete. A model trained with retrieved references and then tuned on multimodal instructions handles all three without tuning at test time, including unseen combinations of subject and style (Instruct-Imagen [12]). On subject-driven generation it follows the prompt slightly better than per-subject fine-tuning but preserves the subject less faithfully. References can also correct an image after generation, fixing historical colors, identities and implausible details, although this has so far been shown only qualitatively [121]. Following the prompt and

preserving an identity are separate goals, and progress on one does not imply progress on the other.

Diversity and distributional fidelity Diversity across many images is a different goal from preserving one identity. One method conditions a frozen diffusion model on features of retrieved images of people, with instructions to transfer demographic attributes. Compared with adding text alone, this increases demographic diversity and keeps images closer to the distribution of real images without hurting agreement with the prompt, and both the references and the instructions contribute [122]. The generated images remain less diverse than the references, since generation changes the distribution it receives. Staying close to the distribution of real images is *distributional fidelity*, a property of many images together, whereas preserving an identity concerns one particular subject. Factual attributes, identity, prompt agreement and diversity each need their own measure, and visual similarity to a reference does not show where a depicted fact came from.

Audio synthesis Retrieved sound can carry an event that the prompt alone fails to produce, typically a rare sound in a busy scene. One generator feeds both the text and the acoustic features of retrieved text–audio pairs into a diffusion model's cross-attention, which improves the generation of rare events; acoustic features alone help little (Re-AudioLDM [13]). A different design extracts information from retrieved audio with an encoder conditioned on the target text, so the references need no captions; its best zero-shot setting lowers the Fréchet Audio Distance (§7.2) by about a fifth relative to the same model without retrieval, and its retrieval choices can trade agreement with the prompt for distributional fidelity [123]. Without captions it can also draw on a large unlabeled collection, and in a comparison that changes size and labels together, the larger unlabeled collection works better. Retrieval thus lets audio generation use sounds that no one has described in words.

Initialization from retrieved sound Retrieved sound can also be the starting point of generation. To reconstruct heard audio from brain recordings, one method partially noises a retrieved sample and starts a frozen diffusion or flow model from that intermediate point [124]. Compared with generating from scratch, starting from the retrieved sample more than doubles how often a reconstruction can be matched to the original stimulus, approaching retrieval alone, and brings the output closer to real audio. An autoregressive generator used as a control gains nothing comparable, which ties the benefit to the initialization. Replaying the retrieved sample is still more realistic, since synthesis alters it. Across images and audio, perceptual quality, requested content, faithfulness to the reference and novelty can move in different directions, so evaluation must check what content generation carried through as well as how good the result looks or sounds.

6 Adaptive search and learning

An initial search can identify the building in the photograph without finding its construction date. The system can then search again with the building's name, look more closely at a page it already has, or answer from what it has. *Adaptive search* lets what the system observes decide its next action during inference (§6.1); training changes how the system matches sources, reads them or chooses actions (§6.2). The two choices are independent: a fixed workflow can contain trained components, and an interactive one can run on frozen models (Figure 4).

6.1 Adaptive search

Systems differ in how much their observations change what they do next. Some fix their stages, or plan the whole sequence, before seeing any results; others choose each search, inspection or answer from what earlier steps returned. We move from the first kind to the second, then to recordings, where actions also decide how closely to look, and end with evidence on when feedback actually helps.

Prescribed stages and one-time plans Dividing work among agents does not by itself make a system adaptive. Text and page-image retrieval can run in parallel before general, critical, modality-specific and summarizing agents process the results in fixed stages (MDocAgent [125]). A planner can instead look at the image and question once and commit to a sequence of image searches, text searches or a direct answer [126]. Later steps may take earlier results as input, but the sequence is never revised. In that study, planning once reduces redundant searches and improves accuracy over iterative planning baselines, so committing early pays off when the first plan is adequate.

Feedback-driven decomposition Feedback matters when the next information need depends on an earlier result. In the running example, identifying the building supplies the name that the date query needs. A planner can choose new subquestions and tools from what a solver returned for the previous ones (OmniSearch [31]). For a question that compares the construction dates of two buildings, two separate searches work better than one query that packs in both buildings and the comparison. Writing each of those searches is itself hard: models struggle to turn what a user asks about an image into a good web query [127]. Inside documents, an inspector can reject retrieved evidence and send feedback, along with the history of what has been explored, back to a seeker [128]. A fixed succession of agents passes each result forward unchanged; such feedback instead changes which evidence is acquired next.

Search and perception actions Missing information may require another source or a better view of one already found. A policy can choose at each step among searching the web with the whole image, searching text with a generated query, and answering, including answering without any search (MMSearch-R1 [32]). Region inspection adds a second kind of action: the policy selects a region, which is cropped from the original page image, re-encoded and added to its context (VRAG-RL [33]). Search can find the history page that gives the building's date; cropping can make the table on that page readable. A small router trained on the model's own signals can also choose between viewing the whole image and zooming in, which avoids repeated patch retrieval when the whole image suffices [91]. Which action helps depends on why the current evidence has not yet produced an answer.

Temporal search For recordings, actions control both where to look and how closely. Within a long supplied video, a policy can choose from its interaction history among browsing the whole video, searching for clips and inspecting individual frames [129]. Searching external video adds a check on the source: after rejecting a retrieved video, a policy can search again with continuous query tokens instead of writing a new text query, which saves generated tokens [34]. For audio, a planner can search speech, speaker, emotion and sound-event records under temporal constraints [46]. Because each action exposes a different amount of source detail, the number of calls says little about how much evidence a policy actually saw.

Limits of feedback Feedback helps only if the policy can tell which observations warrant another action. In a held-out routing comparison, adding features of the retrieval state to a router

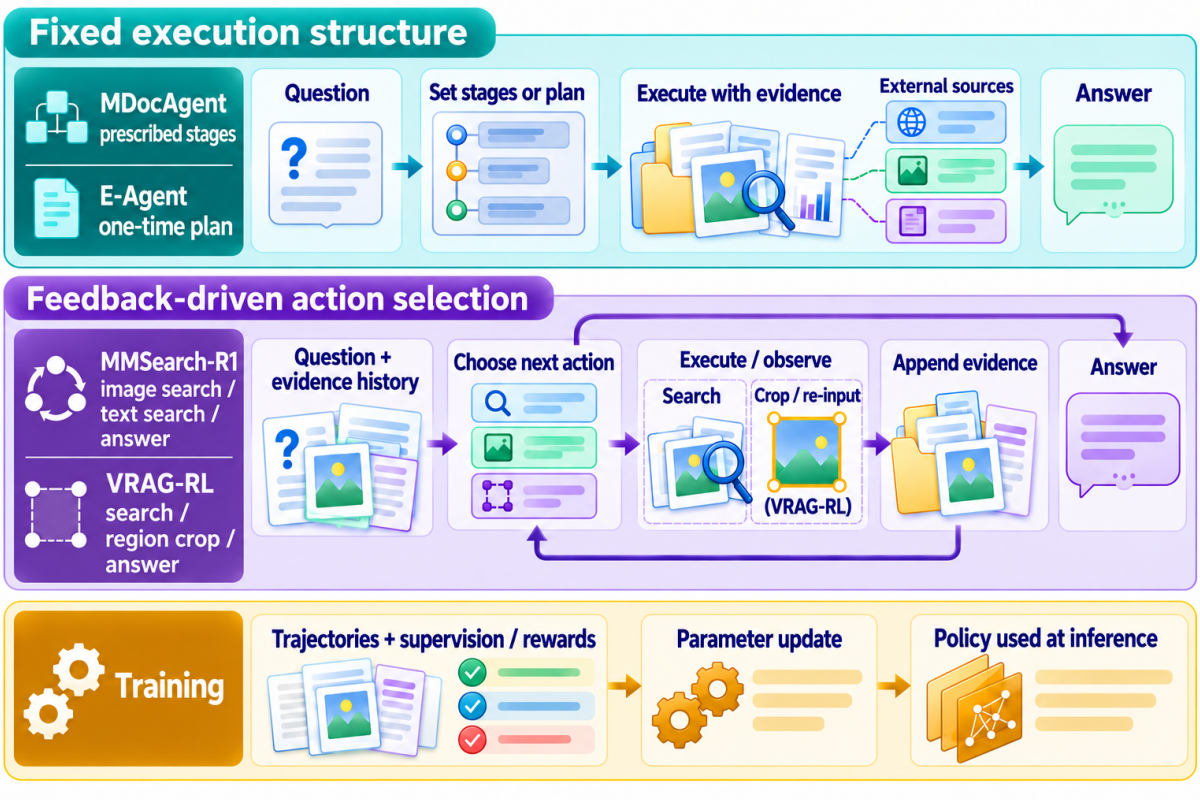


Figure 4: Evidence acquisition with prescribed stages or feedback-driven actions. Whether training updates a component is a separate choice from whether observations change the next action [32, 33, 125, 126].

that otherwise sees only the query gives no reliable improvement [130]. Adaptive search creates the opportunity to correct a search; whether a system takes it depends on how its decisions are learned.

6.2 Training strategies

What training teaches depends on what it supervises. Query–source pairs teach ranking; answers produced with evidence teach reading; preferences between responses teach which outputs, or which reliance on evidence, to favor; and rewards over whole trajectories teach which actions to take. Training also decides which components change. A frozen reader gives a selector a stable consumer to learn for, while updating the reader changes what it can do with the same context.

