SELF-CALIBRATING AGENTIC GENERATION AND ANSWERING OVER MULTIMODAL INSTRUCTIONAL COLLECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Instructional recordings such as lectures, tutorials and talks convey information through speech, slides and source documents like research papers. Question answering (Q/A) over them therefore requires reasoning across modalities, combining shared with modality-specific information. Existing instructional Q/A benchmarks do not test this together with demanding reasoning constraints such as temporal dependencies, multi-hop connections and long-context reasoning. We propose MM-AdvQAGen, a multimodal Q/A generation pipeline that certifies each question's modality, context and multi-hop requirements through a series of falsifiable tests run by blind solvers. It yields MMAv-InstructBench, built from the seed datasets M^{3}AV (Chen et al., 2024) and SlideSpeech (Wang et al., 2024a): 4,624 questions (1,618 and 3,006) for evaluating answering across four dimensions, *temporal grounding*, *cross-modal synthesis*, *long-context reasoning* and *multi-hop reasoning*. For retrieval and answering over such collections we further propose MM-Graph-RAG, a graph-retrieval pipeline that builds one cross-modal graph per recording, linking slide text, speech transcript and related document content, and derives its retrieval settings from the recording itself rather than per-dataset tuning. On MMAv-InstructBench, MM-Graph-RAG improves judge accuracy over the strongest baseline by up to 10.7 percentage points on M^{3}AV and 16.6 on SlideSpeech. On external datasets of other types it likewise wins against the best baselines by good margins.

1 INTRODUCTION

Instructional content is increasingly delivered through recorded multimodal media rather than text-based resources. Such content in the form of recorded lectures, tutorials and talks has become an increasingly important source of knowledge and expertise (Brame, 2016; Noetel et al., 2021; Poquet et al., 2018), thus creating growing demand for assistants that can answer questions over these recordings for tutoring, search, and accessibility. An instructional recording has an underlying structure: information unfolds over time, with concepts distributed across complementary modalities like speech, slides, and visual information and often linked through dependencies spanning distant portions of recording. Handling questions over such content requires a diverse set of reasoning capabilities — *Temporal reasoning:* for localizing and relating events across time; *cross-modal understanding:* to integrate complementary evidence from speech, slides, and other visual content; *long-context reasoning:* for linking information separated by substantial distance; *multi-hop reasoning* to aggregate evidence across multiple concepts or intermediate steps. A reliable multimodal Q/A system associated with instructional content should have answering capabilities across these varied dimensions, instead of only being able to handle questions pertaining to a limited set of modalities.

Current benchmarks provide complementary directions for understanding instructional video content via diverse tasks. Instructional content driven tasks include audio-visual ASR for understanding specialized terms in the SlideSpeech corpus (Wang et al., 2024a), cross-modal retrieval between spoken language and slide figures as part of the Multimodal Lectures Presentation dataset (Lee et al., 2022) and contextual slide generation, script generation, and speech synthesis in the M^{3}AV (Chen et al., 2024) dataset. Despite this progress, these benchmarks primarily evaluate recognition, retrieval, and generation and don't consider question answering tasks that require joint reasoning over

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

evidence distributed across time and multiple modalities. Further, they don't support detailed evaluation of multimodal Q/A systems across key reasoning capabilities, including temporal grounding, cross-modal understanding, long-context and multi-hop reasoning.

Retrieval methods on long video Q/A typically identify a small set of query-relevant moments (Xu et al., 2026; Wang et al., 2025) or iteratively search for relevant evidence through agentic interaction (Wang et al., 2024b; Fan et al., 2024), whereas multimodal RAG approaches retrieve query-relevant frames, segments, or textual representations (Luo et al., 2026; Ren et al., 2026). While these methods are effective at localizing relevant content, the retrieval tasks they are tested on, don't require the retrieved evidence to be jointly integrated across multiple modalities. This makes it difficult to distinguish successful localization from cross-modal reasoning, especially when the evaluation benchmarks consists of unimodal shortcuts. Further, such retrieval pipelines rely on predefined choices for segmentation granularity, retrieval thresholds, and evidence budgets. Such choices might be sensitive to changes in video lengths, information density, and available modalities, which can vary across different instructional collections. This motivates the design of retrieval methods that can adapt their behavior to each recording rather than relying on dataset specific calibration.

To address the lack of an instructional benchmark that comprehensively evaluates multimodal reasoning capabilities, we propose a scalable Q/A generation pipeline called **MM-AdvQAGen** for generating the benchmark **MMAv-InstructBench** consisting of question-answer pairs with gold evidences from seed datasets like M^3AV and SlideSpeech. **MM-AdvQAGen** verifies the assumption that the information required to answer questions is present across multiple modalities and spread across time. This verification is conducted via a series of falsifiable tests. Every drafted question is attacked by blind solvers which is shown a deliberate impoverished view with modalities removed, truncated context, or isolated hops. The question-answer pair with the required multimodal evidence survives if every attack fails. **MM-AdvQAGen** results in a combined corpus called **MMAv-InstructBench** with $4{,}624$ question-answer pairs across the four dimensions: *temporal grounding*, *cross-modal synthesis*, *long-context reasoning*, *multi-hop reasoning*. For retrieval purposes, we propose **MM-Graph-RAG** that builds a cross-modal graph per recording with linkages between slide text, speech transcript, and related document content from the instructional source. This proposed technique determines the retrieval parameters at inference time using the context of the instructional content and given query. We show that **MM-Graph-RAG** results in gains of up to 10.7 and 16.6 percentage points in judge accuracy over the strongest baseline when tested on **MMAv-InstructBench** with M^3AV and SlideSpeech as seed datasets respectively.

2 Related work

Retrieval-augmented generation. RAG grounds generation in query-relevant external evidence (Lewis et al., 2020), but flat chunk retrieval struggles with dispersed or relational evidence, especially in long video. Video-RAG retrieves OCR, ASR, and object signals to augment video-language models (Luo et al., 2026); subsequent work adds graph-based indexing (Ren et al., 2026) or jointly structures retrieval and reasoning (Shen et al., 2026). Multimodal document RAG likewise finds cross-modal evidence selection and integration challenging (Dong et al., 2026). These methods predefine retrieval granularity and budgets as dataset-level hyperparameters, motivating inference-time calibration to each recording.

Video understanding. A parallel line lowers the visual-token cost of long video by selecting frames; generative and adaptive samplers learn to retrieve query-relevant frames at variable density (Yao et al., 2025). They operate on the frame stream alone and build no persistent cross-modal structure linking speech, slides and documents. Tool-using agents orchestrate foundation models over structured memory: an LLM that iteratively gathers evidence (Wang et al., 2024b), a variant with temporal and object memory that invokes localisation and query tools (Fan et al., 2024), and entity-relation graphs that steer frame selection and reasoning (Chu et al., 2025). Such pipelines are expensive, often depend on proprietary models, and rely on hand-specified control logic and thresholds.

Graph-based and knowledge-graph retrieval. Graph RAG organises a corpus into a knowledge graph and retrieves over nodes and edges, gaining on questions that span large corpora (Edge et al., 2024). Granularity-aware variants trade the token cost of entity-level graphs against the relational

detail lost in document-level ones (Wu et al., 2025), and multimodal variants bring knowledge graphs into RAG-based visual question answering (Yuan et al., 2026).

Instructional and lecture datasets. Lecture-scale corpora supply hundreds of hours of presentation video with human annotation of slide text and speech across disciplines (Chen et al., 2024), and knowledge-based VQA has represented multimodal knowledge as triplets (Ding et al., 2022). Building such resources needs costly manual annotation, and none provides difficulty-stratified question–answer supervision across multiple reasoning dimensions. Because annotation is expensive and existing benchmarks leak, letting models answer without retrieval, recent work generates questions synthetically: semi-synthetic pipelines yield leakage-free, difficulty-preserving benchmarks whose difficulty is governed by graph structure (Liu et al., 2026), and LLM frameworks pair generation with multi-instance evaluation for closed-loop quality control (Siro et al., 2024). These target text corpora, do not span the coordinated multimodal narrative of a recording, and rely on single-pass generation rather than closed-loop verification across reasoning dimensions.

3 METHODOLOGY

We first introduce our multimodal reasoning benchmark **MMAv-InstructBench**: its four reasoning dimensions, the adversarial generation pipeline that produces it (§3.1.1) and its statistics (§3.1.2); we then describe **MM-Graph-RAG**, the cross-modal graph retrieval method for answering questions from MMAv-InstructBench over instructional recordings (§3.2).

3.1 MMAV-INSTRUCTBENCH

MMAv-InstructBench draws on two lecture collections, M^{3}AV (Chen et al., 2024), whose recordings carry speech, slides, and a source paper, and SlideSpeech (Wang et al., 2024a), whose recordings carry speech and slides only. MMAv-InstructBench comprises of a total $4{,}624$ question–answer pairs over $1{,}286$ recordings. Throughout, a *segment* is the time-aligned unit at which a collection releases a recording: one short span of speech together with the slide on screen while it is spoken, so that a recording is an ordered sequence of segments. The retrieval method of §3.2 later re-partitions a recording into its own *fine segments*; wherever a reasoning dimension below refers to segments, it means these released units. The benchmark is organized along four *reasoning dimensions*, each isolating one capability and excluding one shortcut (Table 3 in Appendix A). **Temporal grounding** asks what was shown on-screen at a specific moment, and the answer exists only on the slide; the *easy* tier states the time window in the question, so the slide must be read, while the *hard* tier gives no time reference and identifies the slide by its content alone, so it must first be located, and a question is rejected if its answer appears on more than three slides of the recording. **Cross-modal synthesis** asks a question that no single source among paper, slide and speech, can answer in isolation. **Long-context reasoning** asks about a concept whose treatment is spread over non-adjacent segments, so no single segment suffices. **Multi-hop reasoning** places the answer at the end of an ordered chain of two or three facts, so neither stopping early nor jumping to the last hop suffices. In the result tables the two temporal tiers are abbreviated as p1e (easy) and p1h (hard), and cross-modal synthesis, long-context reasoning and multi-hop reasoning as p2, p3, and p4 respectively after the multi-agent pipelines (P1–P4) that generate each (§3.1.1).

3.1.1 MM-ADVQAGEN: MMAV-INSTRUCTBENCH GENERATION PIPELINES

A recording is denoted as $R = (T, S, D)$: a timestamped transcript $T = \{t_i\}$, slides with OCR $S = \{s_i\}$ aligned to the same segments, and an optional document D. A *blind solver* σ_E is an LLM that must answer a question from an evidence subset $E \subset R$ and nothing else. Every pipeline runs GENERATE $\rightarrow$ VERIFY $\rightarrow$ GATE: a generator drafts (q, a) from full evidence, a pipeline-specific *attack set* $\mathcal{A}(q)$ of impoverished views is issued to blind solvers, and the draft is kept only if every attack fails,

$$\mathrm{NoLeak}(q, a) = \bigwedge_{E \in \mathcal{A}(q)} \big[\mathcal{M}_{\mathrm{hyb}}(\sigma_E(q), a) = 0\big], \tag{1}$$

$$\mathsf{Accept} = \mathrm{Shape}(a) \wedge \mathrm{NoLeak} \wedge (P_Q \geq \theta_Q) \wedge \neg\mathrm{Dup} \wedge G. \tag{2}$$

Here $\mathcal{M}_{\mathrm{hyb}}(\hat{a}, a)$ returns 0 for refusals, 1 for exact/contained matches or numeric agreement within 0.5, and $\mathrm{Judge}(\hat{a}, a)$ otherwise, avoiding most model calls; refusals are retried once with abstention

forbidden. Shape requires a non-empty, non-boolean answer of at most six words (three for P2); P_Q is the score a quality-gate agent assigns to the draft and θ_Q the minimum score for acceptance, so the gate must return an explicit PASS with $P_Q \geq \theta_Q$, where $\theta_Q = 7$ for P1 and 8 otherwise; Dup flags repeated answers with same-video or question-word Jaccard > 0.45; and G verifies that a follows from the cited evidence. Q/A generation pipelines differ only in the agents generating them and their attack sets $\mathcal{A}(q)$:

P1, temporal grounding. A scorer agent ranks segments by askable on-screen content, a selector agent picks a rich slide s_i, and a drafter agent writes a value, time, or context question. Attack set $\mathcal{A} = \{T,\ D,\ T \cup D\}$: no text-only view may yield the answer. Acceptance also requires one positive check, that a solver shown only s_i succeeds, which is how we confirm a is legible on that slide. The hard tier further rejects drafts whose moment is cued by the transcript.

P2, cross-modal synthesis. On M^{3}AV only, one collection with a source document. An aligner agent matches segment i to the paper sentences D_i, a disagreement check confirms the three sources say different things, a bridge agent proposes a synthesis spanning ≥ 2 modalities, and a drafter agent writes (q, a). Attack set $\mathcal{A} = \{D_i, \{s_i\}, \{t_i\}\}$: each single modality must fail.

P3, long-context reasoning. Recurrence finder agent locates a concept that reappears across an adaptive gap, a span builder agent assembles 2–5 non-adjacent segments $\{t_{i_1}, \ldots, t_{i_k}\}$, and a drafter agent writes a question over the span. Attack set $\mathcal{A} = \{\{t_{i_j}\} : j = 1..k\}$: each segment alone must fail.

P4, multi-hop reasoning. A chain discoverer agent summarises each segment and links those summaries into an ordered chain $h_1 \to \ldots \to h_n$, $n \in \{2, 3\}$; a drafter agent then writes a question answerable only at h_n. Attack set $\mathcal{A} = \{h_{1..k} : k < n\} \cup \{\{h_k\} : k \leq n\}$ holds every proper prefix of the chain and every hop on its own, so a question survives only if stopping early fails and no single fact, the last one included, gives the answer. Alignment, duplicate detection and answer comparison are rule-based: a solver's answer is compared with the gold answer by string and numeric rules, and a model is consulted only when those rules are inconclusive. The remaining model calls are segment scoring, drafting, chain discovery, blind solving and the quality gate. No prompt is asked to certify that a question is hard: every impoverished view is handed to a solver that must actually fail, so leak-resistance is measured rather than claimed. Model configurations are in Appendix A.1. Appendix A.2 gives each pipeline's agents, question templates and verifier passes in full.

3.1.2 MMAv-InstructBench Corpus

Table 4 in the Appendix summarizes properties of the generated dataset MMAv-InstructBench. Answers are deliberately concise (1.4–3.0 words on average; at most six), so each can be scored by exact match as well as by the LLM judge that Section 4 specifies for evaluation. Every accepted question also clears the quality gate of §3.1.1, an agent that scores the draft and must return an explicit pass; accepted items average 7.0–9.4, and the gating model is named in Appendix A.1. Long-context questions span a median of 3 segments on M^{3}AV and 5 on SlideSpeech, with median maximum inter-segment gaps of 6 and 97; 86–90% of multi-hop questions require three hops; and 79% of cross-modal questions require all three of paper, slide and speech, a dimension that exists on M^{3}AV alone. The corpora provide complementary regimes: M^{3}AV recordings average 5.0 minutes and 14 aligned segments, SlideSpeech 47.9 minutes and 451 segments, contrasting short, document-backed lectures with long-form, slide-based instruction. Overall, for accurate answers, 4,617 questions require slide evidence, 2,699 require speech evidence, 798 require explicit temporal grounding, and 500 require a source document. A stratified 20% sample of the corpus was verified by three independent human verifiers, who accepted 96.6% of items and caught 90.0% of deliberately corrupted controls; the protocol, agreement statistics and guidelines are in Appendices A.5 and A.6.

3.2 MM-Graph-RAG: Self-Calibrating Cross-Modal Graph Retrieval

MM-Graph-RAG builds *one* cross-modal graph per recording and answers every question against it. It has an *offline* phase that uses only embeddings and closed-form statistics from the recording, and an *online* phase that spends at most three model calls per question (Figure 1): a *classifier* that labels what kind of question was asked, a *controller* that decides whether the retrieved evidence is sufficient, and the *answerer* that writes the answer. A fourth call happens in one case only: when the

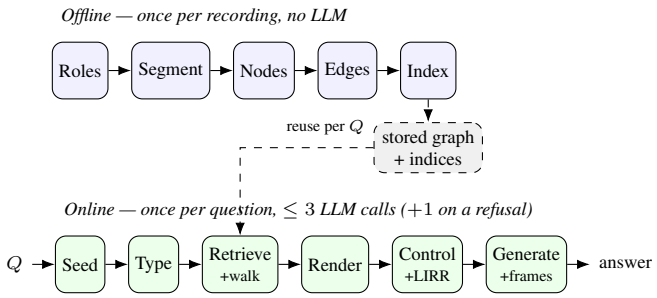

Figure 1: MM-Graph-RAG. *Offline*, a recording becomes one reusable graph from embeddings and closed-form statistics. *Online*, each question is answered against the stored graph with at most three model calls: an optional type classifier, an optional sufficiency controller, and the answerer, which is re-asked once only if it refuses an on-screen value read.

question asks for a value printed on a slide and the answerer replies that it cannot read it, the question is re-asked with more frames. The method is answerer-agnostic: one model plays all three roles, the embedding model (Qwen3-Embedding-8B) is frozen, and either can be swapped without rebuilding the graph (Section 4). Appendix B.1 gives the rationale and motivation for each component below and the measurement supporting it.

