WHEN LATER EVIDENCE CHANGES THE VERDICT: SNAPSHOT-BOUND CITATION EVALUATION FOR STREAMING MULTIMEDIA QA

Anonymous authors
Paper under double-blind review

ABSTRACT

An answer issued while OCR and crop extraction are incomplete can later receive a better citation score even though its text and returned citation handles never change. Completed-index evaluation therefore lets future evidence alter the recorded quality of a past multimedia QA response. Snapshot-bound evaluation exposes this temporal leakage by committing each output to source-native evidence roots and its served snapshot, then replaying those fixed handles through an explicit provenance relation under served and settled revisions. On 72 streams spanning 187 hours and 14,382 verified QA pairs, settled replay raises citation correctness by 1.8 points for ServeTrace, 3.3 for Rolling-ANN, and 4.4 for Tier-less indexing; all stream-block 95% confidence intervals exclude zero. The service mechanism is visible across backlog strata: sketch-only answers rise from 8.3% to 32.4%, served citation correctness falls from 82.4% to 76.1%, and QA F1 on answered queries falls from 72.0 to 69.1. At the prespecified 80% line, 4.1–6.7% of answer-level labels change, with fail-to-pass transitions outnumbering reversals by 11–14 times. The effect transfers to the released Video-RAG pipeline on TVQA+ (+3.4 points, 95% CI [2.4, 4.5]) and to InternVL3-38B. ServeTrace realizes the protocol at 82.1% served citation correctness, 2.1 GB/h median storage, and 2.58 s p95 latency, while remaining within 0.8 QA-F1 points of Store-all. The resulting audit distinguishes evidence verifiable at response time from support that becomes resolvable only after the decision.

1 INTRODUCTION

Consider a response returned while a live video index contains a coarse ASR window but not the lower-third crop that identifies the speaker. Minutes later, asynchronous enrichment attaches the crop and a precise text span to the same source event. Scoring the frozen response only against the completed index now records a precise citation that the service could not expose when it answered. The output is unchanged; the verdict is not.

This mismatch matters wherever evaluation, monitoring, or policy decisions are made after a mutable multimedia index settles. A timestamp alone cannot recover the relevant counterfactual. The evaluator must know which evidence object the response actually committed, which fields were resolvable then, and whether a later crop is a refinement of that object or a different item that retrieval never returned. Existing grounded QA and citation metrics evaluate support in the evidence collection presented to the scorer. They do not preserve the identity and revision history needed to measure post-response changes in verifiability.

Snapshot binding supplies that missing audit object. Each response freezes the answer, citation-handle set, source-native primitive roots, served locators, resolution states, and evidence snapshot. An append-only provenance relation permits one-to-many refinements and many-to-one merges while preventing an unrelated item discovered after the response from entering the cited handle. Human annotation defines output-independent gold support slots and validates ambiguous source mappings; deterministic matching then assigns exact, over-broad, shifted, missing, or wrong-item outcomes. Replaying the same handles under served and settled snapshots yields a paired *settlement correction*:

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.

the change in verifiability caused by evidence-version materialization, with retrieval output, prompts, questions, and answers held fixed.

The measured consequence is large enough to change reported conclusions. Across 72 streams, the correction is +1.8 points for ServeTrace, +3.3 for Rolling-ANN, and +4.4 for Tier-less indexing, with all stream-block confidence intervals above zero. At an 80% reporting line, 4.1–6.7% of answer labels change and fail-to-pass transitions dominate reversals. The same direction appears under a released Video-RAG implementation on the official TVQA+ validation episodes and after replacing Qwen2.5-VL with InternVL3-38B. Backlog supplies the operational link: between the lowest and highest observed strata, sketch-only service rises by 24.1 points while citation correctness drops 6.3 points and answered-query QA F1 drops 2.9 points.

The contributions form one end-to-end audit:

1. **Audit object and scorer.** Snapshot binding combines immutable response handles, retrieval-independent primitive evidence roots, split/merge provenance, alternative gold support slots, deterministic matching, and explicit wrong-item handling.

2. **Paired decision protocol.** Served/settled replay measures version-dependent citation verifiability for fixed outputs at answer, stream, and system levels, with stream-block uncertainty and threshold-integrated decision sensitivity.