Retriever alignment Training can first change what counts as a match. A knowledge retriever can be trained in stages, first mapping visual features into a text retriever's space, then training on knowledge retrieval, and finally training on a mixture of related tasks, each with query–passage pairs and negatives; the knowledge-retrieval stage improves final retrieval (PreFLMR [70]). A document retriever instead first aligns page images with text, through contrastive learning and text reconstruction, and then trains on question–document pairs [37]. In both, training teaches the retriever to rank a misleading source below a useful one before the reader ever sees it.

Reader alignment The reader can be trained separately. A document system can give its retriever and generator separate adapters and train the generator on the images its own retriever returns [37]; a biomedical system can pretrain on image–text data and then train answering with similar cases supplied by frozen retrieval encoders [103]. Teaching a retriever to rank a

misleading source low and teaching a reader to ignore one already in its context are different learning problems, and a deployed system may need both.

Retrieval-aware answer training Training a reader only on clean evidence can leave it unprepared for what retrieval actually returns. On knowledge-based visual question answering, a reader trained on retrieved captions that were reordered to put those containing the answer first does worse at test time, under ordinary retrieval, than a reader trained on captions in their retrieved order [77]. Adding distractor pages beside the evidence page during training gives a small gain in correctness under retrieval [131]. Training with hard-negative images and visual distortions goes further, weighting each answer token by how much its prediction changes between clean and distorted inputs, so learning concentrates on useful visual information (RagVL [7]). Training on the kinds of errors retrieval makes prepares a reader better than training on evidence cleaner than any it will see.

Progressive context training Training can also make the context harder through the number and arrangement of examples. An audio captioner can be trained on a gradually increasing number of interleaved retrieved audio–caption examples, and ablations credit gains to this curriculum together with a retrieval query that the model generates itself [132]. Increasing context complexity during training teaches the reader to use more examples rather than merely receive them at test time.

Preference alignment Preferences can improve selection for a fixed reader or change how the reader responds to evidence. On the selection side, reranker adapters can be trained on pairs of documents with which the frozen deployment reader answers correctly and incorrectly, and refreshing these pairs during training beats keeping the initial ones [81]. On the reader side, preferences target harmful reliance on evidence: training on cases where retrieval overturns an otherwise correct answer teaches the reader to keep the correct response [133]. One medical system addresses both sides [134]. It trains a separate retriever for each medical domain, and it trains the reader on two kinds of preferences: some contrast helpful with harmful retrieval, and others keep the retrieved context but replace the query image with an unrelated, heavily noised one, so the reader learns not to answer from retrieval alone. A selector learns what helps a stable reader; a reader learns when retrieved information should change its answer.

Trajectory optimization A tool call may prove useful only because of evidence it leads to several steps later. Reinforcement learning assigns credit across such interactions. One multimodal policy is rewarded for a correct answer and well-formed actions and penalized for each search (MMSearch-R1 [32]). Training it on a mix of questions that need search and questions that do not is what teaches it to answer directly when it can. The trained policy makes about a third fewer search calls than RAG baselines while remaining competitive with a larger reader. Another first imitates expert trajectories of searches and region inspections and is then rewarded for answer quality and for retrieving relevant images early [33]. During training, both policies see the observations their own actions produce, which supervised training on clean final answers never shows them.

Joint objectives Rewards can also cover ranking and answering together. One method first trains adapters to judge each document's relevance against hard negatives, then rewards correct document identifiers and correct answers jointly over benchmark candidate lists, leaving the retrieval encoder itself untrained [135]. Whatever the reward covers, the policy learns from the candidates and tool responses of its training environment, so its decisions transfer only as far as that environment resembles deployment.

7 Evaluation

Evaluation asks whether a system finds the evidence a question needs, whether its output is correct, whether the evidence supports that output, and, for media outputs, whether the requested attributes survive. RAG evaluation already reports retrieval quality, answer quality and faithfulness separately [2]; multimodal evaluation must also state which visual, audio or temporal information reached the model. We follow the order in which an evaluation is built: the benchmarks that define the task and annotate what success requires (§7.1), the metrics that score it (§7.2), and the controlled comparisons that show why scores change (§7.3).

7.1 Benchmark design

A benchmark fixes what the system receives, where it must search and what counts as success, and its annotations decide which failures can be located. Table 2 compares representative benchmarks along these lines. We group them into four families by setting (knowledge questions about images, documents, recordings and media outputs) and then turn to two design choices that cut across families: making the required evidence impossible to bypass, and keeping the question from revealing where to look.

Knowledge-intensive visual questions Benchmarks in this family pair an image and a question with an external collection, usually Wikipedia. Most ask for facts the image does not show (e.g., InfoSeek and M3-VQA [3, 136]); others provide reference images that help recognize the subject across changes of viewpoint and appearance (MRAG-Bench [137]). What each can check depends on what it annotates. Entity links, evidence sentences and answer spans connect recognition to the fact that justifies the answer, and evidence chains with section indices expose each step of questions about several entities. Ground-truth reference images identify a helpful reference but not a fact that justifies the answer. Only annotations that mark the justifying fact can tell a recognition failure from a retrieval failure.

Documents Document benchmarks differ mainly in how much of the search they leave to the system. Some require finding the relevant PDFs in a collection (M3DocVQA [10]) or the relevant content among distracting documents [36]. Others supply one scientific paper and require evidence chains across its pages, annotated with page and region anchors (BRIDGE [108]), or supply the paper's figures and ask the system to select and read them [38]. Chart collections add annotated paths across charts [60], and suites of real documents score parsing, retrieval and reasoning together [138]. A benchmark that supplies the document isolates understanding, one that supplies a collection also tests discovery, and only region-level anchors show whether a system found the right part of the right page.

Recordings Benchmarks over recordings differ in where the system searches and what its tools reveal. At the narrow end, a long-video benchmark fixes the input video and gives every system the same tools, which return clip captions; answers must cite clip identifiers, and the necessary clips and each hop of the reasoning are annotated (LongVidSearch [139]). Wider benchmarks search a collection: how-to questions with human answers are paired with videos from a large instructional collection [11], and spoken questions synthesized from knowledge-base triples require linking a spoken entity to a fact, in some settings after retrieving from a pool of recordings [35]. Open-web benchmarks widen the search further, from auditory clues to text, image and video sources, but do not specify which sources support each answer [76]. Annotated clip chains let an evaluation locate a failure at a particular hop; without support labels, only the final answer can be scored.

Table 2: Search scope, annotated support and output requirements of representative benchmarks.

Benchmark	Query / evidence / output	Search scope	Annotated support	Principal demand
InfoSeek-Human [3]	Image + text / Wikipedia / text	External Wikipedia collection	Entity links; evidence sentences; answer spans	Identify an entity and recover unseen facts
M3-VQA [136]	Image + text / Wikipedia text + images / text	External Wikipedia collection	Gold sentences, section indices and evidence chains	Multi-entity, parallel and sequential reasoning
MRAG-Bench [137]	Image + text / images / choice	External image collection	Ground-truth reference images	Use visual references across viewpoint and appearance changes
VisDoMBench [36]	Text / document text, tables, charts, slides / text	Supplied document set with distractors	Localized source evidence inherited from constituent datasets	Find relevant content among distracting documents
M3DocVQA [10]	Text / PDF text, images, tables / text	External PDF collection	Supporting source documents from MultimodalQA	Retrieve and combine evidence across documents
BRIDGE [108]	Text / paper text, tables, figures / text + evidence	Supplied scientific paper	Evidence chains with page and region anchors	Ground cross-page, cross-modal reasoning chains
ChartWalker-Bench [60]	Text / chart collection / text	External chart collection	Gold evidence sources; sampled paths and explanations	Integrate chart evidence at different granularities
LongVidSearch [139]	Text / video-clip captions / text + clip IDs	Within a supplied long video	Necessary evidence clips; hop-wise reasoning chains	Plan retrieval of discontinuous temporal evidence
Audiopedia [35]	Audio + text / KB facts; audio set in r-AQA / text	External KB; supplied audio set in r-AQA	Construction triples; relevant/irrelevant audio labels	Link spoken entities to facts and aggregate across audio
Omni-DeepSearch [76]	Audio + text / web text, images, video / text	Open-web search	Support-source labels not specified	Turn auditory clues into cross-modal, multi-hop search
FIG-Eval [5]	Text / web text + images / synthesized image	Open-web search	Expert text/image references; factual feature questions	Preserve perceptual, compositional and temporal facts
MRAMG-Bench [116]	Text / document text + images / interleaved text + images	External document collection	Reference answers with selected and placed source images	Select and place images within coherent answers
VinQA [30]	Text / document pages / interleaved text + visuals	Supplied retrieved page set	Visual boxes, citation tags and grounded reference answers	Cite and integrate visuals; recognize unanswerable queries

Media outputs Benchmarks for media outputs must annotate the artifact. For synthesized images, expert references and questions about factual features check whether perceptual, compositional and temporal facts appear in the image [5]. For interleaved answers, reference answers record which source images were selected and where they were placed [116], while boxes and citation tags further localize reuse and unanswerable questions test whether the system declines [30]. Audio synthesis is evaluated on distributional fidelity and target content in the settings of Section 5.3. In each case, answer accuracy alone cannot show whether the output preserved the source object or attribute it was meant to carry.

Evidence necessity In every family, a benchmark tests multimodal retrieval only if the required information cannot be bypassed. Dropping annotated evidence whose removal does not

change the answer ensures that every remaining clip matters [139]. Removing questions that models can answer without the context blocks answers from memory [140], although it makes difficulty relative to the models used for filtering. Facts that change over time make external search necessary, at the price of making the state of the collection part of the test [31], and noisy web environments without hints about modality make the system discover what to search for [141]. Each design removes a way to answer without finding the evidence, so a high score has to come from search.