Notation. A recording exposes one stream per channel: speech, slides, and a document where one exists. A stream is a list of *units* U_m, a unit being the smallest piece its collection releases, one transcript line or one slide, carrying an optional time span and its text or image; T is the recording's duration. $e(\cdot) \in \mathbb{R}^d$ ($d = 4{,}096$) is a frozen, ℓ_2-normalized text embedding and $\cos(x, y) = e(x)^\top e(y)$. The graph $G = (V, E)$ has four node levels, finest first: fine segments S ($n = |S|$), each a stretch of speech with the slide shown during it, and the unit retrieval returns; aligned items A, the slides themselves; sections C, runs of consecutive segments on one topic; and document passages D. Five edge types $\mathcal{E} = \{\text{TEMPORAL}, \text{ALIGNED}, \text{SECTION}, \text{CROSS_MODAL}, \text{COREF}\}$ join them: three are read straight off this structure and two are similarity-based with per-recording thresholds, so the former move retrieval along the recording and the latter move it across modalities and across time. Appendix B.2 defines each node level and edge type. P_k is the k-th percentile of a sample and $\text{clip}(x, a, b) = \min(\max(x, a), b)$.

3.2.1 Offline: building the graph without a language model

The rationale for each step is in Appendix B.1.

Role planning. Let $\text{coverage}(m)$ denote the fraction of T spanned by modality m, $\text{continuity}(m)$ the fraction of that span covered by its timed units, and $\text{ft}(m)$ the fraction of units with timestamps. We classify m as a timed stream, $m \in \mathcal{T}$, if a majority of its units are timed and its temporal coverage is non-zero; all remaining modalities are treated as *documents*. The *anchor* is the stream that spans the recording, fills it, and is finely divided, $a = \arg\max_{m \in \mathcal{T}} \text{coverage}(m)\,\text{continuity}(m)\ln(1 + |U_m|)$, ties broken deterministically. Other timed streams that span most of the recording are *co-anchors*, fused onto the anchor's units so that transcript and slides jointly drive segmentation; remaining timed streams are *aligned*.

Change-point segmentation. Anchor units are partitioned into *fine segments* at points of information change. Each unit is represented by an embedding of its speech together with the co-anchored OCR active at its midpoint. For each candidate boundary i, we compute four change signals: (i) *semantic drift*, given by $1 - \cos$ between the mean embeddings on either side of the boundary; (ii) *lexical novelty*, $1 - |\tau_i \cap \tau_{i+1}| / \min(|\tau_i|, |\tau_{i+1}|)$, where τ_i denotes the token set of unit i; (iii) *visual shift*, measured as $1 - \cos$ between consecutive key-frame OCR embeddings; and (iv) a binary *discourse cue* indicating transition phrases such as "moving on."

We standardize each signal within the recording and assign boundary i the score $s_i = \sum_x w_x z_{x,i}$, where $z_{x,i}$ is the z-score of signal x. Signal weights are derived from $\tilde{w}_x = \text{KL}(p_x \| \text{Unif}) + \left|\Pr_i[x_i > \bar{x}] - \frac{1}{2}\right|$, with p_x the signal normalized across candidate boundaries, which favors informative, non-uniform signals; the $\tilde{w}_x$ are normalized to a fixed total and clipped to lower and upper bounds, so a weak signal neither vanishes nor dominates. A boundary is inserted when $s_i > \eta^*$,

where η^* is the knee of the recording's sorted boundary scores constrained to an inner percentile range, so a degenerate score distribution produces neither excessive nor no segmentation. Key-frame transitions are always retained as boundaries, after which adjacent segments are merged or split to satisfy duration bounds derived from T.

Nodes on four levels. Each segment forms a *fine* node containing its span, speech, overlapping OCR, key-frame, embedding, and a mild importance prior imp that down-weights meta content, i.e. greetings, outlines, agendas and acknowledgements, by a multiplier on its dense score. Co-anchor/aligned units (e.g. slides) form *aligned* nodes linked to substantially overlapping segments, or the nearest by centre. A coarser pass groups segments into $\sim \sqrt{n}$-bounded *sections*, cutting at low consecutive-segment cosine similarity. Document passages form untimed *document* nodes. Table 12 in Appendix B.1 states what each level is for.

Edges. Five edge types give retrieval five ways to move through a recording, one per way its evidence is distributed. TEMPORAL joins consecutive segments, so retrieval walks forward and backward in time. ALIGNED joins a segment to the slides displayed while it was spoken, turning a question about a spoken moment into the slide that answers it. SECTION joins a section to its members, a coarse unit for questions about a stretch of the lecture. CROSS_MODAL joins a segment to a document passage on the same subject, the only path from lecture to source paper. COREF joins segments sharing a salient term, gathering far-apart mentions of one concept. The first three are read off the structure; the last two are similarity-based, thresholded per recording by Otsu's rule (Otsu, 1979) and by a document-frequency band, with bounded out-degree so long recordings stay sparse. Appendices B.2 and B.1 give both thresholds and map each type to the dimension it serves.

Indexing. The leaves $L = S \cup A \cup D$ are stacked into an embedding matrix and a BM25 index (Robertson & Zaragoza, 2009); graph, indices and imp are stored and shared by every question. The two indices are what tell retrieval where to start: an embedding matches a paraphrase while BM25 matches a rare symbol or number exactly, and lecture material needs both.

3.2.2 ONLINE: ANSWERING WITH A LANGUAGE MODEL IN THE LOOP

Leaf seeding. For $q = e(Q)$, a dense ranking by $\mathrm{imp}(v)\cos(q, v)$ and a BM25 ranking over L are fused by reciprocal rank fusion (Cormack et al., 2009); the top fine segments, verbalised in time order, form a preview E_{pre}.

Question tagging. Each question Q is assigned a type $\tau \in \{\text{TEMPORAL}, \text{CROSS_MODAL}, \text{MULTI_HOP}, \text{LONG_CONTEXT}, \text{GENERAL}\}$. A keyword scan proposes (τ_0, conf) and is accepted when unambiguous; otherwise, a single temperature-0 LLM call classifies (Q, E_{pre}). We additionally extract two question-intrinsic signals: an explicit time window $\omega(Q)$, when present, and a binary temporal flag $b(Q)$, obtained by regex with nearest-prototype cosine fallback. After retrieval, GENERAL is promoted to CROSS_MODAL or LONG_CONTEXT if the corresponding edge type dominates the retrieved subgraph.

Structure-aware retrieval with a typed walk. Budgets scale sub-linearly, $|\mathrm{seeds}|(n) = 2\lceil \log_2 n \rceil$ and $|\mathrm{evidence}|(n) = \lceil 3.5\sqrt{n} + 2 \rceil$ (both n when $n \leq 2$), so a recording twice as long gets more evidence but not twice as much, keeping the decisive segment from being diluted. Seeds come from the same RRF fusion, with nodes inside $\omega(Q)$ prepended. If $b(Q)$, retrieval stays local, adding members of a few best-matching sections; otherwise graph signals widen it with members of the top sections (a fixed fraction of $|C|$), each seed's temporal neighbours, $k_{\mathrm{sp}} \propto n(1-\rho)$ segments spread over the timeline, ρ being the fraction of T the seeds already span, and $\lceil \sqrt{|D|} \rceil$ document and as many aligned nodes chosen by *consensus*: $\frac{1}{2}(\cos(q, v) + \cos(q, s_v^*))$, plus $\min(\cos(q, v), \cos(q, s_v^*))$ when the linked segment s_v^* best matching Q exceeds the median segment cosine, so a passage is seeded for corroborating a relevant spoken moment rather than for its own score. The walk then orders edge types per question: with N_ϵ the nodes within two ϵ-hops of the leading seeds, $\pi(\epsilon) = \mathrm{aff}_\epsilon \cdot \mathrm{nov}_\epsilon \cdot \frac{|N_\epsilon|}{|V|}(1+\nu_\epsilon)(1+b(Q)\phi_\epsilon)$, each factor additively smoothed, for aff_ϵ the mean top $\cos(q, \cdot)$ in N_ϵ, nov_ϵ its mean distance from the seed centroid, ν_ϵ its fraction of node levels absent from the seeds and ϕ_ϵ its timed fraction. Types expand in decreasing π until $|\mathrm{evidence}|(n)$ nodes are collected, with a bounded share of documents.

Evidence rendering. The subgraph is verbalised as one time-ordered block, each segment as $[\mathrm{S}_i$ @ t_0–$t_1]$ SPEECH: … ON-SCREEN TEXT: …, aligned nodes as $[\mathrm{V}_j]$, passages as $[\mathrm{A}_k]$,

followed by its edges written as relations ("then", "shows", "grounds", "same-topic", "part-of"), so the answerer sees how the pieces relate.

Sufficiency control and latent re-retrieval. A controller call at temperature 0 returns $\{\text{sufficient}, \text{followup}\}$; only when a specific fact is missing are the named terms retrieved as above and merged under a fixed cap (Table 21 reports how often this fires). For non-temporal questions answered without frames, *latent iterative re-retrieval* (LIRR) then widens coverage with no model call: the query is shifted toward the retrieved centroid $\bar{c}$, $q' = \text{norm}(\alpha q + (1-\alpha)\bar{c})$, where α increases with the timeline fraction ρ' the evidence spans and with its level diversity δ, so clustered evidence explores further and spread evidence stays near Q; the $\Theta(\sqrt{n})$ sections closest to q' each add their best members under an enlarged budget.

Grounded generation with latent visual verification (LVV). The answerer receives the block, the video and a type-specific instruction and returns a 1–8 word answer at temperature 0, the one call that always runs. When the question asks for an on-screen value ($b(Q)$, a named window, or a visual-locative cue such as "arrow"), a few frames are extracted across $\omega(Q)$ if given, else across the retrieved segment closest to q, and value questions are answered from the frames alone (a refusal is re-asked once with more frames, the only case in which the answerer is called twice); for reasoning types the frames are added to the video and text, and stored key-frames are used when the video file is unavailable. Every online stage is *total*: it always returns a result, the deliberate opposite of the generation pipeline (Section 3.1.1), which rejects and regenerates.

4 EXPERIMENTS

We ask three questions. **(Q1)** Does MM-Graph-RAG outperform strong video-RAG, agentic and visual-retrieval systems on MMAv-InstructBench? **(Q2)** Which component produces the gain, and for which reasoning dimension? **(Q3)** Does it transfer, without recalibration, to benchmarks we did not generate? Inference cost is analysed in Appendix G.

Task and benchmarks: Given a recording R and question q, a system returns a short answer a. We evaluate on both MMAv-InstructBench collections: M^3AV (1,618 questions; all four dimensions and both temporal tiers; the only collection with source documents) and SlideSpeech (3,006 questions; recordings $\sim 10\times$ longer; P1e, P3, P4), Table 4. To test generalization, **MM-Graph-RAG** is also evaluated on two external multiple-choice benchmarks not generated by us: **TVQA** (Lei et al., 2018; Ataallah et al., 2024), full television episodes with subtitles, and **Video-MME** (Fu et al., 2025).

Baselines: We compare five groups: (i) *no retrieval*: **Zero-shot** (uniform frames and full transcript in one call) and **Agentic** (a planner call followed by an answer call over the same input); (ii) *caption-then-read frame selection*: **LLoVi** (Zhang et al., 2024), **VideoAgent** (Wang et al., 2024b), **VideoTree** (Wang et al., 2025), and **VGent** (Shen et al., 2026); (iii) *video RAG*: **Video-RAG** (Luo et al., 2026) and **VideoRAG-Ren** (Ren et al., 2026); (iv) *document/graph RAG*: **RAG-Anything** (Guo et al., 2025); and (v) *retrieval-centric multimodal*: **ColPali** (Faysse et al., 2025), **CLaMR** (Wan et al., 2025), and **M2A** (Feng et al., 2026). Video-only baselines receive decoded frames on every collection. On M^3AV, frame-selection methods and VideoRAG-Ren are run both *faithful* (video+transcript) and *+docs* (also given the source document); Table 1 reports the stronger variant, while Appendix D reports both (f/d; +docs improves results by 6–16 points across answerers). M2A uses tool calling when supported. Full inputs, budgets, and adaptations are in Appendix I.

Models and protocol: All systems use the same answerer, isolating retrieval from generation, and experiments are repeated with three open-weight models at temperature 0: **gemma-3-27b-it** (Team et al., 2025), **LLaVA-OneVision-7B** (Li et al., 2024), and **Qwen3-VL-8B-Instruct** (Bai et al., 2025) (gemma, LLaVA-OV, and Qwen3-VL). MM-Graph-RAG uses the answerer itself as classifier/controller and **Qwen3-Embedding-8B** for dense retrieval. Appendix I gives serving and evaluation details.

Metrics: Open-ended answers use *judge accuracy*, *relaxed exact match* (REM), and *token F1*. The judge, **Llama-3.3-Nemotron-Super-49B** (Bercovich et al., 2025) at temperature 0, receives question, reference, and prediction and returns a binary verdict: synonyms/standard abbreviations count, while partial answers, incorrect additions, or doubt are wrong. REM tests equality after lower-casing and removing articles, punctuation, and extra whitespace; token F1 uses the same normalization. REM and F1 for all systems appear in Appendix D. Multiple-choice benchmarks use judge-free letter accuracy with option-text matching. We report 95% bootstrap intervals (10^4 resamples) and

exact McNemar tests against the strongest baseline on shared questions. An independent human audit of the judge on 20% of the M^{3}AV predictions is reported in Appendix H.1.

5 RESULTS

5.1 MAIN COMPARISON

	M^{3}AV ($n=1{,}618$)			SlideSpeech ($n=3{,}006$)		
System	gemma	LLaVA-OV	Qwen3-VL	gemma	LLaVA-OV	Qwen3-VL
Zero-shot	37.4	33.2	39.2	39.4	34.2	42.0
Agentic	34.3	29.7	37.0	33.2	22.6	26.6
LLoVi	45.3	41.1	47.3	31.5	30.6	34.6
VideoAgent	45.8	39.4	47.3	31.2	27.5	34.4
VideoTree	36.5	36.7	40.7	16.2	21.0	21.3
VGent	37.0	31.2	39.2	21.0	19.0	23.6
Video-RAG	52.4	47.6	58.0	31.8	31.7	41.4
VideoRAG-Ren	41.9	38.0	50.8	36.4	34.7	39.1
RAG-Anything	36.0	8.8	19.8	19.7	14.7	21.6
M2A	41.8	31.8	41.5	35.4	22.1	34.4
ColPali	40.4	37.6	66.8	32.5	28.4	48.6
CLaMR	55.7	52.5	61.2	39.1	41.2	46.0
MM-Graph-RAG	**66.4**	**61.1**	**69.2**	**56.0**	**50.1**	**50.2**
Δ vs. best baseline	$+10.7$	$+8.5$	$+2.4$	$+16.6$	$+8.9$	$+1.8$
McNemar p	2.9×10^{-19}	3.1×10^{-13}	2.3×10^{-3}	1.3×10^{-58}	1.1×10^{-19}	7.9×10^{-2}

Table 1: Judge accuracy (%) on MMAv-InstructBench. Every system answers through the same answerer within a column; frame-selection baselines and VideoRAG-Ren are shown in their stronger (+docs) configuration on M^{3}AV. Δ and the p-value are paired on the questions both compared systems answered, so they need not equal the difference of the rounded column entries. Per-dimension results for every system are in Appendix D.

Table 1 reports judge accuracy on both collections under all three answerers; MM-Graph-RAG is highest in all six columns, with every margin and exact McNemar p-value in the table. Over the strongest baseline, CLaMR, it gains $+10.7$ and $+8.5$ points on M^{3}AV under gemma and LLaVA-OneVision (95% CI $[+8.4, +13.0]$ and $[+6.1, +10.7]$), and on SlideSpeech, whose recordings are ten times longer, the margins widen to $+16.6$ and $+8.9$ ($[+14.6, +18.5]$ and $[+7.1, +10.9]$); under Qwen3-VL the SlideSpeech margin over ColPali is $+1.8$ ($[-0.2, +3.8]$, $p = 0.079$), within noise. Judge-free metrics (Table 19, Appendix D) put MM-Graph-RAG highest on REM and F1 in every column, including the closest Qwen3-VL cases (39.8 vs. 37.2 REM on M^{3}AV, 34.6 vs. 28.8 on SlideSpeech), and across all 78 system–answerer–collection cells 97.5–100% of exact string matches are also judged correct.

Per-Dimension Results. Table 2 splits M^{3}AV by reasoning dimension. Gains over CLaMR are largest on cross-modal synthesis ($+15.1$), multi-hop ($+12.2$) and temporal grounding ($+12.8$ P1e, $+9.8$ P1h), the tiers MM-AdvQAGen certified not text-answerable, and smallest on long-context ($+1.7$). Against the gold-evidence oracle, which bounds retrieval, MM-Graph-RAG closes about three quarters of the zero-shot gap overall and matches it on multi-hop (81.1 vs. 80.2, one question on a 106-item pool). In 72.6% of remaining errors every gold span already sits among the top eight retrieved nodes, so headroom lies in answering, not retrieval; fully covered on only 29.3% of questions (Appendix H).

System (gemma answerer)	P1e	P1h	P2	P3	P4	overall
Zero-shot	24.1	28.9	46.2	41.2	55.7	37.4
ColPali	29.4	32.1	49.4	41.5	57.5	40.4
Video-RAG	43.6	61.6	50.8	53.8	60.4	52.4
CLaMR	48.6	68.2	52.4	52.9	68.9	55.7
MM-Graph-RAG	**61.4**	**78.0**	**67.5**	**54.6**	**81.1**	**66.4**
Gold-evidence oracle	71.7	84.3	81.4	62.0	80.2	75.6

Table 2: MMAv-InstructBench (M^{3}AV) judge accuracy (%) by reasoning dimension. The oracle answers from the gold evidence stored with each item and upper-bounds retrieval. All answerers and systems are in Appendix D; the oracle under LLaVA-OneVision is in Table 27.