3. **Cross-stack evidence and realization.** The correction persists across event shards, rolling windows, Video-RAG, two vision-language backbones, and TVQA+; ServeTrace records the required state while operating at 2.1 GB/h median storage and 2.58 s p95 latency.

2 Related Work

Static grounded QA evaluates support in a fixed document collection, while factuality and quote-grounding metrics test whether generated claims align with scorer-visible evidence (Lewis et al., 2020; Guu et al., 2020; Izacard et al., 2023; Menick et al., 2022; Min et al., 2023). AIS formalizes whether a generation is attributable to an identified source, and ALCE evaluates citation correctness and completeness for retrieved passages (Rashkin et al., 2023; Gao et al., 2023). These protocols establish *what* counts as support; their cited source payload is fixed during scoring. Snapshot binding addresses the orthogonal case in which the cited multimedia object itself gains, loses, splits, or merges locators after the output is committed.

StreamingQA assigns question and knowledge dates and evaluates adaptation as articles enter a corpus (Liska et al., 2022). Its model and knowledge state advance with time. Our paired protocol instead freezes the question, answer, retrieval result, and citation handles, then varies only the versioned evidence materialization available to the scorer. This fixed-output contrast isolates evaluator-state drift from model adaptation and repeated retrieval.

Temporally grounded video QA benchmarks such as TVQA+, together with long-video suites such as MLVU and Video-MME, evaluate localization over completed videos (Lei et al., 2020; Zhou et al., 2025; Fu et al., 2025). Video-RAG enriches long-video models with aligned OCR, ASR, and object text, while VideoRAG retrieves multimodal items from a video corpus (Luo et al., 2025; Jeong et al., 2025). Snapshot-bound evaluation complements these systems by preserving source-native identity and response-time resolution across provisional and settled revisions. Table 1 contrasts the scorer-visible state and the resulting estimand.

3 Scoring Citation Handles Across Evidence Revisions

Snapshot-bound evaluation separates the protocol from its systems realization. The protocol follows source-native evidence roots through versioned materializations; ServeTrace supplies those records for event-sharded, tiered ingest. Fixed-window and Video-RAG outputs use adapters to the same primitive ledger.

Evaluation object	Scorer evidence	Citation unit	Evidence-state comparison
AIS / static grounded QA	Fixed identified source	Quote or text span	Attribution under one scorer-visible source
ALCE	Fixed retrieved passages	Passage citation	Correctness and completeness of generated citations
StreamingQA	Date-bounded evolving corpus	Retrieved text evidence	Questions and model knowledge advance over time
Settled long-video QA	Completed video and annotations	Temporal/spatial region	One completed state supplies the score
Video-RAG / VideoRAG	Retrieved completed assets	Retrieved span or multimodal item	Retrieval is scored after asset preparation
Snapshot-bound scoring	Bound served and settled revisions	Fixed handle over source-native roots	Same output is replayed; version-dependent verifiability is measured

Table 1: Evaluation objects and the evidence state each scorer can inspect. Snapshot-bound scoring adds response-time identity and revision provenance to established attribution and citation-quality criteria.

3.1 Audit record and paired intervention

For query q_i, the service commits an immutable record

$$R_i = (q_i, a_i, C_i, t_i, E_i^{\mathrm{srv}}, z_i, b_i, p_i, \rho_i^{\mathrm{srv}}),$$

where a_i is the fixed answer, C_i the ordered returned handle set, t_i query time, E_i^{srv} the answer-time evidence snapshot, z_i its served-tier identifier, b_i the answer-time age of the oldest relevant sketch shard with unfinished full backfill, p_i the active answer policy, and ρ_i^{srv} the per-handle resolution status. We set $b_i = \perp$ when no relevant shard is in flight. A handle commits its source-native root keys, the locator payload printed in the response, and the served resolution state; all three remain immutable. Snapshot binding is the immutable association between (a_i, C_i) and $(E_i^{\mathrm{srv}}, z_i, b_i, p_i, \rho_i^{\mathrm{srv}})$. Subsequent enrichment appends a distinct E_i^{set} rather than overwriting this record. The settled snapshot is the first materialization after every refinement job reachable from a root in C_i reaches a terminal completed, retracted, or failed state.