Source cues Question wording can also reveal where to look before retrieval begins. A question that names the paper it concerns turns search into lookup, though the system must still find the table that holds the values. Gold labels add a related problem when they mark one supporting page and miss others. Removing explicit source references from questions addresses the cue, and verifying support across the whole corpus addresses the labels [142]. Without such checks, a high retrieval score may reflect the wording of the question more than the retriever.

7.2 Evaluation metrics

Metrics turn these success criteria into numbers. We define each family once: retrieval metrics score whether relevant material was found, answer metrics whether the output is correct, and support metrics whether the evidence justifies what the output says. Media outputs need further measures of attributes and quality.

Retrieval and information coverage Retrieval metrics compare returned items with the annotated relevant items. *Recall at k* is the fraction of relevant items that appear among the first k results, and a *hit at k* records whether at least one does. Rank-sensitive metrics also reward placing relevant items earlier; for example, *mean reciprocal rank* averages, over questions, one divided by the rank of the first relevant item. A hit can succeed while a multi-hop answer stays unsupported, as when the top pages name the building but none gives its date. *Evidence-set completeness* asks instead whether the selected items together contain all the information the answer needs. Coverage metrics count the distinct information units recovered, so a page that repeats a known fact earns nothing.

Coverage and reader utility Covering the right information still does not predict what the reader will say. Correlations between the information retrieved and the information expressed in the output are stronger for linear text pipelines than for iterative ones, and negative for one video pipeline [143]. Utility for a particular reader measures a third thing, the benefit to that reader's answer, for instance as the change in probability of informative answer tokens when a document is added (§4.2) [28]. Access to annotated information, completeness of the set and benefit to the reader should therefore be reported separately.

Answer correctness *Exact match* (EM) gives one when the prediction equals a reference answer after the benchmark's normalization and zero otherwise; *string accuracy* applies a benchmark's own, often looser, matching rule; and *judged accuracy* is the fraction of answers that an evaluator, human or model, accepts as correct. For answers that are sets, *intersection over union* (IoU) divides the number of elements shared by the prediction and the reference by the size of their union. These scores check the answer, not whether the retrieved sources justify it.

Support units and attribution Support metrics must say which unit they count. *Precision* P is the fraction of predicted items judged correct, *recall* R is the fraction of reference items recovered, and *F1* combines them, $F_1 = 2PR/(P + R)$. Answer F1 compares the tokens of the

predicted and reference answers; evidence-selection F1 compares the selected support units with the annotated ones; claim-level measures count verifiable claims. Long answers need claims as units, because one sentence can contain several facts drawn from several sources. One framework splits sentences into independently verifiable subclaims and reports two pairs of scores (MiRAGE [144]). *Information* precision and recall measure how much of the output is supported and how much of the needed content it covers; *citation* precision and recall measure whether the cited sources support each claim and whether each claim is cited. Checking subclaims against textual stand-ins for the media can miss facts that are only visible or audible, so the framework also checks them against the raw media.

Claim and localization metrics Other measures inspect smaller or differently defined units. The relevance of an image to the query can be scored separately from whether each factual statement in the response is supported [145], and extracted claims can be checked against individual text and image items [140]. *Required-phrase recall*, the fraction of specified key phrases a response contains, can be combined with a classification of whether the response answers or abstains to label hallucinations [146]. Localization measures score whether a system recovers annotated support boxes [111, 113], or how much of the time it inspected overlaps a relevant interval. Neither a correct location nor a correct answer establishes support. Measuring temporal grounding separately from support in tool outputs finds correct answers that the inspected sources do not support [20], and web-search trajectories can reach correct answers through wrong evidence or fabricated provenance [147]. Support has to be measured directly.

Judges and nontext outputs Automatic judges need validation for the property they score. Text-based frameworks score faithfulness, answer relevance and context relevance [148], but carrying them over to audiovisual material does not ensure that they agree with human judgments of grounding. Automatic verification against raw sources remains much weaker than human assessment of subclaims [144], and judges that agree strongly on answer correctness agree poorly on some labels for search trajectories, including the use of modality shortcuts [147]. A judge that is reliable for answers is not thereby reliable for evidence use.

Media-output metrics For synthesized images, factual-feature questions check whether perceptual, compositional and temporal attributes appear [5], and reports with figures need judgments of whether each visual supports the analysis [119]. For audio, *Fréchet Audio Distance* (FAD) compares the distribution of generated audio with that of reference audio in an acoustic feature space, with lower values meaning closer distributions, and *stimulus identification* tests whether a reconstruction can be matched to its target [123, 124]. These measure distributional fidelity and target content, not whether a source supports a claim. For reused images, selection and placement annotations show whether the right object accompanies the right explanation [30, 116]: accurate prose can sit beside an irrelevant image, and a correctly chosen image can be explained wrongly. Reuse is checked through that connection, synthesis through which requested attributes survived.

7.3 Experimental controls

Benchmark scores show how well a system does; controlled comparisons show why. Each comparison changes one condition, such as what evidence the model may access, whether the evidence is complete, how much noise it contains or which policy gathers it, and holds the rest fixed. Table 3 collects such comparisons, each scored with the study's own measure from Section 7.2, so scores are comparable within a row but not across rows.

Table 3: Within-study interventions separating evidence access, completeness, noise and control policy.

Model / setting	Baseline → intervention	Outcome	Factor changed	Coupled conditions
Qwen2.5-VL-7B; M^3-VQA [136]	Image + question → entity-name oracle	Answer-set IoU (%): 22.5 → 31.8	Entity-name access	Annotated evidence entities
Qwen2.5-VL-7B; M^3-VQA [136]	Image + question → agentic KB retrieval	Answer-set IoU (%): 22.5 → 31.2	Retrieved-context access	Planner and retrieval tools
Qwen3-VL-8B; MMLongBench-Doc, multipage [17]	Gold pages → remove lowest-ranked gold page	Judged accuracy (%): 38.0 → 18.5	Evidence completeness	Oracle set; ColQwen3 rank
Qwen3-VL-8B; MMLongBench-Doc, one gold page [17]	Gold page → add three distractor pages	Judged accuracy (%): 63.7 → 63.1	Distractor exposure	Gold retained; ranked noise
Gemini-2.5-Pro; DeflectionBench [18]	Gold contexts → gold + two distractors	Judged accuracy (%): 59.8 → 51.0	Context noise	Strict-RAG prompt; curated sample
Qwen2.5-72B; MultiModalQA [150]	Pre-retrieval → post-hoc modality router	String accuracy (0–1): 0.431 → 0.447	Router features	50% target escalation; cached sidecars
Qwen2.5-Omni-7B; document/audio/video [130]	Query-only → query + retrieval-state router	Δ routing score: +0.00288	Retrieval-state features	Fixed actions and cost weights
Sonnet 4.5; MMLongBench-Doc [151]	Static ColQwen2.5, $k = 20$ → six-tool agent	Official accuracy (0–1): 0.508 → 0.625	Evidence-access policy	Catalog and tools; 27k → 20k tokens

Answer-set IoU is intersection over union. Each row compares conditions within one study; scores across rows have different protocols. The routing result is a reported difference rather than an absolute score pair; its 95% confidence interval is $[-0.00726, 0.01107]$.

Access interventions The basic contrast is among three conditions: (i) answering without retrieval, which shows what the model already knows; (ii) answering with oracle evidence, which removes the search; and (iii) answering with retrieved evidence, which tests search and reading together. On a benchmark of visual questions about several entities, telling the model which entities appear raises answer-set IoU by about two fifths over the image and question alone. Agentic knowledge-base access reaches a similar level, and giving the model the complete supporting sentences scores considerably higher than either [136]. The similar scores come from different changes: the identity oracle removes recognition errors, while real retrieval adds a planner and tools. Even oracle evidence leaves integration errors on multi-hop questions [142], and better transcription or preprocessing can close an apparent retrieval gap without changing retrieval at all [149]. An end-to-end gain can be interpreted only when the comparison states whether it changed recognition, retrieval, integration or preprocessing.

Completeness and noise Missing support and added noise affect readers differently. On a multi-page document benchmark, removing the lowest-ranked required page roughly halves judged accuracy, while adding three ranked distractors to a single gold page barely changes it, though the two tests use different subsets of questions [17]. Curated distractors do more harm: keeping the gold contexts but adding two distractors noticeably reduces accuracy even when the model is told to answer only from the evidence [18]. Together with the results on context expansion in Section 4.2 [84], these comparisons show that completeness and tolerance of noise are separate properties of a reader, and a benchmark that tests one says little about the other.

Abstention Abstention changes what a system does when evidence is missing or conflicting, and its evaluation must count both sides. Separating correct answers, incorrect attempts and

refusals shows that stricter refusal instructions suppress answers even on some contexts that contain valid support [18], and false abstentions also appear on answerable document questions [17]. A refusal policy should be judged by the errors it avoids net of the answerable questions it discards.

Conflict and modality Whether a model prefers text or images in a conflict depends on how the conflict is built. Textual distractors override visual support in the configurations of one benchmark [18]. When the same content is rendered as text or as images and set against the model's elicited knowledge, models prefer images given a single piece of evidence but show unstable preferences among conflicting sources [21]. Reordering fixed evidence reveals sensitivity to position alone [152]. Modality preference is therefore a property of the reader, the format, the conflict and the order together, not of the modality.