Where the gain comes from. Because all systems share the same answerer, accuracy differences arise from the context they provide. Three analyses isolate this effect.

Retrieval quality: Figure 2 (right) compares retrieved evidence directly with gold spans i.e., stored evidence segments. At three nodes, the graph walk reaches 76.7% recall versus 63.1% for BM25 and 59.1% for dense retrieval over identical embeddings, with the ordering preserved across budgets. The gap is largest on single-slide lookup (99.7% vs. 54.3%/50.7%), where the walk seeds the nodes inside the time window the question states and follows speech–slide edges, whereas text similarity has no reason to rank a slide first (Table 25). The graph also retrieves more gold evidence per second of video consulted than CLaMR at every budget (Table 26).
Evidence, not prompting: To control for system-specific prompts, we feed MM-Graph-RAG and CLaMR evidence to the same neutral answerer under identical prompts and character budgets (Appendix E.5). MM-Graph-RAG evidence improves accuracy by $+8.7$ ($p = 3.1\times10^{-10}$, $n = 1{,}617$), concentrated on cross-modal ($+22.2$) and multi-hop ($+9.4$) versus $+0.9$, 2.6 elsewhere (Figure 5).
Ablations: Figure 2 (left) isolates each mechanism. Removing vision costs 22.4 points, dwarfing every retrieval ablation (Appendix A.7). Flat dense retrieval in place of the graph costs 2.0 points overall, rising to 3.2 on p3+p4 for gemma and 5.8 for LLaVA-OneVision ($[+2.1, +9.7]$, $p = 0.0041$); removing LVV costs 3.9 and 5.2. Calibration matters little: fixed statistics shift accuracy by < 1 point, halving or doubling the evidence budget by ≤ 0.7, and oracle routing not at all. Four added minutes of irrelevant evidence change nothing either (Appendix E.6), so coverage, not dilution, is the retrieval bottleneck. Figure 4 gives full attribution.

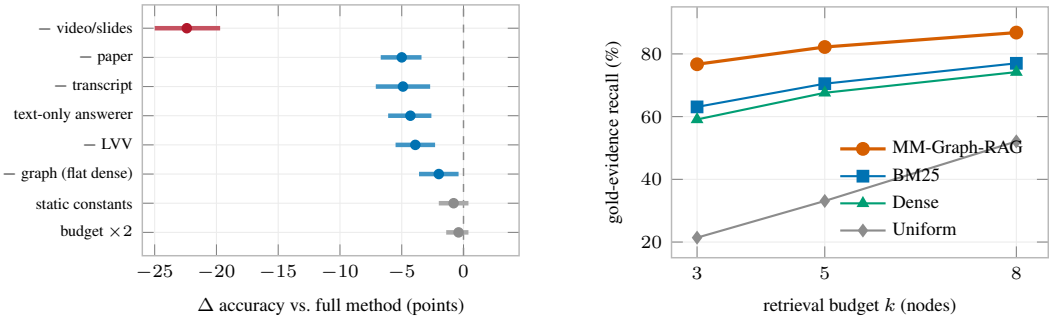

Figure 2: **Left:** paired ablations on M^{3}AV (n=1,618, gemma) with 95% bootstrap intervals; colour marks modality removal (red), retrieval structure (blue) and effects within noise (grey). **Right:** gold-evidence recall against retrieval budget. Full results in Appendices E and H.

5.2 Transfer to public benchmarks

We report MM-Graph-RAG on the public benchmarks **TVQA** and **Video-MME** in Tables 23 and 24 respectively. On **TVQA**, despite having no slides or documents, MM-Graph-RAG leads under all three answerers, over CLaMR/Video-RAG (2.0, $+2.5$) too by relatively small margin, but substantially over zero-shot $+15.9$. On **Video-MME**, it leads with all the models with margins of ($+1.2$, $+2.7$, 3.0).

6 Conclusion

We presented two components for Q/A over instructional recordings. **MM-AdvQAGen** keeps a question only when adversarial solvers deprived of the required evidence fail, so each needs several modalities, long context or several hops; on M^{3}AV and SlideSpeech it yields **MMAv-InstructBench**, 4,624 auditable questions across four reasoning dimensions, each with its validating evidence. **MM-Graph-RAG** builds one cross-modal graph per recording and sets segmentation, linkage and evidence budgets at inference time, under one code path and constants across all three collections. It reaches the highest judge accuracy under every answerer, by 1.8–16.6 points over the strongest of twelve baselines, with self-calibration buying portability rather than accuracy (Table 1).

Limitations. (i) SlideSpeech covers three of five tiers: it has no paper, so cross-modal synthesis is M^{3}AV-only, and we did not generate its hard temporal tier. (ii) We answer in a few words, not explanations. (iii) The questions inherit their generators' biases, which the gates and human verification limit but do not remove.

REPRODUCIBILITY STATEMENT

Appendix A gives the generation configuration and audit trail of MMAv-InstructBench, Appendix B the implementation constants of MM-Graph-RAG, and Appendix I the serving setup, the configuration of every baseline and the evaluation protocol. The corpus, the generation and evaluation code, the per-item audit trail and the graded predictions of every system are released.

AI USE STATEMENT

Generative AI appears in this work in two distinct roles, one as an object of study and one as an assistant, and we disclose both.

As part of the research. Large language and vision-language models are the subject of this paper. MM-AdvQAGen calls GPT-5.4, Gemini 2.5 Pro and Claude Opus 4.8 to draft, attack and gate the questions of MMAv-InstructBench, the evaluated answerers are three open-weight vision-language models, and scoring uses an open-weight LLM judge. Every model, its role and its temperature are listed in Appendix A.1 and Appendix I. These uses are contributions of the work rather than writing assistance, and their outputs were audited by people: three human verifiers checked a stratified 20% sample of the corpus (Appendix A.5), and an independent human auditor re-scored the judge on 20% of the M^{3}AV predictions (Appendix H.1).

As an assistant. We used a general-purpose LLM to help edit the manuscript, chiefly to tighten prose, shorten the abstract and conclusion, and check that numbers quoted in the text agree with the tables, and to help write and debug parts of the evaluation code. Every such passage was read, revised and verified by an author against the underlying results, and the code was reviewed and tested before use. No research idea, design decision, experimental protocol or reported number originates from an AI assistant; all results come from runs the authors configured and executed. We have reviewed all AI-assisted work and we take responsibility for the final content of this paper, including its text, its claims and its released artifacts.

ETHICS STATEMENT

Data. MMAv-InstructBench is built on two publicly released research corpora, M^{3}AV (Chen et al., 2024) and SlideSpeech (Wang et al., 2024a), and the two transfer benchmarks, TVQA (Ataallah et al., 2024) and Video-MME (Fu et al., 2025), are also public. We use each for research within its published terms, and we rely on the consent and licensing arranged by the original dataset creators. Our contribution on top of them is question–answer pairs with evidence pointers; we do not redistribute the source audio or video, which users obtain from the original providers. The recordings are academic lectures and talks. We make no attempt to identify any individual, and every question concerns lecture content rather than the people in the recording.

People who took part. Three verifiers rated a 20% sample of the corpus and one independent auditor re-scored the judge's verdicts. All are researchers in this area who took part voluntarily and without monetary compensation, and all were told how their judgements would be used. We recorded only their per-item ratings and short written reasons, no personal data, and we report their results in aggregate (Appendices A.5 and H.1).

Release. We release the corpus, the generation and evaluation code, the per-item audit trail and the graded predictions of every system. The audit trail makes each accepted question re-checkable item by item, and the 31 sampled items that at least two verifiers rejected are released with the corpus rather than quietly dropped.

Risks and limits. The questions were written by large models and inherit their topical and stylistic biases. The adversarial gates and the human verification bound this but do not remove it, and residual errors will exist in the portion of the corpus that was not sampled for verification. A system that answers questions over lectures could be over-trusted in a teaching setting, so we report accuracy per reasoning dimension and against a gold-evidence oracle rather than a single headline number, which keeps the failure modes visible. We see no dual-use or security concern specific to this work.

Competing interests. The authors declare none. Funding is acknowledged in the camera-ready version.

REFERENCES

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos. In *European Conference on Computer Vision*, pp. 251–267. Springer, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, et al. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*, 2025.

Cynthia J Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE—Life Sciences Education*, 15(4):es6, 2016.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. M3av: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9041–9060, 2024.

Meng Chu, Yicong Li, and Tat-Seng Chua. Understanding long videos via llm-powered entity relation graphs. *arXiv preprint arXiv:2501.15953*, 2025.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5079–5088. IEEE, 2022.

Kuicai Dong, CHANG YUJING, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimomal generation for document question answering. *Advances in Neural Information Processing Systems*, 38, 2026.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Junyu Feng, Binxiao Xu, Jiayi Chen, Mengyu Dai, Cenyang Wu, Haodong Li, Bohan Zeng, Yunliu Xie, Hao Liang, Ming Lu, et al. M2a: Multimodal memory agent with dual-layer hybrid memory for long-term personalized interactions. *arXiv preprint arXiv:2602.07624*, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*, 2025.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. Multimodal lecture presentations dataset: Understanding multimodality in educational slides. *arXiv preprint arXiv:2208.08080*, 2022.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1369–1379, 2018.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Jiayi Liu, Jiaxing Zhang, Bowen Jin, and Jennifer Neville. Generating leakage-free benchmarks for robust rag evaluation. *arXiv preprint arXiv:2605.08838*, 2026.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *Advances in Neural Information Processing Systems*, 38:168008–168033, 2026.

Michael Noetel, Shantell Griffith, Oscar Delaney, Taren Sanders, Philip Parker, Borja del Pozo Cruz, and Chris Lonsdale. Video improves learning in higher education: A systematic review. *Review of educational research*, 91(2):204–236, 2021.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi: 10.1109/TSMC.1979.4310076.

Oleksandra Poquet, Lisa Lim, Negin Mirriahi, and Shane Dawson. Video and learning: a systematic review (2007–2017). In *Proceedings of the 8th international conference on learning analytics and knowledge*, pp. 151–160, 2018.

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2390–2401, 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. *Advances in Neural Information Processing Systems*, 38:100805–100830, 2026.

Clemencia Siro, Yifei Yuan, Mohammad Aliannejadi, and Maarten de Rijke. Agent-cq: Automatic generation and evaluation of clarifying questions for conversational search with llms. *arXiv preprint arXiv:2410.19692*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

David Wan, Han Wang, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Clamr: Contextualized late-interaction for multimodal content retrieval. *arXiv preprint arXiv:2506.06144*, 2025.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. Slidespeech: A large scale slide-enriched audio-visual corpus. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11076–11080, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025.

Yaxiong Wu, Jianyuan Bo, Yongyue Zhang, Sheng Liang, and Yong Liu. Query-centric graph retrieval augmented generation. *arXiv preprint arXiv:2509.21237*, 2025.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models. *IEEE Transactions on Multimedia*, 2026.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17900–17917, 2025.

Xu Yuan, Liangbo Ning, Qingqing Ye, Wenqi Fan, and Qing Li. mkg-rag: Leveraging multimodal knowledge graphs in retrieval-augmented generation for knowledge-intensive vqa. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2274–2285, 2026.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024.

APPENDIX

The appendix is organised so that each part answers one question a reader may raise about the main text. Appendix A documents how MMAv-InstructBench was generated and verified, and tests whether its modality labels predict system behaviour. Appendix B lists the implementation constants of MM-Graph-RAG. Appendix D gives every system on every collection and answerer by reasoning dimension, with judge-free metrics. Appendix E reports all ablations, the per-step attribution, the controller, and three attribution experiments. Appendix F reports the two public benchmarks in full. Appendix G reports inference cost. Appendix H isolates retrieval from generation, bounds the remaining headroom, audits the judge, and tests robustness to recording length. Appendix I gives the serving setup, the baseline configuration and the evaluation protocol. Accuracies are in percent; where an analysis uses fewer questions than the full pool, its caption states the count.

A BENCHMARK CONSTRUCTION AND VALIDATION

Dimension (pipeline)	Capability isolated	Shortcut excluded
Temporal grounding, easy / hard (P1; p1e / p1h)	read one slide at one moment; hard tier must first locate the slide	answer spoken, or in the text
Cross-modal synthesis (P2; p2)	combine ≥ 2 of paper, slide, speech	any one source suffices
Long-context reasoning (P3; p3)	combine ≥ 2 non-adjacent segments	one segment suffices
Multi-hop reasoning (P4; p4)	follow an ordered chain of 2–3 hops	early stop, or a lone hop

Table 3: The four reasoning dimensions of MMAv-InstructBench. Each is an evaluation axis defined by the capability it isolates and the shortcut its generation pipeline rules out.

Corpus	Dimension	Items	Rec.	$\lvert q\rvert$	$\lvert a\rvert$	Num. (%)	$\bar{P}_Q$
M^{3}AV	Temporal, easy	381	60	17.0	1.8	40.9	8.48
	Temporal, hard	305	52	17.7	2.1	15.7	7.00
	Cross-modal	500	201	20.4	2.2	8.0	8.54
	Long-context	326	199	26.8	3.0	1.8	9.19
	Multi-hop	106	90	27.8	2.0	17.0	9.08
SlideSpeech	Temporal, easy	112	33	16.6	1.4	56.2	8.45
	Long-context	1852	647	39.3	2.4	2.1	9.25
	Multi-hop	1042	372	26.0	2.1	8.0	9.43
Total		4624	1286				

Table 4: MMAv-InstructBench statistics per corpus and dimension. Rec. is the number of distinct recordings contributing items; $|q|$ and $|a|$ are mean question and answer length in words; Num. is the share of numeric answers; $\bar{P}_Q$ is the mean quality-gate score of accepted items. Cross-modal synthesis requires a source document and so exists only on M^{3}AV; the structural properties of each dimension are given in the text.

A.1 Generation configuration

Transcripts and slide text are taken as released by each collection, and SlideSpeech slide frames are extracted at the released segment boundaries. Drafting is done by GPT-5.4 at temperature 0.3; every verification call, comprising the blind solvers, the quality gate, the ground-truth verifier and the paraphrase judge inside $\mathcal{M}_{\mathrm{hyb}}$, runs at temperature 0. Blind solving, quality gating and judging use Gemini 2.5 Pro, which also reads slide images; the ground-truth verifier uses GPT-5.4; the synthesis-bridge and chain-discovery agents of P2 and P4 use Claude Opus 4.8; and paper–segment alignment uses $\mathrm{text\text{-}embedding\text{-}3\text{-}large}$. Each pipeline attempts up to five items per recording with at most three regeneration attempts per draft, long-context spans require a gap of at least four segments, and the hard temporal tier admits only drafts whose answer string occurs in the OCR of at most three slides. No evaluation-side model (Section 4) took part in generation or verification, so the evaluated answerers and the judge are independent of the models that wrote the corpus.

Audit trail. Each accepted item stores the evidence that justified it. P1 items keep the slide image, its OCR text, the aligned speech, the time range, the visual-only elements the answer depends on and a statement of why the answer is absent from speech and paper; P2 items keep the matched paper sentences, the synthesis pattern and the per-modality blind-solver answers; P3 items keep the anchor concept, the segment indices and the segment gap; P4 items keep the ordered hop list with the modality and segment of each hop. Every blind-solver verdict, quality score and rejection reason is stored with the item, so the corpus is re-checkable item by item.

A.2 The four generation pipelines in detail

Notation. For a string z, let $W(z)$ be its content words, the lower-cased alphabetic tokens of length ≥ 3 that are not stopwords, and let $J(z, z') = |W(z) \cap W(z')|/|W(z) \cup W(z')|$ be their Jaccard similarity. A recording has N segments; segment i carries OCR text o_i, speech text u_i, key-frame image x_i and span $[\tau_i^-, \tau_i^+]$. Write $|z|$ for character length, $n_w(z)$ for whitespace-token count, $\mathrm{num}(z)$ for the event that z contains a decimal number, and $\mathrm{up}(z)$ for its upper-case character ratio. Two *richness* functions map a string to $[0, 1]$:

$$\rho_{\mathrm{o}}(o) = \mathbf{1}[|o| \geq 10] \cdot \min\Big(1,\ \min\big(1, \tfrac{|W(o)|}{15}\big) + 0.15\,\mathbf{1}[\mathrm{num}(o)]\Big) \cdot 0.4^{\,\mathbf{1}[|o|<25]} \cdot 0.7^{\,\mathbf{1}[\mathrm{up}(o)>0.6]}, \tag{3}$$

$$\rho_{\mathrm{s}}(u) = \mathbf{1}[|u| \geq 15] \cdot \min\Big(1,\ \min\big(1, \tfrac{n_w(u)}{80}\big) + 0.1\,\mathbf{1}[u \text{ contains a full stop}]\Big) \cdot 0.4^{\,\mathbf{1}[n_w(u)<10]}. \tag{4}$$

Both are cheap and deterministic, and both are deliberately forgiving: the number bonus in equation 3 protects slides whose askable content is numeric, while the two multiplicative penalties suppress short fragments and all-caps OCR noise. Write $\mathcal{R}(a, b) = \{i : \rho_{\mathrm{o}}(o_i) \geq a,\ \rho_{\mathrm{s}}(u_i) \geq b\}$ for the segments passing a richness floor, ranked by $0.55\,\rho_{\mathrm{o}}(o_i) + 0.45\,\rho_{\mathrm{s}}(u_i)$. For a recording with

paper sentences $\{d_j\}$, the alignment of segment i is the top-k set

$$A_i^{(k,\tau)} = \text{top-}k_j \left\{ d_j : J(o_i \| u_i, d_j) \geq \tau \right\}, \tag{5}$$

where $\|$ is concatenation. All pipelines share the shape gate, the duplicate gate at question-word Jaccard > 0.45, the matcher $\mathcal{M}_{\text{hyb}}$ of Section 3.1.1, $\text{QA}_{\max} = 5$ attempted items per recording and $\text{retry}_{\max} = 3$ regenerations per draft; every blind solver that refuses is re-issued once under an abstention-forbidden prompt before the refusal counts as a failure. Table 5 collects the pipeline-specific constants.