3.2 Canonical identity and citation score

The identity layer is defined over source primitives rather than retrieval containers. Every frame, crop, OCR span, or ASR span has a canonical key

$$\kappa(u) = (\text{source}, \text{stream}, \text{modality}, \text{source-native primitive id})$$

and an immutable source locator. Event-shard membership, fixed-window membership, retrieval rank, and sketch/full status are versioned attributes, not key components. Retrieval adapters emit the primitive roots contained in a returned event or window; Video-RAG spans use source element ids and timestamps to address the same ledger. If an adapter cannot select a unique primitive root, the handle remains unresolved and is scored rather than dropped.

An append-only provenance relation Π_E records REFINES, SPLITS, MERGES, and RETRACTS edges among primitive keys. It may therefore be one-to-many or many-to-one. For a committed handle c, $M(c, E)$ is the active materialization reachable from its frozen root set under Π_E. A key discovered after response time can contribute only when an explicit provenance path begins at a root already in c; an unrelated item is outside $M(c, E)$. This rule handles event-boundary splits and merges without making event identity part of the canonical key. The response payload remains in R_i for audit, while the paired estimand intentionally compares the handle's version-specific materializations.

Before any method output is exposed, annotators construct K_i method-independent support slots $\mathcal{G}_i = \{G_{i1}, \ldots, G_{iK_i}\}$ from the question, reference answer, and source stream. A slot contains source-native alternatives that are independently sufficient for the same support obligation; for example, either an aligned speaker frame or an equivalent lower-third crop may satisfy a visual-identity slot. Required multimodal evidence occupies separate slots. The ledger is frozen before scoring, preventing an output from defining its own acceptable evidence.

For snapshot E, the scorer materializes and deduplicates the handles in C_i. Two materializations are duplicates only when their frozen root sets and child-locator payloads agree and their intervals overlap after symmetric ± 2 s dilation; their interval union is retained once. Conflicting payloads remain separate. Let $[\ell_c, r_c]$ and $[\ell_g, r_g]$ denote returned and gold intervals. A bipartite edge from a returned materialization to a gold slot is eligible when it reaches an alternative in that slot, is active in E, satisfies $|\ell_c - \ell_g| \le 2$ s and $|r_c - r_g| \le 2$ s, and agrees with every required frame, crop, OCR-span, or ASR-span locator. Maximum-cardinality one-to-one matching gives $\mathrm{TP}_i(E)$.

Diagnostics follow a deterministic priority order. A matched edge is *exact*; an active reachable interval containing a gold alternative but violating endpoint tolerance is *over-broad*; another active reachable record with an endpoint or required-locator conflict is *shifted*; a root with no active sufficient materialization is *missing*; and an active materialization with no edge to any gold slot is *wrong-item*. The compact raw-count table combines missing and wrong-item outcomes as *unverifiable*; both are false positives, while every unmatched gold slot is a false negative. Thus

$$P_i(E) = \frac{\mathrm{TP}_i(E)}{|\bar{C}_i(E)|}, \quad R_i(E) = \frac{\mathrm{TP}_i(E)}{K_i}, \quad g(a_i, C_i; E) = 100 \frac{2P_i(E)R_i(E)}{P_i(E) + R_i(E)},$$

with $g = 0$ when either required set is empty or the denominator is zero. Thus $g \in [0, 100]$, table deltas are percentage-point differences, and the thresholds below use the same percentage scale. The support slots credit valid alternative evidence without weakening recall, and the provenance constraint prevents later unrelated retrievals from receiving credit.

Served and settled scores are

$$s_i^{\mathrm{srv}} = g(a_i, C_i; E_i^{\mathrm{srv}}), \qquad s_i^{\mathrm{set}} = g(a_i, C_i; E_i^{\mathrm{set}}),$$

with paired settlement correction $\Delta_i = s_i^{\mathrm{set}} - s_i^{\mathrm{srv}}$. The ordered pair $(s_i^{\mathrm{srv}}, s_i^{\mathrm{set}})$ is the *paired evidence-state intervention*: E and its reachable materialization change, while the answer, committed handles and roots, response payload, query population, prompt, and retrieval output remain fixed. The estimand is therefore version-dependent verifiability of a fixed output under an unchanged retrieval run. For threshold τ, $V_\tau(s) = \mathbb{1}[s \ge \tau]$ yields the answer-level transition $V_\tau(s_i^{\mathrm{srv}}) \to V_\tau(s_i^{\mathrm{set}})$.