Control policy and budget Adaptive policies should be compared under stated resource and access conditions. One router decides whether to consult the image only after drafting an answer, using features of the draft and of a verifier together with cached image descriptions. At the same budget for consulting images, it gains a little accuracy over a router that decides before retrieval [150]. In a separate study, adding retrieval-state features to a router that otherwise sees only the query gives no reliable gain [130]. In long-document question answering, a tool-using agent improves on a static retriever with the same reader while reading fewer input tokens. The agent also gains a structural catalog of the document and tools for pages, crops and text search, however, and with a smaller reader under matched token budgets the improvement is much smaller and uncertain [151]. Matching tokens controls cost, but the benefit of adaptation still depends on the reader and on what its tools make accessible.

8 Efficiency and robustness

A deployed system pays for work done once, when sources are indexed, and for work repeated with every query. It must also withstand errors and manipulation wherever queries, sources or tool outputs reach the model. We organize this section around cost along the pipeline (§8.1) and around the difference between ordinary retrieval failures and attacks on the evidence and on how it is used (§8.2).

8.1 Resource budgets

Costs follow the evidence path. Indexing, query processing, reading sources, tool calls, context tokens and data transmission can each change when a system changes its representation or policy, and a faster search does not by itself make the whole response cheaper.

Search cost and total response cost Retrieval itself can be made much faster. An inverted-file index cuts page-search latency by more than an order of magnitude over exact search, at a small cost in answer accuracy [10]. Another system searches pooled vectors first and reranks a short list with the full multi-vector representations. For larger encoders, this roughly quadruples retrieval throughput while preserving the short list's ranking, although pooling loses more recall with a smaller encoder [153]. Matching audio directly is several times faster than transcribing it and then retrieving text [8]. Each comparison stops at retrieval, but the reader still has to receive and process the selected pages or recordings. Search latency, throughput, storage and reader tokens measure different stages, so a report must name the cost it moved.

When visual processing occurs Moving visual work between indexing and query time decides whether it is paid once or with every query. A coarse video graph takes less than half as long to build as a fine one and keeps nearly all of the baseline's accuracy, at the price of slightly slower queries as the system recovers fine detail [23]. Deferring vision-model analysis until a page is selected avoids processing pages no query uses. On synthetic engineering-drawing questions full of identifiers, it more than doubles accuracy over descriptions prepared at indexing time, while adding tens of seconds to each query [154]. Deferred analysis repeats its work whenever a source serves many questions, so the right split depends on how often sources are reused as well as on latency.

Hardware and interaction On edge hardware, embedding and transmission can cost more than selection. A system that feeds video from edge devices to a cloud model can cluster frames at ingestion and transmit only selected ones, avoiding continuous embedding and communication. It lowers response latency by more than an order of magnitude relative to its selection baselines with comparable or better accuracy [155]. Keeping video memory on the device saves repeated computation and transmission as well as image input to the cloud model, so the budget that matters covers the whole path to the reader.

Interactive request costs Interactive search has two costs: the number of sequential rounds, which sets latency, and the total number of requests, which sets expense. A search agent that queries several grounded visual entities at once needs roughly five times fewer rounds of tool calls than an earlier multimodal deep-research agent, with higher accuracy [156]. Parallel calls shorten the critical path but still consume requests and reader input. Avoiding a call has a cost too: deciding whether to consult the image spends draft and verifier computation before any image is read [150]. End-to-end accounting must include the decisions that avoid calls as well as the calls themselves.

8.2 Robustness and security

Errors can enter before retrieval, in the source collection, or after evidence reaches the reader. Robustness studies need to distinguish evidence that is unreadable or irrelevant from evidence that someone manipulated, and each defense needs to say which part of the path it protects.

Retrieval noise and distribution shift Irrelevant evidence can change an answer with no adversary involved. A retrieval-augmented captioner does worse with random captions than with none, and sometimes copies words repeated across irrelevant captions even when a relevant one is present. Sampling training captions from a broader retrieved pool reduces this copying and improves captioning outside the training domain [157]. Such failures to use context well differ from the perception failures on degraded documents discussed in Section 3.1. Diagnosing robustness requires knowing whether the needed information was unreadable, never retrieved, or retrieved and then displaced by noise.

Query and corpus control The first security question is what the attacker can change. A medical image submitted as a query can be perturbed without touching the indexed collection: perturbations built on surrogate encoders transfer to black-box retrieval-and-generation systems, and the purification and compression defenses tested do not stop them [158]. Corpus poisoning changes the collection instead, with injected image–text pairs that target one query or attract retrieval broadly [159]. The two attacks cross different boundaries, so a defense that protects the collection says nothing about an untrusted query image.

Metadata and pixel control Poisoning can also target different parts of a source. Editing an image's metadata while leaving its pixels unchanged evades the paraphrasing and image–metadata similarity checks tested [160]. Localized edits to the pixels instead change the very feature a visual comparison asks about, and in black-box tests such images overturn answers the model previously gave correctly, even when generated descriptions mediate retrieval [161]. A defense that checks only the associated text misses attacks on both the metadata and the image.

Poisoned synthesis Synthesis can be steered in the same way. Poisoned captions in a music database can keep the words that make an entry relevant while adding acoustic instructions that contradict the query, so the entry is retrieved and the generated music follows the injected attributes [162]. Relevance and safe use are separate properties, and a defense must state whether it filters queries, protects the collection, inspects retrieved media or verifies the generated output.

Defenses and residual failures Defenses at the retriever act before poisoned content reaches the reader. One defense learns *soft prompts*, trainable vectors added to the input of a frozen retriever that embeds queries and documents separately. Trained adversarially, the prompts push poisoned documents away from queries while keeping clean embeddings in place. They lower poison retrieval and attack success across the attack families evaluated, without optimizing for each query; an answer counts as preserved when it agrees with the clean system's answer, whether or not that answer is correct [163]. Encoding the prompts nearly doubles runtime in one corruption test. Retriever-side training thus keeps many poisoned sources out of the context without per-query optimization, in exchange for a runtime cost paid on every query.

Stateful defenses Defenses can also act later in an agent's workflow. One defense inspects content and tool calls at several points in the agent's workflow, keeps a running record of what looked risky, and approves, alters or blocks each item [164]. On a synthetic suite it reduces attack success more than sixfold on average across threat types, and removing its memory or inspecting only the final generation weakens protection. Attacks that change which tool is called still succeed in most cases, and inspecting at each point more than triples response time. Retriever-side defenses keep some poisoned sources out of the context; stateful inspection catches failures later in the workflow, at a cost in latency and with persistent gaps in semantic judgment.

Privacy Access to source detail also allows disclosure. Black-box prompts can extract retrieved images or reveal their contents through transcription and detailed description [165]. Transforming what the reader receives can protect selected information: replacing faces after retrieval while keeping spatial attributes such as gaze and expression lets the reader reason about a scene without learning who is in it [166]. A privacy intervention must specify which information the task needs and which the reader should no longer receive.

9 Open problems

The surveyed results leave five connected problems. In each, a decision at one stage succeeds only if another stage preserves, exposes or verifies the right information.

Reader-dependent evidence-set utility Selection lacks a reliable estimate of what an item adds to the evidence a particular reader already has. Supervising documents one at a time with the reader's answers (§4.2) cannot tell whether a page completes a partly retrieved chain [28]. Experiments with oracle pages show that acquisition can succeed while integration across pages fails [17], and annotating scientific evidence chains for matching, relevance and logical

order shows that the systems best at answering are not the best at building chains [167]. What is missing is set-valued utility: a prediction of which missing item will let this reader use the support it already holds. In the running example, such a prediction would rate a second page about the building near zero and the history page highly.

Relations that survive reduction Compression must preserve the relation that makes an observation informative, such as the order of events, who spoke, or the alignment of sound and image. Keeping original chunks behind a coarse index preserves access to detail, but merging segments by visual similarity does not mark every important transition [23]: a visually stable interval can contain a decisive change in what is said or heard. Some errors happen before any reduction: recognizing entities from audio and forming cross-modal queries remain bottlenecks [76], and keeping an accurate interval cannot repair a wrong identification. Reduction methods need to represent uncertain perception and dependencies between separated intervals, not only local similarity.

Reliable stopping under cost constraints The next useful action depends on why an answer is still uncertain: missing support calls for search, unreadable detail for inspection, and conflicting sources for reconciliation. In a held-out routing comparison, features of the retrieval state alone give no reliable improvement [130]. An independent inspector makes the judgment of sufficiency explicit, yet longer exploratory traces can still weaken grounding [20]. No current method judges reliably whether another search, crop or verification call is worth its cost as tools, readers and budgets change, and confidence in the answer does not settle the question. Because refusing answerable questions discards usable support [17], stopping, continuing and abstaining need to be evaluated as one decision.

Verification through media transformations A report may reuse a source figure, redraw a chart or synthesize a new visual, and verification has to follow the relevant information through each change. Visual working memory aligned with sources, together with ratings of whether embedded images support the surrounding claims, begin to follow images through a report [168], but checks across a whole report still need claim-level links to the original text and media. Sound poses the same problem without explicit assertions: stimulus identification and distributional fidelity capture different effects of transforming a retrieved sample [124]. Verification must first determine which attributes should survive a transformation and then check their connection to the references, rather than relying on overall similarity.

Trust in changing evidence collections Agreement among sources can be manufactured. Generated text and images that pollute the evidence pool degrade the out-of-context misinformation detectors tested, which motivates cross-modal reranking and explicit reasoning between claims and evidence [169]. A system must therefore distinguish independent corroboration from repeated content with a shared or fabricated origin, and it must keep an auditable record of the source state it actually consulted as collections change. That record can itself disclose sensitive information, since retrieved images can leak directly or through transcription and description [165]. The record that makes an answer verifiable can also disclose what was consulted, so provenance and privacy have to be designed together.