Quantity	P1	P2	P3	P4
Richness floor $\mathcal{R}(a, b)$	$(0.05, 0.08)$	$(0.12, 0.20)$	see equation 12	–
Paper alignment (k, τ) in equation 5	–	$(8, 0.01)$	–	–
Quality bar θ_Q	7	8	8	8
Answer length (words)	≤ 6	≤ 3	≤ 6	≤ 6
Attack-set size $\lvert\mathcal{A}(q)\rvert$	3	3	$k \in [2, 5]$	$2n - 1$

Table 5: Pipeline-specific constants of MM-AdvQAGen. n is the hop count of a P4 chain, k the span length of a P3 question.

P1, temporal grounding. Selection is a three-stage cascade. The candidate pool is $\mathcal{R}(0.05, 0.08)$; its ten highest-ranked members are then scored by a vision model reading x_i, giving an integer v_i, and any segment with $v_i < 3$ is dropped, which removes title and near-blank slides. One segment is drawn from the five survivors with

$$\Pr[i] = \frac{\max(v_i, 0.01)}{\sum_j \max(v_j, 0.01)}, \tag{6}$$

so sampling is proportional to askable visual content rather than deterministic, which diversifies the slides chosen across repeated attempts on one recording. A visual-content agent then returns, from x_i alone, the elements present only visually, the numbers on the slide and a preferred question target. The drafter instantiates one of three templates: TYPE A asks for a value or count on the slide inside the stated window $[\tau_i^-, \tau_i^+]$; TYPE B asks for the time at which a slide matching a description first appears, so the answer is τ_i^- and lives in the metadata rather than in any channel; TYPE C asks for a slide element together with the speech uttered while it is displayed. Verification is four passes:

$$\text{P1}_1 : \mathcal{M}_{\text{hyb}}\big(\sigma_T(q), a\big) = 0 \quad \text{on the words timestamped inside } [\tau_i^-, \tau_i^+], \tag{7}$$

$$\text{P1}_2 : \mathcal{M}_{\text{hyb}}\big(\sigma_D(q), a\big) = 0 \quad \text{(skipped when the recording has no paper)}, \tag{8}$$

$$\text{P1}_3 : \sigma_{\{x_i\}}(q) \text{ succeeds and } \text{VisNec} = c_{\text{vis}} - c_{\text{blind}} > 0, \tag{9}$$

$$\text{P1}_4 : P_Q \geq \theta_Q = 7 \text{ and the timestamp is required.} \tag{10}$$

Pass P1_1 uses word-level timestamps so the leakage test covers exactly the speech spoken over that slide, not the whole transcript. Pass P1_3 is the only *positive* requirement in the pipeline, and VisNec makes it strict: the image must raise the solver's confidence above what the text-only context already gives. An answer occurring in o_i is admissible, since OCR is a function of x_i and reading x_i is the capability under test; leakage is defined against T and D only. The hard tier substitutes three templates that carry no time reference and identify the slide by content, adds a transcript-cue rejection, and enforces uniqueness, $\left|\{ j : a \subseteq o_j \}\right| \leq 3$, so a content description picks out at most three slides of the recording.

P2, cross-modal synthesis. Defined only on recordings with a document, hence M^{3}AV. Candidates are $\mathcal{R}(0.12, 0.20)$, aligned by equation 5 with $(k, \tau) = (8, 0.01)$, and ranked by

$$s_i = 0.7 \underbrace{\left[0.55\rho_{\text{o}}(o_i) + 0.45\rho_{\text{s}}(u_i)\right]}_{\text{richness}} + 0.3 \max_{d \in A_i} J(o_i \| u_i, d), \tag{11}$$

so a segment is preferred for being individually informative first and paper-adjacent second; the floor $\tau = 0.01$ is set far below the useful matching regime on purpose, so that weak pairings still reach the bridge agent rather than being filtered by a string heuristic. A modality isolator computes the uniquely-held content of each channel with no model call, as $W(o_i) \setminus (W(u_i) \cup \bigcup_{d \in A_i} W(d))$ and its two permutations. The bridge agent then proposes a synthesis in one of four shapes, paper+slide, paper+speech, slide+speech, or all three, and the drafter writes (q, a) with $|a| \leq 3$ words. The attack set is $\mathcal{A} = \{A_i, \{x_i\}, \{u_i\}\}$: each channel is issued alone and each must fail. Gate at $\theta_Q = 8$.

P3, long-context reasoning. Every segment receives

$$q_i = 0.4\,\rho_{\mathrm{o}}(o_i) + 0.4\,\rho_{\mathrm{s}}(u_i) + \min\Big(\tfrac{|K_i|}{10},\, 0.2\Big), \qquad q_i \text{ kept iff } q_i \geq 0.15, \tag{12}$$

where K_i is the key-term set a vision model extracts from x_i, so entity density enters the score from the image rather than from OCR; a cheap prefilter $0.3\rho_{\mathrm{o}} + 0.7\rho_{\mathrm{s}} \geq 0.15$ runs first. A term recurring at segment indices $S = \{i_1 < \cdots < i_k\}$ is admitted as a span anchor only if its extent clears an *adaptive* gap,

$$i_k - i_1 \;\geq\; g_{\min}(N) = \max\big(3,\; \lfloor N/6 \rfloor\big), \tag{13}$$

which is the sense in which the gap is adaptive: on a five-minute M^{3}AV lecture it is three segments, on a forty-eight-minute SlideSpeech recording it is tens of segments, so "non-adjacent" scales with recording length instead of being a constant. Lexical diversity across the extremes and the span score are

$$\delta = 1 - \min\big(J(u_{i_1}, u_{i_k}),\, 0.8\big), \qquad \mathrm{score}(S) = \frac{i_k - i_1}{N} \cdot \bar{q} \cdot \delta, \tag{14}$$

with $\bar{q}$ the mean of equation 12 over the span and J taken on the first 400 characters of each speech segment. The cap at 0.8 keeps δ bounded away from zero so that a recurring term in near-identical contexts is down-weighted but not excluded. The pair $(\delta, |S|)$ types the relation: $\delta > 0.7$ with $|S| = 2$ is a contrast, $\delta > 0.5$ with $|S| \geq 3$ an evolution, $\delta \leq 0.5$ with $|S| \geq 3$ a synthesis, and anything else a generic chain. A span builder assembles $k \in [2, 5]$ non-adjacent segments and the attack set is $\mathcal{A} = \{\{u_{i_j}\}\}_{j=1}^{k}$, every member issued alone. Gate at $\theta_Q = 8$.

P4, multi-hop reasoning. Each segment is first compressed to a two-to-three-sentence summary m_i from (u_i, x_i); chain discovery reads only $\{m_i\}$, which is what makes the search tractable when N is large, and when $N > 200$ the summaries are subsampled with stride $\lceil N/200 \rceil$ so the discovery prompt stays bounded. The discovery agent returns an ordered chain $h_1 \to \ldots \to h_n$ with $n \in \{2, 3\}$, sampled at temperature 0.3 rather than 0, since chain finding is a search problem and a deterministic decode collapses to the same chain on every retry. A validator then checks that each h_k is supported in the segment it is assigned to, and a builder materialises its content. Two falsifiers follow:

$$\mathcal{A}(q) = \underbrace{\{h_{1..k}\}_{k=1}^{n-1}}_{\text{prefixes: no early stop}} \;\cup\; \underbrace{\{\{h_k\}\}_{k=1}^{n}}_{\text{lone hops: no shortcut}}, \qquad |\mathcal{A}(q)| = 2n - 1. \tag{15}$$

The prefix falsifier requires a stated confidence of at least 7 before a blind answer counts as a success, while the isolation tester uses the stricter joint condition $\mathcal{M}_{\mathrm{hyb}}(\hat{a}, a) = 1$ *and* $J_w(\hat{a}, a) > 0.5$, with J_w the Jaccard similarity on whitespace tokens. Both directions matter: including $\{h_n\}$ in equation 15 is what rules out a question whose answer is readable from the final hop alone, which is the failure mode that distinguishes a genuine chain from a relabelled single-hop question. Gate at $\theta_Q = 8$.

A.3 A WORKED EXAMPLE: ONE CROSS-MODAL SYNTHESIS ITEM

Figure 3 shows the slide of a cross-modal synthesis item exactly as it is stored in MMAv-InstructBench, together with the speech and the paper sentences the generator aligned to it. The item comes from the M^{3}AV recording CHI-2D617, segment 5, spanning 1m29s to 2m02s; it was accepted under synthesis pattern D, the pattern that requires all three sources, with a quality-gate score of 9.

The item. *Question.* Which platform displayed the generated notes on the EMR screen labeled C during the think-aloud review? *Gold answer.* OscarEMR.

What each channel contributes. The **slide** of Figure 3 lists "Think-aloud review & rating of generated notes" and labels the photograph "A: One-way mirror", "B: Participant", "C: EMR", "D: Microphone", "E: Standardized Patient". It therefore establishes what the screen labelled C is, and it never names the software running on it. The **speech** of the segment repeats the setup verbally, "Label A is a one-way mirror which researchers sat behind to operate the digital scribe, B is the physician participant, C is the EMR, D is the microphone", and describes the review, "We then presented the generated notes to our physician participants for a think-aloud review session". It too stops short of the platform name. The **paper** supplies exactly that missing fact: "The generated



Figure 3: The slide shown during the cited segment of CHI-2D617. The legend fixes the referent the question uses, "C: EMR", and the second bullet establishes the think-aloud review, but nothing on the slide names the software.

notes were presented to participants in OscarEMR and we asked each participant to directly edit the generated notes to the standards of their own while they followed a think-aloud procedure to voice their thoughts and reasoning for the changes they made." The string *OscarEMR* occurs in the paper and nowhere in the slide or the transcript, so the answer is uniquely determined and exactly checkable.

The recorded gate verdicts. Each channel was issued alone to a blind solver, and the item was retained because none produced the gold answer.

Blind view E	$\sigma_E(q)$	Why it is insufficient
Paper sentences	CANNOT_ANSWER	Names OscarEMR as the review platform but has no "screen labeled C" to attach it to
Slide image	CANNOT_ANSWER	Identifies C as the EMR; the platform is not written anywhere on the slide
Speech segment	CANNOT_ANSWER	Repeats the labels and the review procedure; never names the software

Table 6: The attack set $\mathcal{A}(q) = \{A_i, \{x_i\}, \{u_i\}\}$ for the item of Figure 3 and the stored verdict of each blind solver. The item is accepted only because all three fail, per equation 1.

Every accepted item stores this much: the slide image and its OCR, the speech span with its time range, the aligned paper sentences, a statement of what each modality contributes and why each one alone is insufficient, the per-modality blind verdicts, and the quality-gate score. That is what makes an acceptance re-auditable rather than asserted, and it is the record the human verifiers of Appendix A.5 worked from.

A.4 Modality exclusivity of the source recordings

The premise that a single channel misses part of a recording can be checked on the raw material, independently of any question. For each recording we take the set of content terms (alphabetic tokens of four or more characters after stop-word removal) on the slides, in the transcript and, for M^{3}AV, in the source document, and measure what share of each channel's terms never occurs in the others (Table 7). On M^{3}AV, 53% of slide terms are never spoken and 67% of spoken terms never appear on a slide; 72% of the paper's terms appear in neither, and 45% of slide terms appear in neither speech nor paper. On SlideSpeech, whose slides are denser, 80% of slide terms are never spoken. A reader of any one channel therefore misses roughly half of what the others contain, which is the condition under which questions that require combination can exist.

Share of a channel's terms absent from the other channel(s) (%)	M^3AV	SlideSpeech
slide terms never spoken (slide\speech)	53.1	79.5
spoken terms never on a slide (speech\slide)	67.3	72.0
paper terms in neither slides nor speech	71.7	–
slide terms in neither speech nor paper	44.7	–
spoken terms in neither slides nor paper	44.5	–
recordings	402	884

Table 7: Modality exclusivity of the raw recordings (mean over recordings; medians are within one point). Computed before any question is generated.

A.5 Human verification

We human-verified a stratified 20% sample of the benchmark: 923 question–answer pairs (323 M^3AV, 600 SlideSpeech), sampled uniformly within each of the four reasoning dimensions. Three independent verifiers, following Appendix A.6 and given identical multimodal evidence, rated each item for answerability, answer correctness, and acceptance. Fleiss κ was 0.888, 0.921, and 0.890, respectively, with Krippendorff α identical to three decimals; pairwise Cohen κ ranged from 0.860–0.929, and three-way agreement was 96.0%, 95.6%, and 93.7% (Table 8).

That sample was interleaved with 301 unmarked control items, twins of genuine items in which either the answer (153 items) or the cited evidence (148 items) had been swapped, so that the judged set contains both acceptable and defective items and chance-corrected agreement is informative. Each of the three human verifiers judged every one of the $1{,}224$ items independently, without seeing the others' judgements and without knowing which items were controls, so the agreement in Table 8 is computed over all $1{,}224$ items and Table 9 reports how reliably the controls were caught. A majority vote rejects 90.0% of the defective twins while accepting 96.6% of genuine items; evidence swaps are detected almost without exception (98–99%) and answer swaps, which require checking a short answer against the source, at 80–86%. On the genuine sample the majority accepts 95.4–100% of items in every dimension except easy temporal grounding on M^3AV (92.1%); the 31 items (3.4%) rejected by at least two verifiers are released with the corpus for inspection.

Criterion	Fleiss κ	Krippendorff α	Cohen κ (range)	Agreement (%)
Answerable from evidence	0.888	0.888	0.860–0.907	96.0
Gold answer correct	0.921	0.921	0.909–0.929	95.6
Accept as-is	0.890	0.890	0.874–0.910	93.7

Table 8: Inter-verifier agreement on the $1{,}224$ judged items (923 sampled items and 301 control twins). Values above 0.81 are "almost perfect" on the scale of Landis & Koch (1977).

Verifier	Defective rejected	Answer swap	Evidence swap	Genuine accepted	Balanced acc.
A	90.7	82.4	99.3	95.3	93.0
B	92.0	85.6	98.6	96.4	94.2
C	88.7	79.7	98.0	94.6	91.6
Majority vote	90.0	–	–	96.6	93.3

Table 9: Detection of the control twins (%). "Defective rejected" is the share of control items marked not acceptable; "genuine accepted" the share of sampled items accepted as-is.

A.6 Guidelines given to the human verifiers

The three verifiers received the same written guidelines before starting and worked through the $1{,}224$ items independently, with no contact with one another and no indication of which items were control twins. All three are researchers working in this domain who took part on a voluntary basis, without monetary compensation, because they valued the benchmark the study would support. For every item a verifier saw the material of the recording that the item cites: the slide images with their OCR text, the speech segments in time order with their time stamps and, on M^3AV, the source document; together with the question, the gold answer and the stored evidence pointers (the time span, the segment indices, or the ordered hop list). The guidelines were as follows.

Order of inspection. Study the evidence before reading the question. Go through the modalities one at a time. Read every slide image in full and compare it with its OCR text; the OCR may be incomplete or wrong, and the image is authoritative. Read the speech segments in time order and note what is said at each moment. Where a document exists, read the cited passages and relate them to the slides and the speech. Only once the content of every modality is understood, read the question and the gold answer, and locate the exact place in the evidence where the answer is established.

Criterion 1: answerable from the evidence. Mark yes only if the shown evidence alone determines the answer, with no outside knowledge, no guessing and no inference beyond what the material states. Every fact the answer depends on must be present. For temporal grounding, the answer must be visible on the slide at the moment the question describes; for cross-modal synthesis, the combination of the cited sources must yield it; for long-context reasoning, the cited segments taken together must yield it; for multi-hop reasoning, every hop must be supported and the chain must lead to the answer in the stated order.

Criterion 2: gold answer correct. Mark yes only if the stored answer is the correct answer to the question as posed. Numeric answers must match the source exactly, including units and precision; names, symbols and technical terms must match the source; a correct answer written in a different but equivalent form counts as correct; an answer that is partially correct, ambiguous between two readings, or correct for a different question than the one asked counts as incorrect. This criterion is judged independently of Criterion 1: an item can be answerable while its stored answer is wrong.

Criterion 3: accept as-is. Mark yes only if the item could be released without any change: the question is unambiguous and has one answer, it cannot be answered from its own wording, it does not presuppose a fact the evidence contradicts, the evidence pointers point to where the answer is established, and Criteria 1 and 2 both hold. Any defect, however small, is a rejection.

Dimension-specific checks. For the easy temporal tier, confirm that the time window in the question contains the slide that carries the answer; for the hard tier, confirm that the description in the question picks out that slide and no other. For cross-modal synthesis, confirm that the cited paper sentences, slide and speech segment refer to the same object and that the answer follows from combining them rather than from any one of them. For long-context reasoning, confirm that the cited segments are non-adjacent and that each contributes something the others do not. For multi-hop reasoning, confirm that each hop is established where the hop list says it is and that the final hop gives the answer.

General rules. Do not use outside sources, search engines or other tools, and do not consult the other verifiers. Judge the three criteria separately and record all three for every item. When in doubt, reject. Record a short note giving the reason for any rejection.