The evidence journal is append-only, but support is not assumed monotone: a later entry may refine a payload, tombstone a provisional OCR hypothesis, merge duplicates, or split an erroneous event boundary. The provenance relation represents each operation explicitly, so settled materialization permits both positive and negative Δ_i without rewriting R_i. The negative corrections in Table 16 arise from retraction or resegmentation, not from replacing C_i with newly retrieved handles.

Let S_m^e denote method m's aggregate score under evidence state $e \in \{\mathrm{srv}, \mathrm{set}\}$. Its aggregate settlement correction is $D_m = S_m^{\mathrm{set}} - S_m^{\mathrm{srv}}$. Across a prespecified threshold set $\mathcal{T}$, we summarize answer-level sensitivity as

$$I_m = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \frac{1}{n_m} \sum_i \mathbb{1}\left[V_\tau(s_{mi}^{\mathrm{srv}}) \ne V_\tau(s_{mi}^{\mathrm{set}})\right].$$

The continuous D_m is primary; threshold transitions and I_m describe sensitivity without treating a reporting cutoff as a universal acceptance criterion. Cross-method ranks of citation scores are not an estimand because methods can answer different query subsets.

For stream j, let A_j be its citation-required answered records. Aggregate citation tables report the equal-stream macro estimator $\frac{1}{J} \sum_j |A_j|^{-1} \sum_{i \in A_j} g_i$, averaged over five seeds. Backlog tables instead micro-average records within each bin, $\sum_i \mathbb{1}[b_i \in B] g_i / \sum_i \mathbb{1}[b_i \in B]$, so unequal stream contributions do not make a count-weighted average of the displayed bin means equal the aggregate score. Coverage is the fraction of all submitted queries that return an answer.

The observation units are explicit. A submitted query is the QA unit; all queries from one source stream form the resampling block; a seed is one complete randomized replay of those streams. System-level means are computed within seed and then averaged. Integer transition tables use one row per unique citation-required query: its served and settled scores are first averaged across the five replay realizations and thresholded once, so each row's counts sum to its displayed query denominator.

For RQ1, records with numeric $b_i \in [0, 15]$ are partitioned into 0–1, 1–5, and 5–15 minute answer-time queue-age bins. This admission rule covers the 13,234 submitted and 12,112 answered records in

Table 20; the remaining 1,014 of ServeTrace's 13,126 answered records have $b_i = \perp$ or $b_i > 15$ and enter full-workload aggregates but not the endpoint contrast. All 8,624 citation-required ServeTrace answers meet the RQ1 admission rule and are partitioned in Table 21. Confidence intervals use 10 000 matched block-bootstrap replicates: each replicate samples all 72 streams with replacement, applies the same block indices to every method and scorer state, recomputes each seed-level statistic, and averages the five paired seed estimates. We report the 2.5th and 97.5th percentiles. TVQA+ substitutes its 431 episodes for streams. This record is necessary because E_i^{set} can establish eventual support but cannot reconstruct whether C_i localized an auditable item in E_i^{srv}.

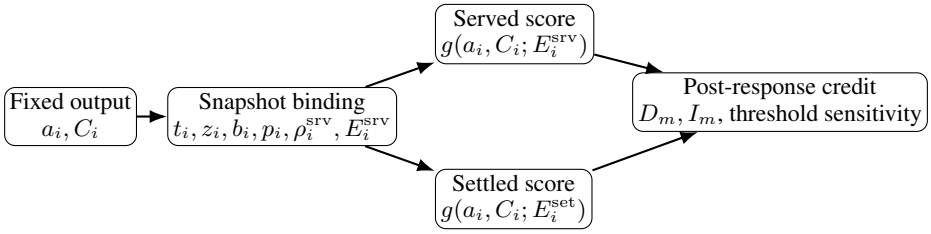

Figure 1: Snapshot-bound scoring. The answer, handles, source-native roots, and response payload remain fixed; served and settled snapshots select different provenance-reachable materializations.