10 Conclusion

Multimodal RAG works only when three steps succeed: finding evidence, keeping what makes it interpretable, and using it in the output. Original pages, frames and recordings preserve detail

but cost reader capacity and still depend on perception; captions, vectors and structured records make search cheaper but can drop labels, relations or sound events unless they link back to the source. Selection must then assemble a complete set, where one more item can complete a chain, repeat a known fact or introduce a conflict. The same choices recur in synthesis, where a retrieved example guides an output without serving as factual support.

Because these steps fail independently, they need separate measurements. Retrieval scores show whether a source or information unit was found, completeness measures whether the set holds what the question requires, support and attribution whether the output uses that material, and media-specific checks whether requested attributes survive reuse or synthesis. Adaptive search and retrieval-aware training improve several of these outcomes for particular readers, tools and budgets without removing the distinctions. Progress depends on methods that preserve relations through compression, recognize when evidence suffices, verify information after media transformations, and track trust and disclosure as collections change.

References

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, 2020.

[2] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv preprint arXiv:2405.07437, 2024.

[3] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? arXiv preprint arXiv:2302.11713, 2023.

[4] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-Imagen: Retrieval-Augmented Text-to-Image Generator. arXiv preprint arXiv:2209.14491, 2022.

[5] Yang Tian, Fan Liu, Jingyuan Zhang, Wei Bi, Yupeng Hu, and Liqiang Nie. Open Multimodal Retrieval-Augmented Factual Image Generation. arXiv preprint arXiv:2510.22521, 2025.

[6] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained Late-interaction Multi-modal Retrieval for Retrieval Augmented Visual Question Answering. arXiv preprint arXiv:2309.17133, 2023.

[7] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. MLLM Is a Strong Reranker: Advancing Multimodal Retrieval-augmented Generation via Knowledge-enhanced Reranking and Noise-injected Training. arXiv preprint arXiv:2407.21439, 2024.

[8] Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. WavRAG: Audio-Integrated Retrieval Augmented Generation for Spoken Dialogue Models. arXiv preprint arXiv:2502.14727, 2025.

[9] Shi Yu, Chaoyue Tang, Bokai Xu, et al. VisRAG: Vision-Based Retrieval-Augmented Generation on Multi-Modality Documents. International Conference on Learning Representations, 2025.

[10] Jaemin Cho, Debanjan Mahata, Ozan İrsoy, Yujie He, and Mohit Bansal. M3DocRAG: Multi-modal Retrieval is What You Need for Multi-page Multi-document Understanding. arXiv preprint arXiv:2411.04952, 2024.

[11] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-Augmented Generation over Video Corpus. arXiv preprint arXiv:2501.05874, 2025.

[12] Hexiang Hu, Kelvin C.K. Chan, Yu-Chuan Su, et al. Instruct-Imagen: Image Generation with Multi-modal Instruction. Computer Vision and Pattern Recognition, 2024.

[13] Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Retrieval-Augmented Text-to-Audio Generation. arXiv preprint arXiv:2309.08051, 2023.

[14] Ruochen Zhao, Hailin Chen, Weishi Wang, et al. Retrieving Multimodal Information for Augmented Generation: A Survey. Findings of the Association for Computational Linguistics: EMNLP 2023, 2023.

[15] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2502.08826, 2025.

[16] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A Survey of Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2504.08748, 2025.

[17] Lewei Xu, Yihao Ding, Zihan Xu, Daniel Yitian Su, Daochang Liu, Siwen Luo, Yifan Peng, and Wei Liu. Locating Failure in Multi-Page Visually Rich Document Understanding: An Empirical Attribution. arXiv preprint arXiv:2608.07943, 2026.

[18] Nicholas Moratelli, Christopher Davis, Leonardo F. R. Ribeiro, Bill Byrne, and Gonzalo Iglesias. Benchmarking Deflection and Hallucination in Large Vision-Language Models. arXiv preprint arXiv:2604.12033, 2026.

[19] Yifan Jia, Kailin Jiang, Yuyang Liang, et al. Benchmarking Multimodal Knowledge Conflict for Large Multimodal Models. arXiv preprint arXiv:2505.19509, 2025.

[20] Chenhao Qiu, Yechao Zhang, Xin Luo, Shien Song, and Xusheng Liu. VideoSEAL: Mitigating Evidence Misalignment in Agentic Long Video Understanding by Decoupling Answer Authority. arXiv preprint arXiv:2605.12571, 2026.

[21] Jungyeon Lee, Yejin Yoon, and Taeuk Kim. Same Semantics, Different Outcome: On the Modality Robustness of Multimodal LLMs under Knowledge Conflict. arXiv preprint arXiv:2609.00550, 2026.

[22] Monica Riedler and Stefan Langer. Beyond Text: Optimizing RAG with Multimodal Inputs for Industrial Applications. arXiv preprint arXiv:2410.21943, 2024.

[23] Zhe Jin, Zhimin Lin, Bin Zheng, Junhua Fang, and Huihua Yang. Coarse Indexing, Fine Evidence: Decoupling Temporal Granularity in Long-Video RAG. arXiv preprint arXiv:2608.23011, 2026.

[24] Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv preprint arXiv:2402.19473, 2024.

[25] Sensen Gao, Shanshan Zhao, Xu Jiang, Lunhao Duan, Yong Xien Chng, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, Jia-Wang Bian, and Mingming Gong. Scaling Beyond Context: A Survey of Multimodal Retrieval-Augmented Generation for Document Understanding. Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026.

[26] Zhuolin Yang, Wei Ping, Zihan Liu, et al. Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning. arXiv preprint arXiv:2302.04858, 2023.

[27] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient Document Retrieval with Vision Language Models. arXiv preprint arXiv:2407.01449, 2024.

[28] Xihang Wang, Zihan Wang, Chengkai Huang, Quan Z. Sheng, and Lina Yao. MEG-RAG: Quantifying Multi-modal Evidence Grounding for Evidence Selection in RAG. arXiv preprint arXiv:2604.24564, 2026.

[29] Ambuj Mehrish and Sebastiano Vascon. Constrained Dominant Sets for Multimodal Document Question Answering. arXiv preprint arXiv:2606.07252, 2026.

[30] Young Rok Jang, Hyesoo Kong, Kyunghwan An, Jae Sub Huh, Gyeonghun Kim, and Stanley Jungkyu Choi. VinQA: Visual Elements Interleaved Long-form Answer Generation for Real-World Multimodal Document QA. arXiv preprint arXiv:2606.16092, 2026.

[31] Yangning Li, Yinghui Li, Xinyu Wang, et al. Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset and Self-adaptive Planning Agent. arXiv preprint arXiv:2411.02937, 2024.

[32] Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. MMSearch-R1: Incentivizing LMMs to Search. arXiv preprint arXiv:2506.20670, 2025.

[33] Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. VRAG-RL: Empower Vision-Perception-Based RAG for Visually Rich Information Understanding via Iterative Reasoning with Reinforcement Learning. arXiv preprint arXiv:2505.22019, 2025.

[34] Seohyun Lee, Seoung Choi, Dohwan Ko, Jongha Kim, and Hyunwoo J. Kim. VideoSearch-R1:

Iterative Video Retrieval and Reasoning via Soft Query Refinement. arXiv preprint arXiv:2607.00446, 2026.

[35] Abhirama Subramanyam Penamakuri, Kiran Chhatre, and Akshat Jain. Audiopedia: Audio QA with Knowledge. arXiv preprint arXiv:2412.20619, 2024.

[36] Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Gowswami, Ryan A. Rossi, and Dinesh Manocha. VisDoM: Multi-Document QA with Visually Rich Elements Using Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2412.10704, 2024.

[37] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. VDocRAG: Retrieval-Augmented Generation over Visually-Rich Documents. arXiv preprint arXiv:2504.09795, 2025.

[38] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers. arXiv preprint arXiv:2407.09413, 2024.

[39] Shuying Chen, Sen Cui, and Zhong Cao. Guideline-grounded retrieval-augmented generation for ophthalmic clinical decision support. arXiv preprint arXiv:2603.21925, 2026.

[40] Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. Semi-Parametric Video-Grounded Text Generation. arXiv preprint arXiv:2301.11507, 2023.

[41] Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen. DRCap: Decoding CLAP Latents with Retrieval-Augmented Generation for Zero-shot Audio Captioning. arXiv preprint arXiv:2410.09472, 2024.

[42] Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities. arXiv preprint arXiv:2504.20734, 2025.

[43] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos. arXiv preprint arXiv:2502.01549, 2025.

[44] Ruisheng Cao, Hanchong Zhang, Tiancheng Huang, et al. NeuSym-RAG: Hybrid Neural Symbolic Retrieval with Multiview Structuring for PDF Question Answering. arXiv preprint arXiv:2505.19754, 2025.

[45] Yifan Zhu, Yu Mi, Yue Lu, Yanchu Guan, and Zhixuan Chu. LFRAG: Layout-oriented Fine-grained Retrieval-Augmented Generation on Multimodal Document Understanding. arXiv preprint arXiv:2605.22829, 2026.

[46] Masao Someki, Chien yu Huang, Siddhant Arora, Samuele Cornell, Markus Müller, Nathan Susanj, Rupak V Swaminathan, Grant P Strimel, Jing Liu, and Shinji Watanabe. PlanRAG-Audio: Planning and Retrieval Augmented Generation for Long-form Audio Understanding. Findings of ACL 2026 (accepted; per arXiv record), 2026.