A.7 Do the modality labels predict system behaviour?

Every accepted item records which modalities its blind solvers found necessary. If those labels describe real dependencies, removing a modality from a system should hurt on the items labelled as needing it and spare the rest. Table 10 tests this with a difference-in-differences on the same M^{3}AV pool (gemma answerer). Removing the source document from MM-Graph-RAG costs 16.4 points on the 500 items the gate marked as document-dependent and 0.1 on the $1{,}117$ others (interaction -16.5). A retrieval-free probe gives the same answer for the other two channels: supplying the transcript to a blind answerer raises accuracy by 15.1 points where the gate says speech is required and by 6.9 where it does not (interaction $+8.2$), and supplying the slides raises the slide-dependent items by 28.7 points. The corpus labels and the system's behaviour were established by independent procedures, so their agreement is an external check on the gates rather than a restatement of them.

B MM-Graph-RAG implementation constants

Offline cost is one embedding per unit plus an $|S| \times |D|$ similarity matrix; online cost is a few hundred dot products and at most three model calls (four on a refused value read), with no separate vision encoder for retrieval and no LLM-built graph. Every threshold and budget above is computed per recording or per question; the remaining constants are clip bands, floors and structural bounds, identical across all collections and answerers in this paper. All thresholds and budgets

Intervention	Gate label	n	before	after	Δ	95% CI
Modality removed from MM-Graph-RAG						
− source document	required	500	67.6	51.2	-16.4	$[-20.4, -12.4]$
− source document	not required	1,117	65.7	65.8	$+0.1$	$[-1.3, +1.4]$
− slides / frames	required	1,616	66.3	43.9	-22.4	$[-25.1, -19.7]$
Modality supplied to a retrieval-free answerer						
+ transcript	required	829	17.9	32.9	$+15.1$	$[+11.2, +19.1]$
+ transcript	not required	788	7.2	14.1	$+6.9$	$[+4.2, +9.5]$
+ slides	required	300	12.0	40.7	$+28.7$	$[+22.3, +35.0]$

Table 10: Gate validation on M^{3}AV (judge accuracy, %). The drop or gain from a modality lands on the items whose gate label says that modality is required (interactions -16.5 for the document and $+8.2$ for the transcript). The transcript is the anchor stream of the graph, so removing it inside MM-Graph-RAG degrades segmentation for every question; the retrieval-free probe isolates its evidential role.

of MM-Graph-RAG are computed per recording or per question (Section 3.2): the segmentation weights and cut, the cross-modal and coreference thresholds, the walk order, the LIRR blend, and the seed, evidence and section budgets. Table 11 lists the constants that bound those quantities or fix the graph's shape. They are identical across every collection and answerer in this paper; the *static constants* row of Table 20 shows that replacing all adaptive quantities by fixed values changes accuracy by under one point.

Constant	**Value**	**Role**
Graph construction		
Anchor admissibility	timed ft ≥ 0.5; co-anchor cov ≥ 0.6, ft ≥ 0.8	which streams may anchor or co-anchor the timeline
Segmentation weights	band $[0.05, 0.60]$; $\sum_x w_x = 2.5$	floor and ceiling on the derived signal weights
Cut threshold	band $[P_{40}, P_{95}]$; P_{75} fallback	bounds the knee on a degenerate distribution
Segment duration	≥ 8 s; $\bar{n} = \mathrm{clip}(T/30\,\mathrm{s}, 4, 60)$	duration-derived merge/split limits
Slide attachment	overlap ≥ 0.40	a slide joins each segment it covers
Section pass	cut at $\mu - 0.5\sigma$; sizes 2–8 around $\sqrt{n}$	coarse topic grouping
Cross-modal edges	band $[0.20, 0.55]$; 0.34 fallback; top-2	bounds Otsu's cut; keeps the graph sparse
Coreference edges	df $\in [2, \mathrm{clip}(P_{85}, 3, 30)]$; degree ≤ 4	term must recur but not be ubiquitous
Question answering		
Leaf preview	8 segments; RRF $k = 60$	evidence shown to the type classifier
Keyword acceptance	conf ≥ 0.90	skips the classifier call on clear cues
Seeds / evidence	$2\lceil \log_2 n \rceil$; $\lceil 3.5\sqrt{n} + 2 \rceil$	sub-linear in recording size
Section expansion	$\lfloor \lvert C \rvert / 10 \rfloor$ (temporal); $\lceil 0.3 \lvert C \rvert \rceil$ (other); $\max(\lvert C \rvert / 2, \sqrt{n})$ (LIRR)	how many sections seed retrieval
Timeline spread	$\lceil 0.25n(1-\rho) \rceil$, in $[2, n/2]$	fills uncovered parts of the timeline
Walk smoothing	$+0.1$, $+0.1$, $+0.01$	keeps every edge type's priority non-zero
Document share	$\leq \lvert \mathrm{evidence} \rvert(n)/3$	a paper cannot crowd out the recording
LIRR	$\alpha = 0.35 + 0.20\rho' + 0.10\delta$; 2 members per section; $+12$ nodes	query shift and coverage budget
Controller	merge cap 24 nodes	one bounded second round
Visual frames	3 (5 on refusal); 768 px	on-screen-value reads
Answer format	1–8 words; temperature 0	extractive, deterministic

Table 11: Implementation constants of MM-Graph-RAG. Each bounds a data-derived quantity or fixes a structural choice; all are shared across every experiment.

B.1 Design rationale and motivation

Section 3.2 specifies MM-Graph-RAG. This subsection gives the reasoning behind each component and the measurement in this paper that bears it out.

Amortised construction. A recording is queried many times, so every quantity that does not depend on the question is computed once. Segmentation, nodes, edges and indices are built offline from embeddings and closed-form statistics with no model call, and the online phase spends a few hundred dot products and at most three model calls, four when a refused value read is re-asked. That count is independent of recording length, so a forty-eight-minute SlideSpeech recording costs the same per question as a five-minute M^{3}AV one, which is why the method sits on the efficient frontier of latency, prompt tokens per question and context per call simultaneously (Appendix G).

Segmentation from both timed streams. Cutting on the transcript alone is blind to roughly half the content: 53% of M^{3}AV slide terms are never spoken, 67% of spoken terms never appear on a slide, and on SlideSpeech the first figure rises to 79% (Table 7). Fusing the co-anchored slide stream onto the anchor before cutting inserts a boundary when either channel changes, so a retrieved unit is coherent in both, and placing boundaries at information change rather than at a fixed length aligns that unit with the unit of exposition the corpus cites as evidence. With the embedding model held fixed this recovers 76.7% of gold evidence at three nodes against 63.1% for BM25 and 59.1% for dense retrieval (Figure 2), which attributes the gap to the units and the structure rather than to the representation.

What the four levels are for. The four node levels exist because the evidence a question needs comes in four different sizes, and the evidence budget is bounded. A single-slide lookup needs one image. A long-context question needs two to five segments scattered over the recording. A cross-modal question needs a spoken moment and a paper passage together. With one granularity these cannot all be served: make every node a fine segment and a slide is reachable only through the segment that mentions it, arriving as OCR rather than as an image; make every node a section and the one decisive line arrives buried in a topical block. Four levels let the same budget buy the unit the question actually needs, and the walk's level-diversity term rewards expanding into a level the seeds do not yet contain, so the levels are used rather than merely present.

Level	One node is	The need it answers	Serves
Fine segment S	a stretch of speech with the slide shown during it	the default answer unit, matching the granularity at which evidence is cited	all
Aligned item A	one slide, kept whole	a slide must be retrievable on its own and readable as an image, not only through its OCR	p1e, p1h
Section C	a run of consecutive segments on one topic	a question can concern a passage of the lecture rather than an instant	p3
Document passage D	a chunk of the source paper	lecture and paper content must meet somewhere	p2

Table 12: Each node level, what a single node of it contains, and the retrieval need it answers. Aligned items exist wherever a recording carries slides; document passages exist only for M^{3}AV.

The aligned level is the clearest case. A slide kept as its own node can be ranked, retrieved and read directly, which is why single-slide lookup reaches 99.7% recall at three nodes against 54.3% for BM25 and 50.7% for dense retrieval over identical embeddings (Table 25): a text-similarity retriever has no reason to rank a slide highly, since its OCR seldom resembles the question, whereas the walk arrives through the speech-to-slide edge. The section level is the complementary case. It is safe to retrieve a coarse unit only because the answerer tolerates surrounding material, which the dilution result establishes: padding gold evidence with four minutes of unrelated content from the same recording leaves accuracy unchanged, 55.2 to 56.5 with $p = 0.15$ (Figure 6). Coverage, not compactness, is the binding constraint, so a level that trades precision for reach is worth having.

What each edge type is for. The introduction argues that answering over a recording needs four distinct abilities: locating a moment, combining modalities, linking distant material, and following a chain. The edge set exists so that one graph can serve all four, because each type encodes a different

way evidence is distributed in a lecture, and the walk of §3.2.2 can prefer whichever the question needs.

Type	Joins	Property it gives retrieval	Serves
TEMPORAL	consecutive segments	order and adjacency in time	p1e/p1h, p4
ALIGNED	a segment and its slides	binds what was said to what was shown at that instant	p1e/p1h
SECTION	a section and its members	a topical unit larger than a moment	p3
CROSS_MODAL	a segment and a passage	the only lecture-to-paper bridge	p2
COREF	segments sharing a term	reaches distant mentions of one concept	p3, p4

Table 13: Each edge type, the relation it encodes and the reasoning dimension it principally serves.

Two of these carry the properties that flat retrieval cannot express. ALIGNED is a relation between a spoken instant and a displayed image, which no text-similarity ranking can recover, since a slide's OCR rarely resembles the question; this is why single-slide lookup reaches 99.7% recall at three nodes against 54.3% for BM25 and 50.7% for dense retrieval over identical embeddings (Table 25). COREF is a relation between segments that a similarity ranking scores independently and therefore never connects, which is why replacing the graph with flat dense retrieval costs most on the dispersed dimensions, 3.2 points on p3+p4 under gemma and 5.8 under LLaVA-OneVision, against 2.0 overall (Table 20). The per-step attribution of Figure 4 shows the same split from the other side: graph formation is the only step with a positive column on p3 and p4 under both answerers.

How this graph differs from earlier graph retrieval. Graph RAG and its granularity-aware variants (Edge et al., 2024; Wu et al., 2025) organise a text corpus into a knowledge graph whose nodes are entities or documents and whose edges are semantic relations extracted by a language model; multimodal variants carry that construction into visual question answering (Yuan et al., 2026), and video systems use graph indexing or entity-relation structure to steer retrieval and frame selection (Ren et al., 2026; Chu et al., 2025). Three properties separate the graph used here. Its nodes are time-aligned multimodal units at four granularities rather than entities, so an edge can assert *displayed while this was said*, a relation an entity graph has no place for. Its construction calls no language model: every node, edge and threshold comes from embeddings and closed-form statistics, which is what keeps graph cost proportional to the corpus rather than to extraction calls. And its two similarity thresholds are fitted to each recording instead of being set once for a dataset, which is the practice the introduction identifies as the reason existing pipelines transfer poorly across collections of different length and density. The consequence is the transfer result: the same code, constants and prompts run unchanged on MMAv-InstructBench, TVQA and Video-MME.

Per-recording thresholds and sparsity. Similarity distributions are not comparable across recordings, so one global cut over-connects a lecture with a closely matched paper and disconnects one whose paper is broader. Otsu's rule (Otsu, 1979) selects the cut that best separates each recording's own two populations, and the coreference band $[2, P_{85}]$ asks for the same relative property, recurrence without ubiquity, at any length. Bounded out-degree keeps the graph linear: with at most two document neighbours per segment and coreference degree at most four (Table 11), the handshake identity gives $|E_{\text{CROSS_MODAL}}| \leq 2|S|$ and $|E_{\text{COREF}}| \leq 2|S|$, while the structural types contribute one edge per consecutive pair, per section member and per slide overlap. The edge count is $O(n)$ rather than $O(n^2)$, so the walk cannot exhaust its budget inside one dense neighbourhood.

Sub-linear budgets and coverage. The evidence a question needs does not grow with the recording: long-context questions span a median of three segments on M^{3}AV and five on SlideSpeech, for recordings differing in length by an order of magnitude. A budget proportional to n would add distractors at the rate the recording grows, so $|\text{evidence}|(n) = \lceil 3.5\sqrt{n} + 2 \rceil$ grows enough to keep distant evidence reachable without diluting it. Two measurements locate the operating point: halving or doubling the budget moves accuracy by at most 0.7 points (Table 20), and padding gold evidence with four minutes of unrelated material leaves accuracy unchanged, 55.2 to 56.5 with $p = 0.15$ (Figure 6). Coverage of the right spans, rather than compactness of the context, is therefore the binding constraint. LIRR widens coverage in the embedding space already computed, at no model call, and the controller spends its one optional call where coverage is lowest, requesting a second

round on 73% of long-context questions, the dimension with 29.3% first-pass coverage and the only one where a second round raises accuracy, 47.1 to 55.2 (Tables 21 and 28).

Typed expansion. Relations are not interchangeable: on-screen questions are served by speech-to-slide edges, dispersed questions by coreference and section edges, and combination questions by cross-modal edges. Ordering the types per question by affinity, novelty and coverage lets one graph behave differently for different questions with no per-dataset tuning. Figure 4 shows the resulting division of labour, with visual verification carrying the slide-reading dimensions, adaptive segmentation the chained dimension, and graph formation the dispersed ones.

Visual verification. OCR omits diagram labels, axis values and callouts, and the corpus is constructed so that this matters, since a temporal-grounding item is retained only when no text-only view yields the answer. The channel is correspondingly load-bearing: removing it costs 22.4 points, an order of magnitude more than any retrieval ablation, and a retrieval-free probe handed the slides gains 28.7 points on exactly the items whose gate label says slides are required (Tables 20 and 10). Answering value questions from the frames alone, with one re-ask at higher frame count on a refusal, is what converts that signal into accuracy.

Attribution. Four comparisons isolate the structure from every other factor. Against flat retrieval over identical embeddings the walk recovers more gold evidence at every budget. Against a retriever with longer units it recovers more per second of video placed in context, 79.8 against 77.5 at 45 s and 90.6 against 89.2 at 120 s (Table 26). Replacing the walk with uniform sampling inside the pipeline costs 29.4 points (Table 22). And one neutral answerer given our evidence rather than the strongest baseline's scores 42.2% against 33.5% under an identical prompt and character budget (Appendix E.5). Each holds fixed what the others vary, so the result is attributable to the retrieval structure rather than to the embedder, the prompt, the answerer or the budget. Since one model serves as classifier, controller and answerer, the embedder is frozen, and replacing every data-derived quantity with a fixed constant moves accuracy by under one point, that structure carries to a new collection without retuning.

B.2 Graph and pipeline vocabulary

This subsection expands the terms of §3.2.

Unit. The smallest piece a collection releases for a stream: one transcript line for speech, one slide for the visual stream, one paragraph for a document. Units of the timed streams carry a start and an end time.

Node levels. *Fine segments* (S) are what retrieval returns: one contiguous stretch of speech together with the slide on screen while it is spoken, cut where the content changes. *Aligned items* (A) are the slides themselves, kept as their own nodes so a slide can be retrieved alone and read as an image rather than through its OCR. *Sections* (C) group consecutive segments that stay on one topic, giving retrieval a coarse handle when a question concerns a whole passage of a lecture rather than a moment in it. *Document passages* (D) are chunks of the source paper and exist only for M^{3}AV.

The importance prior. $\mathrm{imp}(v)$ multiplies the dense score at seeding, so it reorders candidates without removing any. It starts at 1 and subtracts 0.12 when the segment matches a fixed meta phrase list (greetings, self-introductions, outline, agenda, table of contents, acknowledgements, thanks), 0.05 when the segment falls in the first twentieth of the recording, and 0.05 when its text is shorter than forty characters, with the result floored at 0.8. The floor is what keeps the prior safe: an introduction or a short segment that happens to carry the answer is nudged down but can still be seeded.

Edge types. Three are read directly off the structure. TEMPORAL joins consecutive segments, so retrieval can step forward or back in time. ALIGNED joins a segment to the slides it overlaps, which is how a question about a spoken moment reaches the slide that displays the answer. SECTION joins a section to its members. Two are similarity-based and thresholded per recording. CROSS_MODAL joins a segment to a document passage about the same thing, which is what allows paper and lecture to be combined. COREF joins segments that share a salient term such as a symbol or a dataset name, which is what allows far-apart mentions of one concept to be gathered. The takeaway is a

division of labour: the structural types carry movement along the recording, the similarity types carry movement across modalities and across distant moments.

The online model calls. Three calls run per question. The *classifier* assigns the question one of five types, which sets how retrieval widens. The *controller* reads the rendered evidence and returns whether it suffices, naming a missing fact when it does not. The *answerer* writes the answer. A fourth call occurs only on a refusal: when a question asks for a value printed on a slide, the answerer is shown three frames from the relevant moment, and when the answerer itself reports that the value is not legible the question is re-asked once with five frames (Table 11). The re-ask exists because a small number can be unreadable at three frames and readable at five, which is a retrieval failure rather than a wrong answer.

LIRR. The query is shifted toward the centroid $\bar{c}$ of the retrieved evidence, $q' = \mathrm{norm}(\alpha q + (1 - \alpha)\bar{c})$, where α grows with the timeline fraction ρ' that the evidence already spans and with its level diversity δ (Table 11). Evidence clustered in one part of a recording therefore receives a larger shift and explores further, while evidence already spread over the timeline and over node levels keeps q' near the question.