Figure 1 locates the audit boundary at response commitment. Segmentation organizes retrieval but does not define identity; source primitives and provenance define the citation unit across event and window implementations. In a representative trace, a handle rooted in a coarse briefing interval is over-broad when served, then becomes exact after its reachable four-second ASR span and lower-third crop materialize. A crop rooted in a neighboring montage remains a wrong item even if it is relevant, because it is outside the committed handle's provenance closure. This worked case illustrates the score change without permitting post hoc retrieval.

3.3 Evidence-addressable event shards

Each event shard stores 4–12 keyframes, salient crops, timestamped OCR entities, ASR spans, multimodal embeddings, and a citation index from answer spans to minimal evidence items. The index makes timestamp, frame, crop, OCR, and ASR support directly addressable by the scorer.

3.4 Event boundary detection

We combine ASR topic shifts (embedding drift + CUSUM), visual drift (CLIP cosine distance), and OCR/ASR entity churn. Hysteresis prevents over-segmentation during noisy bursts and makes downstream citation errors traceable to concrete boundary decisions. Cheap signals yield the boundary score $s_t = w_a \cdot \Delta\text{ASR}(t) + w_v \cdot \Delta\text{CLIP}(t) + w_e \cdot \Delta\text{EntityChurn}(t)$, where windowed embedding drift or churn rate is denoted by Δ. Emission requires s_t to cross the high threshold ($\tau_h = 0.72$) and stay above it throughout a 3 s dwell time; to reduce burst fragmentation, merging occurs once the score drops below the low threshold ($\tau_l = 0.35$). The same threshold pair is fixed for every stream. Distinct events can be merged by missed boundaries, while false-positive boundaries can split an event; downstream retrieval and QA suffer in both cases. Table 8 quantifies detection quality and the resulting impact on downstream QA.

3.5 Tiered ingest

During spikes, backlog is controlled by storing compressed shards (embeddings + sparse keyframes + coarse entities) in the sketch tier while the full tier asynchronously backfills refined OCR and crops. Provisional retrieval with tier-tagged citations uses sketch-tier shards; richer evidence from full backfill permits finer-grained citations at crop level. Recording both snapshots instantiates the paired intervention above. At query time, backlog b_i is measured as the elapsed wall-clock time since creation of the oldest answer-candidate sketch shard whose full-tier job is still unfinished; this is observable before the answer is committed.

Under GB/hour caps, a fixed utility controller uses query frequency, novelty, entity importance, evidence quality, and recency to choose keep/compress/evict. Its utility is $U(s) = w_d \cdot \text{Demand}(s) + w_q \cdot \text{Quality}(s) + w_n \cdot \text{Novelty}(s) - w_c \cdot \text{Cost}(s)$; the default weight setting is $(w_d, w_q, w_n, w_c) = (0.35, 0.30, 0.20, 0.15)$. A hard override retains every shard that has recent citation hits; otherwise, fixed thresholds select `KEEP` for $U \geq 0.70$, `COMPRESS` for $0.45 \leq U < 0.70$, or `EVICT` for $U < 0.45$. One set of weights and thresholds is fixed before evaluation and applied to every stream.

With retention constrained by a GB/hour cap, the three available choices are {`KEEP`, `COMPRESS`, `EVICT`}; a fixed entity-frequency heuristic assigns the rare-event tag for both the evaluation subsets and the retention override protecting rare evidence.

Citation fidelity constrains compression: ServeTrace retains, in each compressed shard, (i) keyframes/crops that past answers referenced, (ii) entities of high importance, and (iii) a minimal timestamp set covering event subclaims. Write $E(s)$ for the evidence items of shard s (keyframes, crops, OCR spans, ASR windows), and $E_{\text{cite}}(s) \subseteq E(s)$ for those items to which the citation index of previously served answers points. The compression operator produces $E'(s) \subseteq E(s)$ subject to $E_{\text{cite}}(s) \subseteq E'(s)$, retaining subclaim coverage (timestamp windows for each answer span) within a tolerance of ($\pm$2s). At answer-span level, we measure compression's citation correctness and missing-evidence events through timestamp agreement within ± 2 s, valid frame ids, and the presence of crop ids. The missing-evidence rate counts the fraction of answer spans for which the cited item is missing from memory or is invalid for the tier that served the answer.