[47] Zackary Rackauckas and Julia Hirschberg. VoxRAG: A Step Toward Transcription-Free RAG Systems in Spoken Question Answering. arXiv preprint arXiv:2505.17326, 2025.

[48] Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking Multimodal Retrieval for Long Documents. arXiv preprint arXiv:2501.08828, 2025.

[49] Emre Kuru and Mehmet Onur Keskin. Evaluating Modern RAG: Textual, Multimodal, Dense, and Late Interaction Pipelines. arXiv preprint arXiv:2608.23176, 2026.

[50] Alexander Most, Joseph Winjum, Ayan Biswas, Shawn Jones, Nishath Rajiv Ranasinghe, Dan O'Malley, and Manish Bhattarai. Lost in OCR Translation? Vision-Based Approaches to Robust Document Retrieval. arXiv preprint arXiv:2505.05666, 2025.

[51] Elias Lumer, Alex Cardenas, Matt Melich, Myles Mason, Sara Dieter, Vamse Kumar Subbiah, Pradeep Honaganahalli Basavaraju, and Roberto Hernandez. Comparison of Text-Based and Image-Based Retrieval in Modern Multimodal Retrieval Augmented Generation Large Language Model Systems. arXiv preprint arXiv:2511.16654, 2025.

[52] Rashid Mushkani, Hugo Bérard, and Shin Koseki. Evidence-Grounded Constraint Checking in Construction Documents. arXiv preprint arXiv:2607.29058, 2026.

[53] Joongmin Shin, Jeongbae Park, Jaehyung Seo, and Heuiseok Lim. M3DocDep: Multi-modal, Multi-page, Multi-document Dependency Chunking with Large Vision-Language Models. arXiv

preprint arXiv:2605.18774, 2026.

[54] Yilong Zuo, Xunkai Li, Jing Yuan, Qiangqiang Dai, Hongchao Qin, and Ronghua Li. MAGE-RAG: Multigranular Adaptive Graph Evidence for Agentic Multimodal RAG in Long-Document QA. arXiv preprint arXiv:2606.15906, 2026.

[55] Yongyue Zhang and Yaxiong Wu. MLDocRAG: Multimodal Long-Context Document Retrieval Augmented Generation. arXiv preprint arXiv:2602.10271, 2026.

[56] Ainulla Khan, Moyuru Yamada, and Srinidhi Akella. BRIT: Bidirectional Retrieval over Unified Image-Text Graph. arXiv preprint arXiv:2505.18450, 2025.

[57] Xin He, Yili Wang, Wenqi Fan, Qing Li, Qinggang Zhang, Yi Chang, and Xin Wang. HVM-GraphRAG: Holistic-View Multimodal Graph Retrieval-Augmented Generation on Complex Document. arXiv preprint arXiv:2607.24861, 2026.

[58] Jacob Si, Mike Qu, Michelle Lee, Marek Rei, and Yingzhen Li. TabRAG: Improving Tabular Document Question Answering for Retrieval Augmented Generation via Structured Representations. arXiv preprint arXiv:2511.06582, 2025.

[59] Alex Dantart and Marco Kóvacs-Navarro. Topo-RAG: Topology-aware retrieval for hybrid text-table documents. arXiv preprint arXiv:2601.10215, 2026.

[60] Ning Tang, Chenghan Xie, Hanyang Yuan, Yi Li, Renhong Huang, Qian Kou, Xiaofeng Shi, Hua Zhou, and Jiarong Xu. ChartWalker: Benchmarking the Cross-Chart RAG Task with Hierarchical Knowledge Graphs. arXiv preprint arXiv:2606.23997, 2026.

[61] Kevin Dela Rosa. Video Enriched Retrieval Augmented Generation Using Aligned Video Captions. arXiv preprint arXiv:2405.17706, 2024.

[62] Yannis Tevissen, Khalil Guetari, and Frédéric Petitpont. Towards Retrieval Augmented Generation over Large Video Libraries. arXiv preprint arXiv:2406.14938, 2024.

[63] Yongdong Luo, Xiawu Zheng, Guilin Li, et al. Video-RAG: Visually-aligned Retrieval-Augmented Long Video Comprehension. arXiv preprint arXiv:2411.13093, 2024.

[64] Honghao Fu, Miao Xu, Yiwei Wang, Dailing Zhang, Jun Liu, and Yujun Cai. VideoStir: Understanding Long Videos via Spatio-Temporally Structured and Intent-Aware RAG. arXiv preprint arXiv:2604.05418, 2026.

[65] Yifan Zhu, Xinyu Mu, Tao Feng, Zhonghong Ou, Yuning Gong, and Haoran Luo. OmniRAG-Agent: Agentic Omnimodal Reasoning for Low-Resource Long Audio-Video Question Answering. arXiv preprint arXiv:2602.03707, 2026.

[66] Le Zhang, Hao Chen, Vlad Roznyatovskiy, Jianzhong Zhang, and Ke Sun. EgoCITE: Context-Augmented Indexing and Time-Aware Retrieval for Long-Horizon Egocentric Memory. arXiv preprint arXiv:2608.12627, 2026.

[67] Kartik Hegde, Arvind Krishna Sridhar, Naveen Vakada, Yinyi Guo, and Erik Visser. Event-Grounded Question Answering over Long Audio via Structured Retrieval. arXiv preprint arXiv:2602.14612, 2026.

[68] Chenghao Zhang, Guanting Dong, Xinyu Yang, and Zhicheng Dou. Towards Mixed-Modal Retrieval for Universal Retrieval-Augmented Generation. arXiv preprint arXiv:2510.17354, 2025.

[69] Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, Suofei Feng, Ryan Rossi, and Zhengzhong Tu. mRAG: Elucidating the Design Space of Multi-modal Retrieval-Augmented Generation. arXiv preprint arXiv:2505.24073, 2025.

[70] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers. Annual Meeting of the Association for Computational Linguistics, 2024.

[71] Michael Günther, Saba Sturua, Mohammad Kalim Akram, et al. jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval. arXiv preprint arXiv:2506.18902, 2025.

[72] Arun Reddy, Alexander Martin, Eugene Yang, Andrew Yates, Kate Sanders, Kenton Murray, Reno Kriz, Celso M. de Melo, Benjamin Van Durme, and Rama Chellappa. Video-ColBERT: Contextualized Late Interaction for Text-to-Video Retrieval. arXiv preprint arXiv:2503.19009, 2025.

[73] Meftun Akarsu, Recep Kaan Karaman, and Christopher Mierbach. From BM25 to Corrective RAG: Benchmarking Retrieval Strategies for Text-and-Table Documents. arXiv preprint arXiv:2604.01733, 2026.

[74] Joongmin Shin, Gyuho Shim, Jeongbae Park, Jaehyung Seo, and Heuiseok Lim. HiKEY: Hierarchical Multimodal Retrieval for Open-Domain Document Question Answering. arXiv preprint arXiv:2605.29606, 2026.

[75] Jiacheng Tao, Qingyun Sun, Haonan Yuan, Ziwei Zhang, and Jianxin Li. DualG-MRAG: Decoupling Macro-Reasoning and Micro-Matching for Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2607.28580, 2026.

[76] Tao Yu, Yiming Ding, Shenghua Chai, et al. Omni-DeepSearch: A Benchmark for Audio-Driven Omni-Modal Deep Search. arXiv preprint arXiv:2605.08762, 2026.

[77] Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. Multimodal Reranking for Knowledge-Intensive Visual Question Answering. arXiv preprint arXiv:2407.12277, 2024.

[78] Guanchen Wu, Jiayuan Ding, Subhabrata Mukherjee, and Carl Yang. What the Reranker Sees: Multi-Aspect Page Annotation for Long-Document Multimodal Question Answering. arXiv preprint arXiv:2608.14841, 2026.

[79] Inho Kim and Sumyeong Ahn. A Reranker for Orchestrating Heterogeneous Speech and Text Retrievers. arXiv preprint arXiv:2608.26194, 2026.

[80] Yuho Lee, Jisu Shin, Nicole Hee-Yeon Kim, Jihwan Bang, Juntae Lee, Kyuwoong Hwang, Fatih Porikli, and Hwanjun Song. Rethinking RAG in Long Videos: What to Retrieve and How to Use It? arXiv preprint arXiv:2606.13141, 2026.

[81] Zhan-Lun Chang, Dong-Jun Han, Seyyedali Hosseinalipour, Mung Chiang, and Christopher G. Brinton. Bridging the Semantic-Utility Gap in Multimodal RAG via Generator-in-the-Loop Alignment. arXiv preprint arXiv:2609.08188, 2026.

[82] Mengdan Zhu, Senhao Cheng, Guangji Bai, Yifei Zhang, and Liang Zhao. Cross-modal RAG: Sub-dimensional Text-to-Image Retrieval-Augmented Generation. arXiv preprint arXiv:2505.21956, 2025.

[83] Ambuj Mehrish and Sebastiano Vascon. FLOWREADER: Min-Cost Flow Optimization for Multi-Modal Long Document Q&A. arXiv preprint arXiv:2606.07235, 2026.

[84] Jiankun Wang, Yisen Gao, Ziwei Zhang, Xingcheng Fu, Jiaxin Bai, and Chen Gao. Does More Retrieved Evidence Help Visual Retrieval-Augmented Generation with Diffusion Language Models? arXiv preprint arXiv:2608.07006, 2026.

[85] Jaeyoon Jung, Yejun Yoon, and Kunwoo Park. Is a Picture Worth a Thousand Words? Adaptive Multimodal Fact-Checking with Visual Evidence Necessity. arXiv preprint arXiv:2604.04692, 2026.