When LVV fires. A question counts as an on-screen value read when its temporal flag $b(Q)$ is set, when it names a time window $\omega(Q)$, or when it carries a visual-locative cue such as "arrow". Frames are then taken across $\omega(Q)$ when one is given and otherwise across the retrieved segment closest to q.

C WHY THESE FUNCTIONAL FORMS

Section 3.2 states what each quantity is; this section states why it has the form it has. Three constraints apply throughout. Every threshold must be a statistic of one unlabelled recording, since there is no tuning split. It must be unit-free, because the signals combined have incomparable ranges ($1-\cos \in [0, 2]$, token novelty in $[0, 1]$, a binary cue, unbounded BM25 scores), which is why quantities are combined after z-scoring, or through ranks, or through percentiles of the recording's own distribution: a constant such as 0.34 on a cosine is calibrated only for a recording of average topical breadth. And it must fail boundedly over a two-order-of-magnitude range in length (14 aligned segments on M^{3}AV, 451 on SlideSpeech), which is what the clip bands of Table 11 provide. Table 14 collects the local choices; the four with a non-obvious justification follow.

Segmentation weights. From $\tilde{w}_x = \mathrm{KL}(p_x \| \mathrm{Unif}) + |\Pr_i[x_i > \bar{x}] - \frac{1}{2}|$ over m candidate boundaries,

$$\mathrm{KL}(p_x \| \mathrm{Unif}) = \sum_i p_{x,i} \log \frac{p_{x,i}}{1/m} = \log m - H(p_x), \tag{16}$$

so the first term is an entropy deficit: it vanishes iff the signal is flat across candidates, and a flat signal carries no information about where to cut. It is maximized by a single spike, which is why it is not used alone; the second term is minimized when the signal splits candidates into two comparable populations, so it moves weight from spiked signals to bimodal ones. The terms are summed, not multiplied, so a signal scoring zero on one is down-weighted rather than deleted. Renormalizing to $\sum_x w_x = 2.5$, the mass of the fixed weights this replaces, keeps s_i on its original scale, so the cut band retains its meaning.

Cross-modal threshold. For pooled $\cos(s, d)$ the total variance decomposes at any cut θ as

$$\sigma^2_{\mathrm{total}} = \underbrace{w_0\sigma_0^2(\theta) + w_1\sigma_1^2(\theta)}_{\text{within class}} + \underbrace{w_0 w_1(\mu_0 - \mu_1)^2}_{\text{between class}}, \tag{17}$$

with $\sigma^2_{\mathrm{total}}$ independent of θ, so maximizing the between-class term (Otsu, 1979) is identical to minimizing pooled within-class variance: θ^* is the cut making both populations as homogeneous as the data allow, under the assumption that pairs are a mixture of related and unrelated ones, and with no prior on where the cut sits. A unimodal histogram has no useful maximizer, hence the clip band and fallback; the top-2 restriction keeps the edge count linear in $|S|$ rather than $|S|\,|D|$.

Sub-linear budgets. For $k(n) = n^\beta$, $\beta = 0$ covers a vanishing fraction of a long recording, so coverage of the decisive segment falls to zero; $\beta = 1$ keeps coverage but leaves that segment a fixed small share of an unboundedly growing block. Only $0 < \beta < 1$ escapes both, and $\beta = \frac{1}{2}$ balances

the two regimes symmetrically in $\log n$. The exponent is also structural: the section pass yields $\Theta(\sqrt{n})$ sections, so $|\text{evidence}|(n) = \lceil 3.5\sqrt{n} + 2 \rceil$ is the smallest budget that can place a node in every coarse topic. Seeds are entry points the walk expands, so $|\text{seeds}|(n) = 2\lceil \log_2 n \rceil$ suffices and gives $|\text{seeds}|/|\text{evidence}| \to 0$, keeping the block structural rather than a flat ranking. Halving or doubling $|\text{evidence}|(n)$ moves accuracy by at most 0.7 points (Table 20), so the form need only land in the right regime.

Walk priority. In $\pi(\epsilon) = \text{aff}_\epsilon \, \text{nov}_\epsilon \, \frac{|N_\epsilon|}{|V|}(1 + \nu_\epsilon)(1 + b(Q)\phi_\epsilon)$ each factor is a necessary condition, so the product gives a near-zero factor veto power that a weighted sum would outvote; $\text{aff} \cdot \text{nov}$ is the relevance–redundancy trade-off written so neither can be traded away. The availability factor is what lets one code path serve a collection with no document stream: CROSS_MODAL reaches nothing, its priority falls to the smoothing floor, and the order adapts with no special case. Additive smoothing keeps every priority positive so an empty neighbourhood cannot make the ordering degenerate.

Form	Property it guarantees	Alternative that fails
Generation		
Accept a conjunction of five tests (Eq. 2)	monotone acceptance: a test added later cannot enlarge the released set, and each rejection has one named cause	a weighted score lets a fatal leak be offset by a high quality rating
NoLeak a conjunction over all of $\mathcal{A}(q)$ (Eq. 1)	the exclusion holds per item, so one released question can be audited alone	a majority rule releases items one impoverished view still answers
$\mathcal{M}_{\text{hyb}}$ rules before Judge	conservative: a firing rule makes an attack succeed, so it can only reject, never admit an easier item	judge-first spends calls without that guarantee
$\theta_Q = 7$ (P1), 8 (P2–P4)	P1 has a positive check (a solver given only s_i must succeed), so the gate carries less load	one bar ignores that P2–P4 have no such certificate
Duplicate cut, content-word Jaccard > 0.45	symmetric, bounded, needs no per-embedding calibration; above chance overlap, below template reuse	embedding cosine needs a cut per model
Gap $\max(3, \lfloor N/6 \rfloor)$	separation is a constant fraction of the recording, with a floor where the fraction degenerates	a fixed gap is unsatisfiable at $N = 14$ and trivial at $N = 451$
Graph		
Anchor coverage continuity $\ln(1+\lvert U_m \rvert)$	veto per factor: neither spanning with holes nor dense on a fraction can anchor; granularity with diminishing returns	a sum lets one factor compensate; a linear count favours a shredded stream
Boundary score $\sum_x w_x z_{x,i}$	unit-free weights, cut comparable across recordings	raw signals make weights range-dependent
Cut at the knee of sorted s_i	invariant to affine rescaling; segment count bounded by $[P_{40}, P_{95}]$	$\mu + k\sigma$ over-cuts long recordings, as scores are heavy-tailed
$\text{df} \in [2, P_{85}]$, degree ≤ 4	recurrence required; a ubiquitous term would induce a $\Theta(\text{df}^2)$ near-clique carrying no grouping information	a fixed count or stop-word list ignores per-speaker reuse
imp multiplicative, floor 0.8	breaks ties without outranking a clearly better match	an additive or unbounded prior overrides the query
Retrieval		
RRF over dense and BM25	rank-only, so invariant to both scales at once	a score sum needs per-recording calibration
Consensus $\frac{1}{2}(\cos(q, v) + \cos(q, s_v^*)) + \min(\cdot, \cdot)$ above the median	the min is a soft AND, large only under cross-modal corroboration; the median keeps it on the recording's scale	the mean alone seeds a passage that merely shares vocabulary
$\alpha = 0.35 + 0.20\rho' + 0.10\delta$	least committed monotone map; $\alpha \in [0.35, 0.65]$, so neither Q nor $\bar{c}$ is ever ignored; clustered evidence explores further	a fixed blend cannot depend on coverage; a new query costs a model call

Table 14: Local design choices in Section 3.2 and §3.1.1, each with the property that motivates its form and the alternative that lacks it.

These forms are chosen for invariance and bounded failure, not because each maximizes accuracy: replacing every data-derived threshold and budget by fixed values moves accuracy by -0.8 points

under gemma and $+0.6$ under LLaVA-OneVision, and every design-robustness row of Table 20 sits within a point of zero. What the forms buy is that one code path runs on a five-minute document-backed lecture and a fifty-minute slide-only one with no threshold reset; the accuracy comes from modality grounding, visual verification and graph structure, in that order.

D COMPLETE RESULTS ON MMAV-INSTRUCTBENCH

Tables 15–17 report M^{3}AV under the three answerers by reasoning dimension, Table 18 reports SlideSpeech, and Table 19 gives the judge-free metrics for every system. Superscript d marks the +docs configuration of a baseline, which receives the source document that exists only for M^{3}AV, and f its faithful configuration, which follows the published method with the inputs it prescribes. M2A is a memory agent populated through tool calls; gemma-3 and Qwen3-VL support tool calling, LLaVA-OneVision does not, so for that answerer the main row is the original configuration and the tool-calling run is listed separately. Bootstrap 95% intervals have half-width 2.0–2.4 points on M^{3}AV and 1.5–1.8 on SlideSpeech for every row.

System	p1e	p1h	p2	p3	p4	overall
MM-Graph-RAG	**61.4**	**78.0**	**67.5**	**54.6**	**81.1**	**66.4**
CLaMR	48.6	68.2	52.4	52.9	68.9	55.7
Video-RAG	43.6	61.6	50.8	53.8	60.4	52.4
VideoAgent^d	40.7	44.3	48.2	41.5	69.8	45.8
LLoVi^d	37.5	40.7	49.4	46.2	65.1	45.3
VideoRAG-Ren^d	30.7	42.3	45.2	43.1	62.3	41.9
M2A	35.7	62.3	34.8	38.3	49.1	41.8
ColPali	29.4	32.1	49.4	41.5	57.5	40.4
M2A (no memory)	37.0	64.9	29.2	26.2	34.0	37.5
Zero-shot	24.1	28.9	46.2	41.2	55.7	37.4
VGent^d	26.8	35.4	40.6	38.5	57.5	37.0
VideoTree^d	27.8	35.7	43.0	30.5	57.5	36.5
RAG-Anything	42.0	58.7	24.8	24.3	37.7	36.0
VideoAgent^f	22.3	20.7	46.6	38.0	67.9	35.7
LLoVi^f	22.0	18.0	47.0	40.2	66.0	35.5
VideoRAG-Ren^f	15.2	19.0	47.0	45.1	60.4	34.7
Agentic	23.1	28.9	42.0	36.6	46.2	34.3
VGent^f	8.9	4.9	34.0	29.1	44.3	22.3
VideoTree^f	3.9	2.3	39.6	26.7	49.1	22.2

Table 15: M^{3}AV, gemma answerer: all nineteen system configurations by reasoning dimension (judge accuracy, %). MM-Graph-RAG leads every dimension. Under gemma the +docs configuration lifts every baseline that receives it by 7 to 15 points, which is why the main table reports it.

System	p1e	p1h	p2	p3	p4	overall
MM-Graph-RAG	**55.6**	**67.9**	**62.8**	**51.7**	**81.1**	**61.1**
CLaMR	42.8	56.1	55.8	47.4	77.4	52.5
Video-RAG	34.9	54.8	50.0	45.8	66.0	47.6
LLoVi^d	28.6	30.8	50.8	42.5	65.1	41.1
VideoAgent^d	32.7	32.8	44.0	39.4	61.3	39.4
VideoRAG-Ren^d	27.0	32.5	44.4	39.1	60.4	38.0
ColPali	23.4	23.0	50.8	37.8	67.9	37.6
VideoTree^d	24.1	33.4	43.4	35.7	62.3	36.7
Zero-shot	20.5	24.9	42.8	34.2	54.7	33.2
M2A	29.9	55.4	25.0	24.0	27.4	31.8
VideoRAG-Ren^f	15.5	13.8	44.6	40.5	55.7	31.8
VGent^d	22.6	26.6	33.8	34.2	53.8	31.2
LLoVi^f	13.4	12.8	46.2	37.1	58.5	31.1
VideoAgent^f	14.7	11.8	43.6	35.3	57.5	30.1
Agentic	20.5	24.9	34.2	33.5	43.4	29.7
VideoTree^f	6.6	8.2	36.6	28.2	50.0	23.4
VGent^f	7.3	6.9	28.4	26.4	33.0	19.3
M2A (tool-calling)	11.3	13.8	9.0	9.5	6.6	10.4
RAG-Anything	5.8	6.6	10.2	11.1	13.2	8.8

Table 16: M^{3}AV, LLaVA-OneVision answerer. The absolute level falls by about five points relative to gemma but the ordering of the top three systems is unchanged.

System	p1e	p1h	p2	p3	p4	overall
ColPali	64.6	**84.6**	61.2	56.7	81.1	66.8
MM-Graph-RAG	**66.6**	77.4	**68.8**	**60.1**	**85.2**	**69.2**
CLaMR	50.9	68.9	60.8	59.2	84.0	61.2
Video-RAG	44.4	64.3	59.6	58.6	79.2	58.0
VideoRAG-Ren^{d}	39.9	57.0	50.8	50.3	73.6	50.8
VideoAgent^{d}	35.7	43.6	54.0	45.1	75.5	47.3
LLoVi^{d}	36.5	42.0	53.6	46.6	73.6	47.3
VideoRAG-Ren^{f}	26.8	37.4	48.2	48.5	64.2	42.2
M2A	38.3	66.9	33.6	32.2	46.2	41.5
VideoTree^{d}	28.1	38.4	49.8	35.3	66.0	40.7
VGent^{d}	28.9	36.7	43.8	38.3	65.1	39.2
Zero-shot	24.1	27.9	51.2	41.4	62.3	39.2
Agentic	23.6	27.9	46.6	39.9	56.6	37.0
VideoAgent^{f}	19.2	19.3	51.2	41.1	66.0	36.6
LLoVi^{f}	19.2	17.7	50.8	42.3	67.0	36.5
VideoTree^{f}	4.5	4.9	42.6	30.4	54.7	24.8
VGent^{f}	5.0	3.9	36.4	35.3	51.9	23.7
RAG-Anything	18.1	20.3	20.2	20.6	19.8	19.8

Table 17: M^{3}AV results with Qwen3-VL answerer.

	gemma				LLaVA-OV				Qwen3-VL			
System	p1e	p3	p4	overall	p1e	p3	p4	overall	p1e	p3	p4	overall
MM-Graph-RAG	**66.1**	**55.6**	**55.6**	**56.0**	**65.2**	**49.5**	**49.6**	**50.1**	**75.9**	**50.2**	47.5	**50.2**
Zero-shot	42.0	43.4	32.2	39.4	28.6	36.9	29.9	34.2	46.4	44.8	36.7	42.0
CLaMR	34.8	38.8	40.1	39.1	33.0	39.3	45.6	41.2	36.6	45.4	48.2	46.0
VideoRAG-Ren	32.1	37.1	35.5	36.4	25.0	34.6	35.9	34.7	30.4	39.4	39.5	39.1
M2A	42.9	37.1	31.6	35.4	34.8	23.7	17.9	22.1	42.9	34.3	33.7	34.4
Agentic	35.7	37.6	25.0	33.2	19.6	24.8	18.9	22.6	44.6	28.7	21.1	26.6
ColPali	13.4	30.4	38.2	32.5	17.9	28.6	29.1	28.4	60.7	47.8	**48.6**	48.6
Video-RAG	16.1	34.1	29.5	31.8	17.0	33.1	30.7	31.7	20.5	42.4	41.8	41.4
LLoVi	30.4	32.8	29.5	31.5	21.4	31.5	29.9	30.6	25.9	35.9	33.2	34.6
VideoAgent	31.2	32.5	29.0	31.2	21.4	27.9	27.4	27.5	20.5	35.0	34.6	34.4
VGent	13.4	21.1	21.7	21.0	7.1	19.5	19.4	19.0	7.1	24.1	24.6	23.6
RAG-Anything	25.9	18.9	20.4	19.7	17.0	14.2	15.5	14.7	32.1	20.3	22.7	21.6
VideoTree	4.5	16.6	16.7	16.2	4.5	21.2	22.5	21.0	2.7	22.1	21.8	21.3

Table 18: SlideSpeech: every system and answerer by reasoning dimension (judge accuracy, %) on the complete pool. The p1e column holds 112 questions and is the noisiest; p3 and p4 carry the weight of the table. The systems that summarise before retrieving (VideoTree, VGent) degrade most on these ten-times-longer recordings.

	M^{3}AV			SlideSpeech		
System (REM / F1)	gemma	LLaVA-OV	Qwen3-VL	gemma	LLaVA-OV	Qwen3-VL
Zero-shot	14.3 / 29.1	16.3 / 27.6	15.3 / 30.8	21.6 / 37.2	22.2 / 32.4	22.9 / 38.5
Agentic	13.0 / 27.1	15.3 / 25.7	13.8 / 28.8	16.9 / 31.5	13.8 / 21.9	12.8 / 25.0
LLoVi	20.5 / 34.9	13.0 / 24.6	22.5 / 36.5	17.9 / 28.8	11.6 / 21.4	20.8 / 31.7
VideoAgent	22.9 / 36.6	16.7 / 28.7	24.7 / 38.5	18.2 / 28.4	14.6 / 23.9	20.3 / 31.1
VideoTree	17.2 / 29.3	14.3 / 24.7	20.1 / 31.9	8.4 / 14.8	7.3 / 15.1	11.0 / 18.5
VGent	15.6 / 28.4	11.0 / 20.1	19.4 / 31.3	10.1 / 17.6	5.8 / 13.1	13.0 / 21.0
Video-RAG	21.6 / 40.0	26.3 / 39.8	27.9 / 45.1	17.1 / 32.7	20.0 / 30.3	23.0 / 39.2
VideoRAG-Ren	18.9 / 32.3	10.1 / 20.6	26.3 / 40.2	20.2 / 33.0	12.6 / 23.3	24.3 / 36.0
RAG-Anything	17.4 / 27.4	0.4 / 4.9	5.4 / 12.7	6.8 / 15.0	1.3 / 7.8	8.4 / 17.1
M2A	15.2 / 26.8	4.0 / 13.6	24.4 / 33.8	12.2 / 23.1	1.4 / 9.7	18.7 / 30.0
ColPali	13.5 / 30.4	18.2 / 30.3	37.2 / 54.6	14.8 / 29.9	17.2 / 26.8	28.8 / 44.2
CLaMR	24.7 / 44.2	30.6 / 44.0	30.1 / 47.9	21.3 / 39.0	26.5 / 39.5	24.8 / 43.2
MM-Graph-RAG	**37.6 / 55.2**	**38.3 / 51.9**	**39.8 / 54.8**	**35.7 / 53.0**	**33.7 / 46.8**	**34.6 / 47.6**

Table 19: Judge-free scoring of every main-table cell: relaxed exact match / token F1 (%). Both metrics use the same normalisation (lower-casing; articles, punctuation and extra whitespace removed); REM requires identity and F1 is the token-overlap harmonic mean. Across all 78 cells, 97.5–100% of predictions that match the reference exactly are also accepted by the judge, so the judge never overrules an exact match. Verbose systems (RAG-Anything, M2A) score low on both metrics by construction.