For a routed question, the reference design retrieves candidate shards at query time from inexpensive event-level embeddings and entity signals; it subsequently expands just the required modalities (OCR, frames, audio). By determining the evidence available to cite within the SLA, this path operationally defines the served tier. Answers render explicit (`timestamp`, `frame_id`, `crop_id`) locators; the snapshot-binding record associates each rendered locator with its immutable primitive roots and served resolution state.

4 Experimental Setup

The experiments answer three questions in order. **RQ1** asks how incomplete backfill appears at service time across observed backlog strata. **RQ2** measures the decision consequence by replaying fixed citation handles under served and settled evidence revisions. **RQ3** tests whether the correction survives changes in retrieval organization, workload family, source dataset, answer backbone, and annotation sample.

The benchmark's 72 streams supply 187 hours of audio-visual material: a news-video subset aligned with CNN/DailyMail article metadata (Hermann et al., 2015), meetings from AMI (Carletta et al., 2006), and crisis-event compilations from YFCC100M (Thomee et al., 2016); annotation provides 14,382 manually verified QA pairs plus 2,847 gold event boundaries for a stratified subset of 30 streams. Each question must have an answer supported by in-stream evidence and comes with minimal timestamped citations. The verified annotation pool comprises these 14,382 QA pairs; the later Table 14 describes sampling runtime queries from this larger pool for streaming evaluation, not the annotation count. Two annotators independently construct gold support slots and validate ambiguous source-primitive mappings; Cohen's $\kappa = 0.76$ for ServeTrace measures agreement on these mapping decisions. After the gold and mapping ledgers are frozen, citation correctness and the five diagnostic outcomes are computed deterministically.

The 72 streams form an evaluation-only pool: zero benchmark streams are used to fit or select models, scorer rules, boundary weights, controller weights, storage targets, or threshold grids. These choices are frozen before the first primary query and remain fixed across five randomized stream replays. The TVQA+ training split is used only for its separate configuration stage; all transfer results use the untouched official validation episodes. The secondary adjudication pool is sampled before scoring from records absent from the primary annotation pass, stratified by source family, backlog bin, and required modality. Annotators do not see method identities or primary-pool outcomes.

Each QA pair has method-independent support slots containing acceptable timestamp/frame/crop/OCR/ASR alternatives within the ± 2 s rule in Sec. 3. Table 6 reports macro-averaged per-answer citation F1 under the served tier, whereas Table 22 exposes item-level precision and recall on the citation-required subset. Method-specific κ values differ because each

retrieval adapter surfaces a different set of ambiguous unit-to-primitive mappings. Two annotators validate these mappings independently; disagreements go to a third pass with duplicate-window, boundary, sketch-insufficiency, or OCR/ASR-omission reason codes. The resulting ledger is frozen before the deterministic matcher runs. Appendix B gives the rubric, a worked case, and the citation-required admission rule. System-level QA F1 uses the common submitted-query population and assigns zero to abstentions, timeouts, and execution failures. Only rows explicitly labeled "on answered queries," including the RQ1 backlog slices, condition QA F1 on a returned answer.

Across methods, we hold VLM/ASR/OCR versions, prompts, embeddings, the environmental enrichment schedule, caching budget, and the 3 s latency SLA identical; the varying factors are memory organization, retention, retrieval refresh, and exposure of query-time tier state to the answer policy. The memory comparison includes Store-all and Sliding summary, along with Summary-only, Non-event segmentation, Tier-less indexing, and ServeTrace. Served evidence state is examined along three comparator axes: density of raw evidence (Store-all vs. summary baselines), alignment to events (fixed windows vs. event shards), and tier semantics (comparing Tier-less indexing with ServeTrace). For method comparisons, the scorer reads only the answer text, citations returned, coverage, and tags indicating the served tier; it accesses no ServeTrace-specific internals. Unless stated otherwise, results report five-seed mean$\pm$standard deviation, with median [p25,p75] and p95/p99 for heavy-tailed quantities. Retaining event shards but omitting explicit served-tier tags makes Tier-less indexing the nearest comparator for isolating tier semantics. Appendix E supplies three layers of corroboration: Table 23 reports the nearest comparator's error mass; Table 24 provides a rolling-ANN baseline at a matched budget, using fixed windows of 30 s with a 15 s hop, refreshes every 2 s, and HNSW/FAISS-style retrieval (Malkov & Yashunin, 2020; Johnson et al., 2021) with the same scorer and backbone stack; and Table 26 repeats the core evaluator-state measurements in a stratified second pool, with adjudicators blinded to method identities and primary-pool outcomes. Appendix D adds threshold sensitivity and scorer perturbations. Table 5 separately evaluates the official 3,017-question TVQA+ validation split with the released Video-RAG retrieval pipeline.