[86] Chan-Wei Hu and Zhengzhong Tu. Region-R1: Reinforcing Query-Side Region Cropping for Multi-Modal Re-Ranking. arXiv preprint arXiv:2604.05268, 2026.

[87] Jeonghwan Kim, Renjie Tao, Sanat Sharma, et al. Pixel-Grounded Retrieval for Knowledgeable Large Multimodal Models. arXiv preprint arXiv:2601.19060, 2026.

[88] Zhengren Wang, Dongsheng Ma, Huaping Zhong, Jiayu Li, Wentao Zhang, Bin Wang, and Conghui He. AgenticOCR: Parsing Only What You Need for Efficient Retrieval-Augmented Generation. arXiv preprint arXiv:2602.24134, 2026.

[89] Jon Saad-Falcon, Joe Barrow, Alexa Siu, Ani Nenkova, David Seunghyun Yoon, Ryan A. Rossi, and Franck Dernoncourt. PDFTriage: Question Answering over Long, Structured Documents. arXiv preprint arXiv:2309.08872, 2023.

[90] Xihang Wang, Zihan Wang, Chengkai Huang, Cao Liu, Ke Zeng, Quan Z. Sheng, and Lina Yao. Purifying Multimodal Retrieval: Fragment-Level Evidence Selection for RAG. arXiv preprint arXiv:2604.27600, 2026.

[91] Oanh N. Tran, Thanh Quoc Hung Le, Oscar Chew, Kuan-Hao Huang, and Khoa D. Doan. Look Before You Zoom: Adaptive Routing for the Resolution-Context Trade-off in Visual RAG. arXiv preprint arXiv:2606.21968, 2026.

[92] Adrien Mialland, Marc Plantevit, Julien Gallois, and Céline Robardet. ViSAR: Training-Free Adaptive-k Retrieval for Visual Document Question Answering. arXiv preprint arXiv:2609.02486, 2026.

[93] Yucheng Shen, Lingyong Yan, Jiulong Wu, Shuaiqiang Wang, Jianmin WU, Dawei Yin, and Min Cao. Navigating Sparse Evidence: Agentic Visual RAG via Explicit Context Selection and Consolidation.

arXiv preprint arXiv:2609.15800, 2026.

[94] Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to Filter Context for Retrieval-Augmented Generation. arXiv preprint arXiv:2311.08377, 2023.

[95] Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. arXiv preprint arXiv:2310.04408, 2023.

[96] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. arXiv preprint arXiv:2310.06839, 2023.

[97] Jiayang Yu, Jialun Zhong, and Lei Zou. RAGOCR: Optical Compression of Retrieval-Augmented Text via Visual Representation. arXiv preprint arXiv:2608.00765, 2026.

[98] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. arXiv preprint arXiv:2404.15406, 2024.

[99] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. MMHQA-ICL: Multimodal In-context Learning for Hybrid Question Answering over Text, Tables and Images. arXiv preprint arXiv:2309.04790, 2023.

[100] Minda Zhao, Fangyu Hu, Yan Luo, et al. Representation-guided in-context learning for medical image interpretation with multimodal large language models. arXiv preprint arXiv:2609.24057, 2026.

[101] Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong. Fact-Aware Multimodal Retrieval Augmentation for Accurate Medical Radiology Report Generation. arXiv preprint arXiv:2407.15268, 2024.

[102] Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation. arXiv preprint arXiv:2502.15040, 2025.

[103] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. RAMM: Retrieval-augmented Biomedical Visual Question Answering with Multi-modal Pre-training. arXiv preprint arXiv:2303.00534, 2023.

[104] Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, and Liqiang Nie. RA-BLIP: Multimodal Adaptive Retrieval-Augmented Bootstrapping Language-Image Pre-training. arXiv preprint arXiv:2410.14154, 2024.

[105] Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. ViPCap: Retrieval Text-Based Visual Prompts for Lightweight Image Captioning. arXiv preprint arXiv:2412.19289, 2024.

[106] Zhi Zheng, Ziqiao Meng, Hao Luan, Wei Liu, and Wee Sun Lee. One Token per Multimodal Evidence: Latent Memory for Resource-Constrained QA. arXiv preprint arXiv:2606.10572, 2026.

[107] Yang Tian, Fan Liu, Jingyuan Zhang, Victoria W., Yupeng Hu, and Liqiang Nie. CoRe-MMRAG: Cross-Source Knowledge Reconciliation for Multimodal RAG. arXiv preprint arXiv:2506.02544, 2025.

[108] Biao Xiang, Soyeon Caren Han, and Yihao Ding. BRIDGE: Benchmark for multi-hop Reasoning In long multimodal Documents with Grounded Evidence. arXiv preprint arXiv:2603.07931, 2026.

[109] Junyu Xiong, Yonghui Wang, Rongjian Gu, Chenyu Liu, Bing Yin, Wengang Zhou, and Houqiang Li. HIEVI-RAG: Hierarchical Evidence-Driven Reasoning for Long Document Understanding. arXiv preprint arXiv:2607.04625, 2026.

[110] Delvin Ce Zhang, Suhan Cui, Zhelin Chu, Xianren Zhang, and Dongwon Lee. MEVER: Multi-Modal and Explainable Claim Verification with Graph-based Evidence Retrieval. arXiv preprint arXiv:2602.10023, 2026.

[111] Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. VISA: Retrieval Augmented Generation with Visual Source Attribution. arXiv preprint arXiv:2412.14457, 2024.

[112] Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu, and Enhong Chen. Look as You Think: Unifying Reasoning and Visual Evidence Attribution for Verifiable Document RAG via Reinforcement Learning. arXiv preprint arXiv:2511.12003, 2025.

[113] Zhuchenyang Liu, Yao Zhang, and Yu Xiao. Evidence Attribution in Visual Document Understanding

without Coordinates or Region Labels. arXiv preprint arXiv:2607.24651, 2026.

[114] Baiyu Chen, Wilson Wongso, Xiaoqian Hu, Yue Tan, and Flora Salim. Multi-Stage Verification-Centric Framework for Mitigating Hallucination in Multi-Modal RAG. arXiv preprint arXiv:2507.20136, 2025.

[115] Vishwas Sathish, Viresh Ranjan, Xinliang Zhu, Arnab Dhua, and Douglas Gray. Eliciting Self-Verification in Multimodal Reasoning Agents with Reinforcement Learning. arXiv preprint arXiv:2609.08025, 2026.

[116] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. MRAMG-Bench: A Comprehensive Benchmark for Advancing Multimodal Retrieval-Augmented Multimodal Generation. arXiv preprint arXiv:2502.04176, 2025.

[117] Yin Li, Ziyang Hu, Zhiyu Guo, Xiangyu Liu, Wenbin Li, Boo-Ho Yang, Rav Lawana, Ziyue Li, Wei Zeng, and Fugee Tsung. HAM-RAG: Hierarchy-Aware Multimodal RAG for Structure-Faithful Interleaved Generation. arXiv preprint arXiv:2608.14032, 2026.

[118] Fangda Ye, Zhifei Xie, Yuxin Hu, Yihang Yin, Shurui Huang, Shikai Dong, Jianzhu Bao, and Shuicheng Yan. Deep-Reporter: Deep Research for Grounded Multimodal Long-Form Generation. arXiv preprint arXiv:2604.10741, 2026.

[119] Xinkai Ma, Zhiqi Bai, Dingling Zhang, et al. TVIR: Building Deep Research Agents Towards Text-Visual Interleaved Report Generation. arXiv preprint arXiv:2606.02320, 2026.

[120] Zhuofan Shi, Peilun Jia, Baoqin Sun, Haiyang Shen, Sixiong Xie, Yun Ma, and Xiang Jing. ViDR: Grounding Multimodal Deep Research Reports in Source Visual Evidence. arXiv preprint arXiv:2605.13034, 2026.

[121] Youngsun Lim and Hyunjung Shim. Addressing Image Hallucination in Text-to-Image Generation through Factual Image Retrieval. arXiv preprint arXiv:2407.10683, 2024.

[122] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. FairRAG: Fair Human Generation via Fair Retrieval Augmentation. arXiv preprint arXiv:2403.19964, 2024.

[123] Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation. arXiv preprint arXiv:2411.05141, 2024.

[124] Ambuj Mehrish and Sebastiano Vascon. RAG-Audio: Retrieval-Augmented Generation for Faithful Brain-to-Audio Reconstruction. arXiv preprint arXiv:2608.09331, 2026.

[125] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding. arXiv preprint arXiv:2503.13964, 2025.

[126] Yuechen Wang, Yuming Qiao, Dan Meng, Jun Yang, Haonan Lu, Zhenyu Yang, and Xudong Zhang. Efficient Agent: Optimizing Planning Capability for Multimodal Retrieval Augmented Generation. arXiv preprint arXiv:2508.08816, 2025.

[127] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, et al. MMSearch: Unveiling the Potential of Large Models as Multi-modal Search Engines. arXiv preprint arXiv:2409.12959, 2024.

[128] Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. ViDoRAG: Visual Document Retrieval-Augmented Generation via Dynamic Iterative Reasoning Agents. Conference on Empirical Methods in Natural Language Processing, 2025.

[129] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep Video Discovery: Agentic Search with Tool Use for Long-form Video Understanding. arXiv preprint arXiv:2505.18079, 2025.

[130] Qiaomu Li, Qiuyuan Zhang, and Nong Ming. Beyond the Query: Do Retrieval Signals Improve Adaptive Multimodal RAG Routing? arXiv preprint arXiv:2609.12437, 2026.