E Ablations and attribution

E.1 Ablation grid

Table 20 reports the ablations under both answerers. Only the named mechanism is disabled in each row; code, servers, prompts and answerer are otherwise identical to the *full method* row. The three regimes of Section 5 hold under LLaVA-OneVision as well: modality removal dominates (-19.9 for the visual channel, -4.9 for the transcript), visual verification comes next (-5.2 for LVV), and every design-robustness row sits within a point of zero in both directions, which is the portability result: a method whose accuracy is insensitive to its own thresholds can be moved to a new collection without retuning. Per dimension, removing the visual channel costs 43.6 and 61.3 points on the two slide-reading dimensions and nothing on cross-modal synthesis, which remains answerable from paper and speech, and removing the paper costs 16.4 on cross-modal synthesis and nothing elsewhere: the ablation profile reproduces the modality structure the gates certified at generation time (Appendix A.7).

	gemma		LLaVA-OV	
Configuration	overall	Δ [95% CI]	overall	Δ [95% CI]
Full method	66.3	–	60.7	–
Modality grounding				
− video/slides	43.9	−22.4 [−25.0, −19.7]	40.8	−19.9 [−22.6, −17.3]
− transcript	61.4	−4.9 [−7.0, −2.7]	55.8	−4.9 [−6.9, −2.8]
− paper	61.2	−5.0 [−6.7, −3.5]	59.0	−1.7 [−3.3, −0.1]
Text-only answerer (no frames or video)	61.9	−4.3 [−6.0, −2.6]	57.3	−3.4 [−5.1, −1.5]
Retrieval structure				
− latent visual verification (LVV)	62.4	−3.9 [−5.6, −2.4]	55.5	−5.2 [−6.8, −3.6]
Flat dense retrieval (no graph)	64.3	−2.0 [−3.8, −0.3]	59.4	−1.3 [−3.1, +0.5]
Design robustness				
Static constants (no data-derived thresholds or budgets)	65.5	−0.8 [−2.0, +0.4]	61.2	+0.6 [−0.6, +1.7]
Budget ×0.5	65.7	−0.6 [−1.6, +0.4]	61.4	+0.7 [−0.6, +2.1]
Budget ×2	65.8	−0.4 [−1.4, +0.4]	60.0	−0.7 [−2.1, +0.6]
Oracle dimension routing	65.3	−0.9 [−2.2, +0.3]	60.8	+0.1 [−1.0, +1.1]

Table 20: Ablation grid on the full M^{3}AV pool ($n = 1{,}618$), both answerers: judge accuracy and paired difference against the full method with 95% bootstrap intervals. The full-method row is the reference run of the ablation campaign.

E.2 Which method step serves which reasoning dimension

Figure 4 removes one method step at a time and measures the accuracy lost on each dimension, for both answerers. Two readings follow. The top rows are close to a division of labour: visual verification is confined to the slide-reading dimensions (+11.0 and +9.2 for gemma, +13.9 and +11.1 for LLaVA-OneVision, within noise elsewhere), adaptive segmentation peaks on the chained dimension (+7.5), and graph formation is the only step with a positive column on the dispersed dimensions for both answerers (+5.5 on p3 and +6.6 on p4 under LLaVA-OneVision). Under gemma, graph formation is the only step that helps the long-context column, consistent with the retrieval-coverage reading of that dimension in Appendix H.

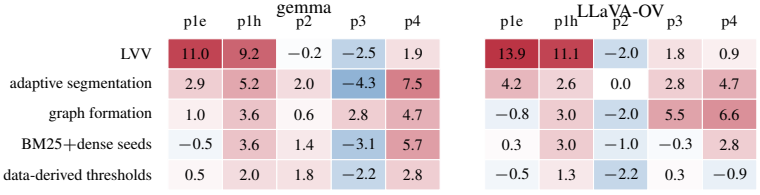

Figure 4: Accuracy points *lost* on each reasoning dimension when a single method step is removed (M^{3}AV, paired on the same questions). Red means the step helps that dimension, blue means the dimension is marginally better without it. Visual verification owns the slide-reading dimensions p1e and p1h; adaptive segmentation owns the chained dimension p4; graph formation carries the dispersed-evidence dimensions, most visibly for LLaVA-OneVision.

E.3 CONTROLLER BEHAVIOUR

Table 21 describes the sufficiency controller. It requests a second retrieval round most often on long-context questions (73% under gemma), and that is the only dimension where the second round raises accuracy ($47.1 \rightarrow 55.2$). Elsewhere the one-round subset scores higher, which is selection rather than damage: the controller asks for more evidence precisely on the questions it finds hard. Under LLaVA-OneVision the controller rarely fires, and the step ranking of Figure 4 is nevertheless preserved, so the division of labour belongs to the retrieval design rather than to the controller's decisions.

	gemma				LLaVA-OV			
Dimension	2nd round	acc. 1 rnd	acc. 2 rnds	nodes	2nd round	acc. 1 rnd	acc. 2 rnds	nodes
p1e	58%	76.2	51.6	21.2	23%	60.0	40.7	19.2
p1h	56%	88.7	70.3	22.2	8%	67.0	87.0	22.6
p2	59%	69.1	66.6	26.0	1%	61.7	85.7	32.5
p3	73%	47.1	55.2	25.6	6%	51.0	61.1	34.2
p4	54%	81.6	80.7	26.9	8%	77.3	88.9	33.1

Table 21: Controller behaviour on M^{3}AV: how often a second retrieval round is requested, judge accuracy on the questions answered with one and with two rounds, and the mean number of evidence nodes placed in the context, which exceeds $|\text{evidence}|(n)$ because the controller's second round and LIRR each add nodes under their own caps (Table 11).

E.4 IS THE GAIN RETRIEVAL? SWAPPING THE RETRIEVER INSIDE THE PIPELINE

Table 22 keeps every other stage of MM-Graph-RAG fixed, the answerer, the answer prompt, the evidence budget, the frames and the judge, and replaces only the retriever. Uniform sampling with no retrieval costs 29.4 points, so the pipeline's accuracy is retrieval-borne, and the graph walk's advantage over dense retrieval over identical embeddings is significant ($+2.0$, $p = 0.020$); the per-dimension and small-budget comparisons of Appendix H show where the walk separates from a flat retriever.

Retriever inside MM-Graph-RAG	Judge acc.	95% CI	vs. graph walk	McNemar p
Uniform sampling (no retrieval)	36.9	[34.6, 39.4]	-29.4	7.6×10^{-90}
Dense only	64.3	[61.8, 66.7]	-2.0	0.020
Graph walk (full method)	**66.3**	[64.0, 68.5]	–	–

Table 22: Retriever ladder on M^{3}AV (1,617 paired questions, gemma answerer). Everything except the retriever is held fixed.

E.5 WHOSE EVIDENCE IS BETTER, WITH THE ANSWER STAGE HELD CONSTANT?

The main comparison lets each system answer with its own prompt, so a margin could in principle be attributed to prompt design rather than to retrieval. To remove that degree of freedom, MM-Graph-RAG and CLaMR each select evidence for every M^{3}AV question; both evidence sets are then handed to one neutral answerer (gemma) with an identical system prompt, an identical template and the same 3,000-character budget, and scored by the same judge. The two evidence sets are the same size to within 2% (2,506 vs. 2,555 characters). Our evidence yields 42.2% against 33.5% on 1,617 paired questions; Figure 5 gives the per-dimension breakdown behind the numbers of Section 5. Absolute values are below the main table by construction: this neutral answer stage sees text evidence only, with no frames, no controller round and no visual verification, so the experiment attributes the gain rather than reproducing the headline result.

E.6 WHAT IRRELEVANT CONTEXT COSTS

Figure 6 holds the gold evidence fixed and pads it with increasing amounts of unrelated material from the same recording, with the same answerer, prompt and judge (1,617 paired questions); the answerer receives the evidence as text, without frames or visual verification, so absolute values sit below the oracle of Table 27. Accuracy is flat: from 55.2 with the gold evidence alone to 56.5 with four additional minutes of unrelated content ($p = 0.15$), and no dimension moves by more than a

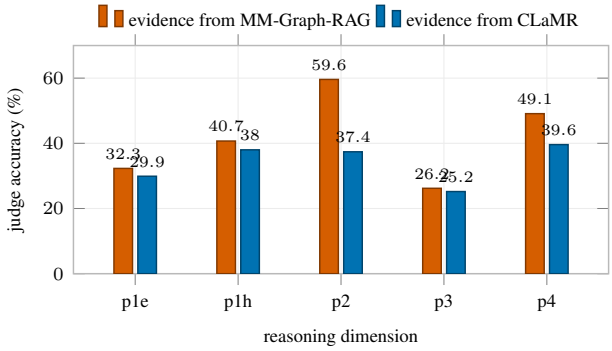

Figure 5: Evidence swap on M^{3}AV: one neutral answerer, one prompt, one character budget; only the retrieved evidence differs. The gain concentrates on cross-modal synthesis (+22.2) and multi-hop reasoning (+9.4).

few points. Two conclusions follow. The answerer is robust to dilution once the right spans are present, so the evidence budget is not the binding constraint, which agrees with the budget rows of Table 20; and what a retriever must deliver is coverage of the gold spans rather than a compact context, which is the quantity on which the graph walk separates from other retrievers (Table 25).

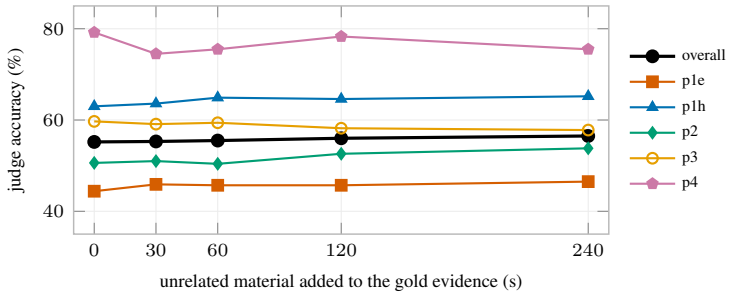

Figure 6: Dilution on M^{3}AV: gold evidence padded with unrelated material from the same recording (gemma answerer). Accuracy is unchanged within noise at every padding level.

F TRANSFER BENCHMARKS IN FULL

Neither benchmark was produced by our generator, so together they test the alternative explanation that the gains in Table 1 reflect an affinity between our question generator and our retriever. Both are multiple-choice and scored by letter with no judge.

TVQA. Table 23 reports every system. TVQA is the strongest test of the generator-affinity explanation: a narrative domain with no slides and no document. MM-Graph-RAG leads under all three answerers (Section 5.2); Figure 7 shows that the advantage over zero-shot is uniform across series while the margin over the retrieval baselines is series-dependent, and that every system finds *The Big Bang Theory* easiest, a property of the source material rather than of any method.

Video-MME. Table 24 reports every system on Video-MME. MM-Graph-RAG is the most accurate system under all the models; the systems that summarise before retrieving (VideoTree, VGent) and the memory agent M2A fall furthest behind.

G INFERENCE COST

Cost. Figure 8 reports inference cost, measured single-threaded on a dedicated server over a fixed 20-lecture subset so that the numbers are comparable across systems, against M^{3}AV accuracy on three axes: latency, prompt tokens per question, and prompt tokens per model call (prompt tokens per question divided by model calls per question), the context the answerer reads in a single call.

System	gemma	LLaVA-OV	Qwen3-VL
MM-Graph-RAG	**55.7**	**60.2**	**60.4**
CLaMR	53.7	57.7	59.0
Video-RAG	51.7	56.4	58.4
VideoRAG-Ren	14.3	56.7	56.6
RAG-Anything	36.9	47.6	47.1
Agentic	40.9	42.2	41.7
Zero-shot	39.8	43.0	44.8
M2A	36.9	37.3	37.8
ColPali	36.7	43.4	40.4
VGent	16.9	34.1	29.2
VideoAgent	11.7	44.8	27.8
LLoVi	4.3	45.1	26.3
VideoTree	3.6	42.9	24.4

Table 23: TVQA, every system and answerer (letter accuracy, %). Bootstrap 95% half-widths are about ± 3.2 points.

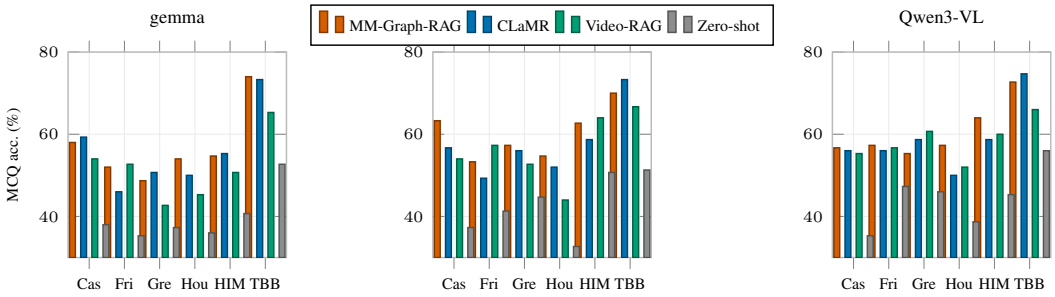

Figure 7: TVQA accuracy by series for the leading systems under each answerer (Cas: Castle, Fri: Friends, Gre: Grey's Anatomy, Hou: House M.D., HIM: How I Met Your Mother, TBB: The Big Bang Theory). The advantage over zero-shot is uniform across series; the margin over CLaMR and Video-RAG is small and series-dependent.

MM-Graph-RAG sits on the efficient frontier of all three axes: every system that is faster or reads fewer tokens is less accurate, the systems that spend more time or more tokens per question score 26 to 30 points lower, and per call it presents a smaller context than the zero-shot answerer, CLaMR, ColPali, VideoRAG-Ren and RAG-Anything while scoring 10.7 to 30 points higher.

H Retrieval quality and headroom

Because every system answers through the same answerer with the same prompt, any accuracy difference must originate in the evidence that reaches the context window. This appendix measures that evidence directly against the gold spans stored with each item, with no answerer involved, and then bounds what better retrieval could still buy.

Recall at fixed budgets. Table 25 extends Figure 2 (right) to every dimension and to the mean reciprocal rank of the first gold hit. The per-dimension gap is sharpest on single-slide lookup, where the walk seeds the nodes inside the time window stated by an easy-tier question and arrives at the

System	gemma	LLaVA-OV	Qwen3-VL
MM-Graph-RAG	**63.9**	**65.0**	**65.0**
CLaMR	62.7	61.0	62.0
VideoRAG-Ren	13.3	62.3	61.3
Video-RAG	60.3	55.7	59.0
Zero-shot	57.3	51.7	55.7
ColPali	57.0	53.0	56.3
RAG-Anything	55.7	56.0	51.7
Agentic	55.0	49.0	53.3
VideoAgent	37.0	53.7	52.7
LLoVi	6.7	53.3	54.7
VGent	27.7	47.3	41.7
VideoTree	3.0	49.0	42.7
M2A	27.0	34.3	36.7

Table 24: Video-MME: every system and answerer (letter accuracy, %).