News/meeting streams lasting 1–6 h and crisis streams lasting 1–3 h make up the workload, with injected topic drift, modality dropout, and burst spikes; queries arrive in cold start instead of waiting for ingest to stabilize. For ingest/indexing, experiments allocate two NVIDIA A100 80 GB GPUs; QA inference receives four NVIDIA A100 80 GB GPUs, using Qwen2.5-VL-32B-Instruct (Bai et al., 2025), Qwen2.5-32B-Instruct (Qwen Team, 2024), Whisper-large-v3 (Radford et al., 2023), PaddleOCR v2.7.3 (Du et al., 2020), CLIP ViT-L/14 (Radford et al., 2021), and BGE-M3 (Chen et al., 2024); Appendix B and Appendix C supply further definitions of metrics, schedules, the query mix, parity controls, and notes on reproducibility.

5 Results

5.1 RQ1: Auditability across backlog strata

Backlog (min)	Sketch-only answers (%)	Citation corr. (%)	Coverage (%)
0–1	8.3 ± 2.1	82.4 ± 1.4	93.7 ± 1.1
1–5	18.6 ± 3.8	79.8 ± 1.7	91.4 ± 1.3
5–15	32.4 ± 5.2	76.1 ± 2.1	88.2 ± 1.8

Table 2: ServeTrace under the tier-aware policy across observed answer-time queue-age strata. Higher-backlog records contain more sketch-only answers and lower served citation correctness and coverage.

The low- and high-backlog strata differ sharply in all three observed outcomes. Citation correctness on citation-required answered records drops 6.3 points (95% CI [5.4, 7.1]), whereas coverage drops 5.5 points and QA F1 drops 2.9 points (95% CI [2.1, 3.8]) on answered queries. The low/medium/high bins contain 3226, 3411, and 1987 citation-required answers from 54, 61, and 43 streams, respectively. Under matched stream-block resampling, the citation drop exceeds the aligned QA drop by 3.4 points (95% CI [2.4, 4.6]). This separation motivates the fixed-output analysis in RQ2. The 82.4% low-backlog value is a bin-micro average, whereas the 82.1% ServeTrace operating score is an equal-stream macro average over the full workload; Sec. 3 fixes both estimators.

5.2 RQ2: FIXED-OUTPUT VERDICT CHANGES

Method	Served	Settled	Δ [95% CI] (points)	80% reporting label	Cit.-req. n
ServeTrace	82.1	83.9	+1.8 [0.9, 2.7]	above→above	8624
Rolling-ANN	77.4	80.7	+3.3 [2.2, 4.4]	below→above	8516
Tier-less indexing	76.8	81.2	+4.4 [3.2, 5.5]	below→above	8624

Table 3: Paired scorer-state intervention on fixed answers, handles, primitive roots, and response payloads. Scores are citation correctness (%); only the evidence revision and its provenance-reachable materialization change. The 80% column is a prespecified sensitivity label, not a cross-method rank. Each row fixes its own citation-required population across states.

Table 3 gives the direct fixed-output test. The same immutable handles change by +1.8 to +4.4 points under settled provenance materialization, and all stream-block intervals exclude zero. At the illustrative 80% reporting line, both baselines change labels; the continuous corrections do not depend on that cutoff. The error decomposition explains the larger baseline corrections: Tier-less indexing returns 275 more over-broad and 260 more shifted items than ServeTrace (Appendix Table 23), so its versioned handles have more localization detail to gain after those answers are served.