[131] Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. M-LongDoc: A Benchmark for Multimodal Super-Long Document Understanding and a Retrieval-Aware Tuning Framework. arXiv preprint arXiv:2411.06176, 2024.

[132] Changin Choi, Sungjun Lim, and Wonjong Rhee. Enhancing Retrieval-Augmented Audio Captioning with Generation-Assisted Multimodal Querying and Progressive Learning. arXiv preprint arXiv:2410.10913, 2024.

[133] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao.

RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models. arXiv preprint arXiv:2407.05131, 2024.

[134] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. arXiv preprint arXiv:2410.13085, 2024.

[135] Shengwei Zhao, Jingwen Yao, Sitong Wei, Linhai Xu, Yuying Liu, Dong Zhang, Zhiqiang Tian, and Shaoyi Du. MMRAG-RFT: Two-stage Reinforcement Fine-tuning for Explainable Multi-modal Retrieval-augmented Generation. arXiv preprint arXiv:2512.17194, 2025.

[136] Jiatong Ma, Longteng Guo, Yuchen Liu, Zijia Zhao, Dongze Hao, Xuanxu Lin, and Jing Liu. M3-VQA: A Benchmark for Multimodal, Multi-Entity, Multi-Hop Visual Question Answering. arXiv preprint arXiv:2604.25122, 2026.

[137] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. MRAG-Bench: Vision-Centric Evaluation for Retrieval-Augmented Multimodal Models. arXiv preprint arXiv:2410.08182, 2024.

[138] Yulong Hui, Yao Lu, and Huanchen Zhang. UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis. arXiv preprint arXiv:2406.15187, 2024.

[139] Rongyi Yu, Chenyuan Duan, and Wentao Zhang. LongVidSearch: An Agentic Benchmark for Multi-hop Evidence Retrieval Planning in Long Videos. arXiv preprint arXiv:2603.14468, 2026.

[140] Yuelyu Ji, Wuwei Lan, and Patrick NG. MRAG-Suite: A Cross-Domain Diagnostic Benchmark and Tooling for Visual Retrieval-Augmented Generation. arXiv preprint arXiv:2509.24253, 2025.

[141] Han Wang, David Wan, Hyunji Lee, Thinh Pham, Mikaela Cankosyan, Weiyuan Chen, Elias Stengel-Eskin, Tu Vu, and Mohit Bansal. MERRIN: A Benchmark for Multimodal Evidence Retrieval and Reasoning in Noisy Web Environments. arXiv preprint arXiv:2604.13418, 2026.

[142] Wenxuan Shen, Mingjia Wang, Yaochen Wang, Dongping Chen, Junjie Yang, Yao Wan, and Weiwei Lin. Are We on the Right Way for Assessing Document Retrieval-Augmented Generation? arXiv preprint arXiv:2508.03644, 2025.

[143] Saron Samuel, Alexander Martin, Eugene Yang, Andrew Yates, Dawn Lawrie, Ian Soboroff, Laura Dietz, and Benjamin Van Durme. Beyond Relevance: On the Relationship Between Retrieval and RAG Information Coverage. arXiv preprint arXiv:2603.08819, 2026.

[144] Alexander Martin, William Walden, Reno Kriz, Dengjia Zhang, Kate Sanders, Eugene Yang, Chihsheng Jin, and Benjamin Van Durme. Seeing Through the MIRAGE: Evaluating Multimodal Retrieval Augmented Generation. arXiv preprint arXiv:2510.24870, 2025.

[145] Matin Mortaheb, Mohammad A. (Amir) Khojastepour, Srimat T. Chakradhar, and Sennur Ulukus. RAG-Check: Evaluating Multimodal Retrieval Augmented Generation Performance. arXiv preprint arXiv:2501.03995, 2025.

[146] Samuel Hildebrand, Curtis Taylor, Sean Oesch, James M. Ghawaly, Jr., Amir Sadovnik, Ryan Shivers, Brandon Schreiber, and Kevin Kurian. FATHOMS-RAG: A Framework for the Assessment of Thinking and Observation in Multimodal Systems that use Retrieval Augmented Generation. arXiv preprint arXiv:2510.08945, 2025.

[147] Zhengxian Wu, Junjie Gao, and Kai Yang. Silent Failures in Multimodal Agentic Search: A Diagnostic Taxonomy and Cross-Judge Evaluation. arXiv preprint arXiv:2607.19793, 2026.

[148] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv preprint arXiv:2309.15217, 2023.

[149] Martin Asenov, Kenza Benkirane, Dan Goldwater, and Aneiss Ghodsi. Retrieval or Representation? Reassessing Benchmark Gaps in Multilingual and Visually Rich RAG. arXiv preprint arXiv:2603.04238, 2026.

[150] Xue Li and Yiming Gai. Modality Relevance is not Modality Utility: Post-hoc Selective Modality Escalation for Cost-Aware Multimodal RAG. arXiv preprint arXiv:2607.05438, 2026.

[151] Kenan E. Ak, Jay Mohta, Gwang Gook Lee, Yan Xu, and Dimitrios Dimitriadis. An Empirical Study of VLM Pipelines for Long-Document QA. arXiv preprint arXiv:2609.29933, 2026.

[152] Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. Who is in the Spotlight: The Hidden Bias Undermining Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2506.11063, 2025.

[153] Ara Yeroyan. Visual RAG Toolkit: Scaling Multi-Vector Visual Retrieval with Training-Free Pooling and Multi-Stage Search. arXiv preprint arXiv:2602.12510, 2026.

[154] Tao Xu. Index Light, Reason Deep: Deferred Visual Ingestion for Visual-Dense Document Question Answering. arXiv preprint arXiv:2602.14162, 2026.

[155] Shengyuan Ye, Bei Ouyang, Tianyi Qian, Liekang Zeng, Mu Yuan, Xiaowen Chu, Weijie Hong, and Xu Chen. Venus: An Efficient Edge Memory-and-Retrieval System for VLM-based Online Video Understanding. arXiv preprint arXiv:2512.07344, 2025.

[156] Guankai Li, Jiabin Chen, Yi Xu, Xichen Zhang, and Yuan Lu. HyperEyes: Dual-Grained Efficiency-Aware Reinforcement Learning for Parallel Multimodal Search Agents. arXiv preprint arXiv:2605.07177, 2026.

[157] Wenyan Li, Jiaang Li, Rita Ramos, Raphael Tang, and Desmond Elliott. Understanding Retrieval Robustness for Retrieval-Augmented Image Captioning. arXiv preprint arXiv:2406.02265, 2024.

[158] Yingjia Shang, Yi Liu, Huimin Wang, Furong Li, Wenfang Sun, Chengyu Wu, and Yefeng Zheng. Medusa: Cross-Modal Transferable Adversarial Attacks on Multimodal Medical Retrieval-Augmented Generation. arXiv preprint arXiv:2511.19257, 2025.

[159] Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. MM-POISONRAG: Disrupting Multimodal RAG with Local and Global Knowledge Poisoning Attacks. arXiv preprint arXiv:2502.17832, 2025.

[160] Kennedy Edemacu and Mohammad Mahdi Shokri. Hidden in the Metadata: Stealth Poisoning Attacks on Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2603.00172, 2026.

[161] Rujin Liang, Zhongpu Chen, Yuhao Lei, and Xin Miao. Vis-Poison: Poisoning Visual Knowledge in Multimodal Retrieval-Augmented Generation. arXiv preprint arXiv:2608.20756, 2026.

[162] Yizhu Wen, Shuhao Zhang, Nan Zhang, Long Cheng, and Hanqing Guo. Mental Damage: Caption Poisoning Attacks on Retrieval-Augmented Text-to-Music Generation. arXiv preprint arXiv:2605.30365, 2026.

[163] Chang Liu, Yuni Lai, Mingyue Cui, Cong Tian, Yunyan Zhang, Xian Wu, Kai Zhou, and Bin Xiao. DSPrompt: Dynamic Soft Prompt Defense Against M-RAG Corruption. arXiv preprint arXiv:2608.16536, 2026.

[164] Inderjeet Singh, Vikas Pahuja, Aishvariya Priya Rathina Sabapathy, et al. Adversarial Intent is a Latent Variable: Stateful Trust Inference for Securing Multimodal Agentic RAG. arXiv preprint arXiv:2602.21447, 2026.

[165] Jiankun Zhang, Shenglai Zeng, Jie Ren, Tianqi Zheng, Hui Liu, Xianfeng Tang, Hui Liu, and Yi Chang. Beyond Text: Unveiling Privacy Vulnerabilities in Multi-modal Retrieval-Augmented Generation. arXiv preprint arXiv:2505.13957, 2025.

[166] Zehua Cheng, Wei Dai, and Jiahao Sun. Identity-Decoupled Anonymization for Visual Evidence in Multi-modal Retrieval-Augmented Generation. arXiv preprint arXiv:2604.23584, 2026.

[167] Yiming Ren, Junjie Wang, Yuxin Meng, et al. SIN-Bench: Tracing Native Evidence Chains in Long-Context Multimodal Scientific Interleaved Literature. arXiv preprint arXiv:2601.10108, 2026.

[168] Chenghao Zhang, Guanting Dong, Yufan Liu, Tong Zhao, Xiaoxi Li, and Zhicheng Dou. Towards Verifiable Multimodal Deep Research: A Multi-Agent Harness for Interleaved Report Generation. arXiv preprint arXiv:2605.29861, 2026.

[169] Zehong Yan, Peng Qi, Wynne Hsu, and Mong Li Lee. Mitigating GenAI-Powered Evidence Pollution for Out-Of-Context Misinformation Detection. arXiv preprint arXiv:2501.14728, 2025.