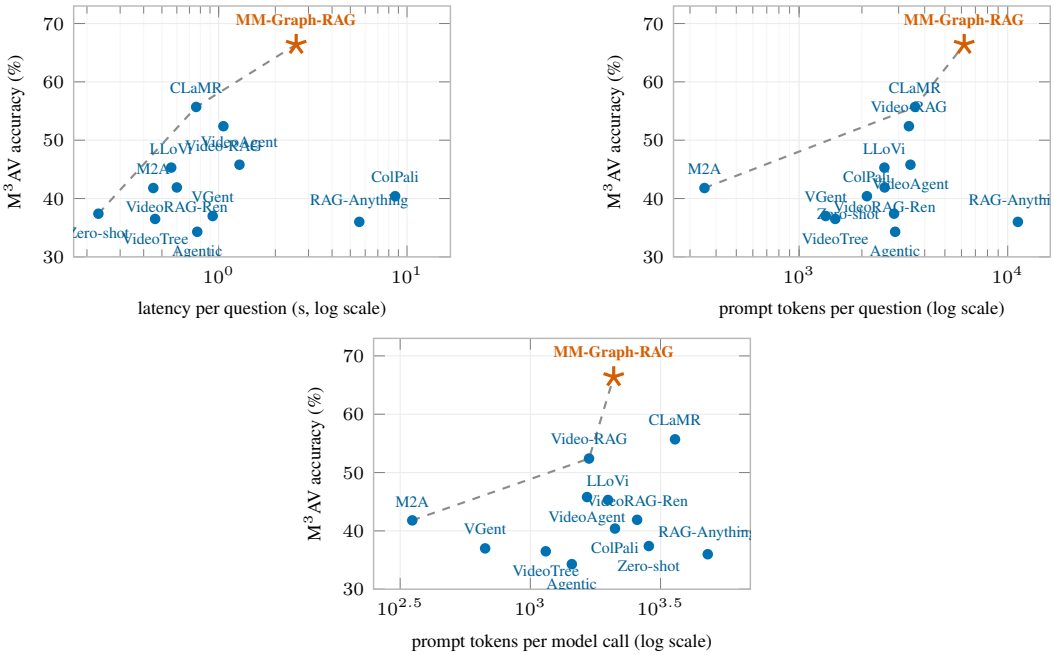

Figure 8: Accuracy against inference cost on M^{3}AV (gemma answerer; log scales): latency per question (top left), prompt tokens per question (top right) and prompt tokens per model call (bottom), the context the answerer reads in a single call, obtained as prompt tokens per question divided by model calls per question. The dashed line traces the efficient frontier in each panel; MM-Graph-RAG sits on it in all three.

slide through the speech–slide alignment edge, while a text-similarity retriever has no reason to rank a slide first. BM25 keeps a slightly better first-hit rank on long-context questions (0.817 against 0.803), so lexical anchors remain competitive for locating a recurring concept.

Retriever	p1e	p1h	p2	p3	p4	overall
Recall@3						
MM-Graph-RAG	**99.7**	**82.6**	**80.8**	43.7	**58.3**	**76.7**
BM25	54.3	79.0	75.0	**43.9**	51.7	63.1
Dense	50.7	73.8	71.6	38.3	51.7	59.1
Uniform	22.6	21.0	18.4	27.5	13.1	21.4
Recall@8						
MM-Graph-RAG	**99.7**	**90.8**	**91.4**	**64.5**	**75.3**	**86.8**
BM25	69.0	89.2	87.2	62.0	69.0	77.0
Dense	63.8	84.3	86.8	59.1	70.1	74.2
Uniform	47.0	52.5	54.0	55.3	50.9	52.1
Mean reciprocal rank of the first gold hit						
MM-Graph-RAG	**0.830**	**0.752**	**0.683**	0.803	**0.810**	**0.763**
BM25	0.486	0.729	0.659	**0.817**	0.804	0.673
Dense	0.435	0.651	0.597	0.705	0.784	0.603
Uniform	0.214	0.218	0.195	0.584	0.271	0.287

Table 25: Gold-evidence retrieval quality on M^{3}AV, independent of any answerer. Recall is the fraction of a question's gold evidence moments overlapped by the retrieved nodes, averaged over questions; MRR is the mean reciprocal rank of the first gold hit. The two recall blocks fix the budget at k nodes (Recall@5 lies between the two shown and is omitted).

Fair-budget recall against CLaMR. CLaMR retrieves longer units than the graph's segments, so a comparison at equal item counts favours it and a comparison at equal seconds of video consulted is the fair one. Table 26 reports both on the 1,465 M^{3}AV questions where both retrievers return evidence. At three items CLaMR covers 115 s of video against 41 s for the graph walk and reaches higher recall; at an equal amount of video consulted the graph walk reaches higher recall at every

budget (79.8 vs. 77.5 at 45 s, 86.3 vs. 84.1 at 75 s, 90.6 vs. 89.2 at 120 s) while using two to three times as many, and correspondingly shorter, items. The walk therefore retrieves more of the gold evidence per second of recording it places in the context, which is the quantity that matters once the dilution result of Appendix E.6 is taken into account.

Equal item count			*Equal seconds of video consulted*		
Budget	MM-Graph-RAG	CLaMR	Budget	MM-Graph-RAG	CLaMR
3 items (41 s vs. 115 s)	77.3	**88.3**	45 s (5.5 vs. 1.8 items)	**79.8**	77.5
5 items (57 s vs. 183 s)	82.6	**93.9**	75 s (8.2 vs. 2.9 items)	**86.3**	84.1
8 items (80 s vs. 270 s)	87.1	**97.1**	120 s (10.7 vs. 4.7 items)	**90.6**	89.2

Table 26: Gold-evidence recall (%) of the graph walk and of CLaMR on M^{3}AV at matched item counts (left) and at matched seconds of video consulted (right); parentheses give the mean video covered, or the mean items used, by each retriever at that budget.

Headroom. Table 2 and Table 27 bound what better retrieval could still buy. Under gemma the method has closed roughly three quarters of the zero-shot-to-oracle distance (Section 5) and under LLaVA-OneVision 84% of it (66.6 against 61.1 and 33.2). The remaining points are not evenly spread: multi-hop reasoning is at or above the oracle under both answerers (81.1 against 80.2 and 75.5, differences within the interval of a 106-question pool), so the recoverable headroom sits in temporal grounding and cross-modal synthesis.

Context (LLaVA-OV answerer)	p1e	p1h	p2	p3	p4	overall
Zero-shot (no retrieval)	20.5	24.9	42.8	34.2	54.7	33.2
MM-Graph-RAG	55.6	67.9	62.8	51.7	**81.1**	61.1
Gold-evidence oracle	**61.7**	**75.7**	**71.4**	**53.4**	75.5	**66.6**

Table 27: Headroom on M^{3}AV under LLaVA-OneVision: zero-shot, MM-Graph-RAG, and an oracle that answers from the gold evidence with the same answerer and prompt (judge accuracy, %). The corresponding gemma rows are in Table 2.

Where the remaining errors live. Table 28 joins per-question retrieval coverage with correctness on 1,615 M^{3}AV questions under gemma; coverage is measured on the graph walk's first eight nodes. Most of the remaining headroom is in the answer stage (Section 5). The exception is long-context reasoning, the one dimension where full coverage is the minority outcome; it is also the dimension on which the controller most often requests a second round (Table 21) and on which graph formation is the only step that helps (Figure 4), so the three analyses agree on where retrieval itself can still improve.

Retrieval outcome	share of questions	accuracy	share of errors
All gold spans retrieved	77.9%	68.7	72.6%
Partially retrieved	16.8%	57.0	21.5%
Gold evidence missed	5.3%	62.4	5.9%

	p1e	p1h	p2	p3	p4
Questions with all gold spans retrieved (%)	99.7	90.8	91.4	29.3	47.2
Accuracy on those questions	61.6	80.9	67.3	58.9	86.0
Errors that occur with the evidence present (%)	99.3	79.1	92.0	26.5	35.0

Table 28: Error decomposition on M^{3}AV (gemma answerer; 543 errors on 1,615 questions; coverage measured on the first eight retrieved nodes). Top: questions grouped by whether the retrieved evidence covered every gold span. Bottom: per-dimension coverage, accuracy when covered, and the share of that dimension's errors that occur with the evidence present.

H.1 Independent audit of the judge

Design. We drew 20% of the M^{3}AV questions (323; stratified by reasoning dimension, seed 0) and took every main-table prediction for them: 13 systems under 3 answerers, 12,595 predictions in total. Each prediction was re-scored by an independent human auditor (following the guidelines in

Appendix A.6), distinct from the authors and given the question, the reference answer, the prediction, and the full evidence stored with the item: the moment in the recording, the on-screen text of the slide at that moment, the speech at that moment, the matched paper sentences, the cited segments, and the reasoning chain. The auditor evaluated each case blind to the LLM judge's initial decision (*blind* verdict). On every disagreement, the auditor was shown the judge's verdict and stated justification to re-examine the case and issue a final consensus verdict, overriding their initial assessment only if the judge's stated reason exposed a genuine misinterpretation (*adjudicated* verdict).

Auditor guidelines. The human auditor was instructed to (1) read the evidence before the prediction, derive the answer themselves from the evidence, and record it; (2) check that the reference answer is supported by the evidence, treating the reference as ground truth for scoring in every case but flagging it when the stored text evidence does not confirm it; (3) compare the prediction with the reference under the benchmark's rubric: a prediction is correct when it conveys the same answer as the reference, where synonyms, standard abbreviations, equivalent numeric forms, unit spelling, capitalisation, articles and word order do not matter; it is incorrect when it is partial, adds a claim that contradicts or changes the meaning of the reference, hedges or refuses, gives a different value, answers a different question, or lists several candidates without committing to one; a numeric prediction must match the reference value, and a different precision is accepted only if it rounds to the reference; extra words that merely restate the question do not make a correct value incorrect, extra words that introduce a different fact do. Each verdict was returned with a concise justification.

Results. Table 29 reports agreement between judge and the human auditor. The blind verdicts agree with the judge on 92.3% of predictions (Cohen $\kappa = 0.840$), and after adjudication on 96.0% ($\kappa = 0.917$); of the 976 blind disagreements, adjudication resolved 462 in the judge's favour and 514 in the auditor's. Against the auditor's blind verdicts the judge's precision and recall for "correct" are 91.7% and 89.5%, and its errors are balanced: it accepted 425 predictions the auditor rejected and rejected 551 the auditor accepted, so it is marginally conservative (40.6% of predictions accepted against the auditor's 41.2%). Agreement is highest on the slide-reading dimensions ($\kappa = 0.92$) and lowest on cross-modal synthesis ($\kappa = 0.77$), whose short synthesised answers admit more paraphrase. The auditor flagged the reference as not confirmable from the stored *text* evidence on 3.6% of items, almost all temporal-grounding items whose value exists only in the slide image. Table 30 re-states every system under the human auditor: the ordering of the 13 systems is preserved under every answerer (Spearman $\rho = 0.96$, 0.98 and 1.00 against the judge's ordering on the same sample), no system moves by more than 4.0 points, and MM-Graph-RAG remains first with a margin of $+9.6$, $+13.0$ and $+6.2$ points over the strongest baseline under gemma, LLaVA-OneVision and Qwen3-VL. The judge therefore neither favours our answer style nor changes any conclusion drawn from Table 1. The verdicts and justifications are released with the graded predictions.

	n	Blind agreement	κ (blind)	Adjudicated agreement	Judge acc.	Auditor acc.
All predictions	12,595	92.3 [91.8, 92.7]	0.840	96.0	40.6	41.2
gemma	4,199	91.2	0.819	95.4	41.0	42.2
LLaVA-OV	4,197	93.4	0.858	96.8	35.5	36.1
Qwen3-VL	4,199	92.1	0.841	95.7	45.2	45.2
p1e	2,962	96.3	0.915	97.6	31.7	31.0
p1h	2,379	95.8	0.916	98.2	44.1	44.9
p2	3,900	88.8	0.768	94.9	41.0	40.1
p3	2,535	90.0	0.795	93.9	40.1	43.4
p4	819	90.5	0.794	95.1	62.0	65.4

Table 29: Judge audit on 20% of the M^{3}AV questions, all systems and answerers. Agreement (%) and Cohen κ between the judge and the independent human auditor, before (blind) and after adjudication; the bracket is a 95% bootstrap interval. Judge and auditor accuracy are the share of predictions each accepts (auditor: adjudicated verdicts).

Robustness to recording length. Figure 9 splits M^{3}AV into tertiles by number of segments. Every system degrades with length; the question is the slope. MM-Graph-RAG loses 5.5 points from the shortest to the longest tertile against 13.3 for zero-shot and 14.7 for ColPali; CLaMR is flatter still (3.1) but nine to twelve points lower in every bucket, and MM-Graph-RAG is the most accurate system in each of the three. The two steepest curves belong to the systems that sample frames uniformly or rank by flat similarity, which is the behaviour the graph walk replaces; the SlideSpeech results, where recordings are ten times longer, extend the same pattern (Table 18).

	gemma		LLaVA-OV		Qwen3-VL	
System	judge	auditor	judge	auditor	judge	auditor
Zero-shot	34.4	35.9	34.1	34.1	40.2	39.0
Agentic	29.7	32.2	30.0	30.7	38.1	38.4
LLoVi	42.1	42.1	34.7	35.0	44.0	44.6
VideoAgent	42.7	42.4	34.9	35.2	41.8	40.6
VideoTree	33.4	33.7	31.9	31.9	35.0	33.7
VGent	34.4	34.4	30.3	30.3	36.5	36.5
Video-RAG	48.3	51.7	46.1	46.7	55.7	55.7
VideoRAG-Ren	41.2	41.5	35.6	37.2	47.1	46.7
RAG-Anything	38.4	37.2	9.0	10.2	19.2	16.7
M2A	38.4	41.5	30.7	32.8	41.2	40.2
ColPali	34.4	38.4	35.9	36.2	63.8	64.4
CLaMR	51.7	53.9	48.6	48.3	58.2	59.8
MM-Graph-RAG	**64.1**	**63.5**	**60.4**	**61.3**	**66.6**	**70.6**

Table 30: Every main-table system on the audited 323 M^{3}AV questions: accuracy (%) under the judge and under the independent human auditor (adjudicated verdicts). The ordering is preserved under every answerer and MM-Graph-RAG remains first.

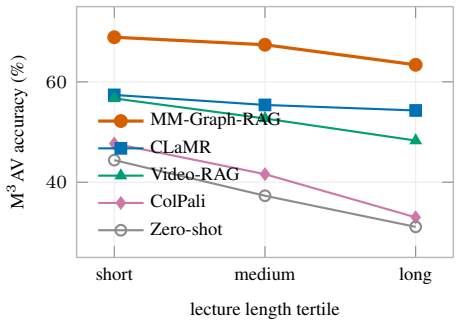

Figure 9: M^{3}AV accuracy by lecture-length tertile (gemma answerer; short ≤ 12 segments, medium ≤ 16, long > 16). Systems that rely on uniform frame sampling or on flat similarity degrade fastest.

I Reproducibility and evaluation protocol

Infrastructure. All models are served with vLLM on H200 GPUs: gemma-3-27b-it, LLaVA-OneVision-7B and Qwen3-VL-8B-Instruct as answerers, Qwen3-Embedding-8B (4,096-d) for dense retrieval and Llama-3.3-Nemotron-Super-49B as the judge. Every system in a table is served by the same endpoints. Work is sharded by recording, never by question, so each shard indexes a disjoint corpus and no two writers share a retrieval store. Predictions are written first and graded in a single uniform pass, so every row of every system is scored by the same judge under the same settings.

Baseline configuration. Every baseline reads the same released transcript, slide images with OCR and, on M^{3}AV, source paper as MM-Graph-RAG, answers through the shared answerer endpoint at temperature 0, and uses Qwen3-Embedding-8B wherever it needs a text embedding; a system that hands the video to the answerer supplies eight frames sampled evenly across the recording, the same for every such system. All settings below were fixed before evaluation and are identical across answerers and collections. LLoVi, VideoTree, VGent, VideoAgent and Video-RAG are our implementations of the published methods on the shared endpoints; RAG-Anything and M2A run from their released packages, and ColPali and CLaMR use their released retrievers with our answer step. *Zero-shot* sends the full transcript, the paper where one exists and the frames in one call. *Agentic* makes one planner call that writes a retrieval plan (up to six slide indices, a speech window, the full transcript or a paper keyword search), assembles that context and answers in a second call with the frames. *LLoVi* captions eight evenly sampled frames with the answerer and answers from the captions and the transcript. *VideoTree* builds, once per recording, a one-sentence summary of each 120-word transcript leaf and a two-sentence overview, then expands the three leaves closest to the question. *VGent* decomposes the question into at most three sub-questions, retrieves the two best-matching 100-word transcript chunks for each, and answers after a verification step. *VideoAgent* starts from five captioned frames and, for up to two rounds, either answers or names a region of the

video from which four further frames are captioned. *Video-RAG* builds OCR and speech databases from the aligned segments and an 80-word-chunk paper database, has the answerer decouple the question into OCR and speech requests, retrieves the five OCR chunks, five speech chunks and three paper chunks above cosine 0.3, and answers with the retrieved text and the frames. *VideoRAG-Ren*, the VideoRAG of Ren et al. (2026), retrieves the six best 110-word transcript chunks on its text channel and the four best of sixteen evenly sampled frames by CLIP ViT-B/32 similarity on its visual channel, and fuses both. *RAG-Anything* is the released package with its LightRAG knowledge graph built per recording from the slides (image items captioned by their OCR), the transcript and the paper, queried in its mix mode with up to eight retrieved slide images and 28k characters of context. *ColPali* encodes every slide keyframe with ColQwen2.5 as a multi-vector document, ranks slides by MaxSim against the question and shows the five best to the answerer with the transcript and paper. *CLaMR* applies the released late-interaction adapter on its Qwen2.5-VL-3B base to encode each aligned segment (slide image, OCR and speech) and each 80-word paper chunk as a document and retrieves the eight best; the release is a retriever only, so the answerer receives the retrieved segment text, paper chunks and the frames. *M2A* is the released memory agent with fresh stores per recording, populated by replaying slide OCR, speech and paper as a synthetic conversation and queried through its hybrid dense and BM25 retrieval, with up to two slide images shown to the answerer. In Appendix D, M2A (tool-calling) routes the agent's memory lookups through the answerer's tool-calling interface and M2A (no memory) answers without a memory lookup. The +docs configurations append the slide OCR and a paper excerpt to the prompt without changing a method's own retrieval.

Protocol notes. Beyond the fairness controls of Section 4, judge inputs are truncated to 3k characters so that every prediction fits the judge's context, and every main-table cell uses the complete question pool of its benchmark; paired analyses use the questions on which every compared configuration returned a prediction, and their captions give the count. The corpus, the generation and evaluation code, the per-item audit trail and the graded predictions of every system are released.