Method	Fail→pass / pass→fail	Changed label (%)	Median Δ [p25,p75]	75–85 integrated (%)	95% CI
ServeTrace	329/23	4.1	1.4 [0.0, 3.1]	3.6	[3.0, 4.2]
Rolling-ANN	412/35	5.2	2.7 [0.4, 5.0]	4.8	[4.1, 5.6]
Tier-less indexing	528/49	6.7	3.8 [0.9, 6.8]	6.1	[5.3, 7.0]

Table 4: Answer-level threshold transitions. "Integrated" is I_m, the changed-label rate averaged over the prespecified 75–85 range; its stream-block interval is shown in the last column.

For threshold set $\mathcal{T} = \{75, 77.5, 80, 82.5, 85\}$, the distributional view in Table 4 aggregates every cutoff rather than selecting one: fail-to-pass transitions outnumber reversals by 11–14 times, paired medians are positive, and Tier-less indexing has the largest integrated instability. Appendix Table 16 reports pass/pass, fail/fail, increase/tie/decrease counts, all five thresholds, per-stream distributions, and intervals. Figure 2 foregrounds the stream-level effect: the median, interquartile range, and number of positive streams for each retrieval organization.

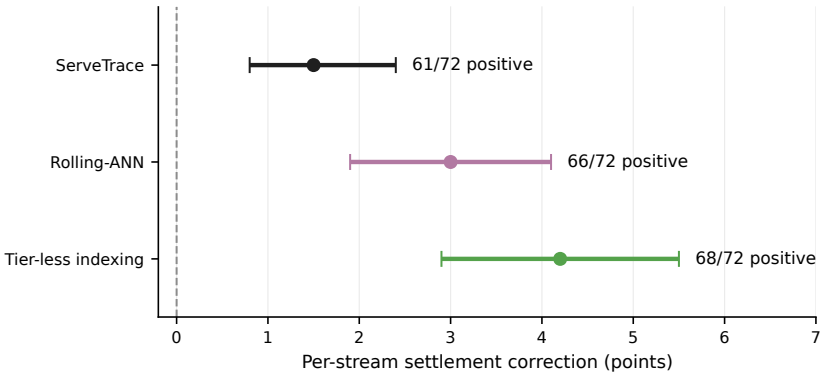

Figure 2: Stream-level settlement corrections on fixed outputs. Points are the median per-stream correction, bars span the interquartile range, and labels report streams with positive corrections.

5.3 RQ3: TRANSFER BEYOND THE SHARED STACK

We replay the official 431-episode, 3,017-question TVQA+ validation split incrementally without changing question or episode membership. Frames and subtitles arrive in timestamp order: the sketch state exposes sparse frames and available subtitle/ASR spans, and the settled state adds official object regions and dense crop/OCR materializations reachable from the same source interval. Questions are issued after their annotated moments become visible; source frame, subtitle, and region ids provide primitive keys, while ambiguous adapters remain in the score. The training split configures the replay schedule only.

ServeTrace, Rolling-ANN, and the released Video-RAG pipeline share Qwen2.5-VL-32B-Instruct, prompts, the nominal 2.5 GB/h target, the 3 s SLA, backlog bins, and seeds $\{13, 37, 61, 71, 97\}$. Video-RAG spans are mapped to source roots without changing its retrieval policy; episode-block intervals quantify uncertainty. QA F1 is the shared open-answer endpoint, and TVQA+'s temporal and region annotations define the served/settled citation-support slots.

Method	QA F1	Served	Settled	Δ [95% CI] (points)	Backlog drop [95% CI]	80%	Answered / cit.-req.
ServeTrace	71.5	81.0	82.8	+1.8 [1.0, 2.7]	5.8 [4.6, 7.0]	pass/pass	2876/2449
Rolling-ANN	70.2	76.9	80.1	+3.2 [2.1, 4.2]	5.4 [4.1, 6.7]	fail/pass	2842/2391
Video-RAG	72.3	78.0	81.4	+3.4 [2.4, 4.5]	5.9 [4.7, 7.2]	fail/pass	2891/2463

Table 5: Independent TVQA+ transfer. Served and settled columns report fixed-output citation correctness (%); backlog drop compares the 0–1 and 5–15 minute bins.

The transfer preserves the fixed-output pattern: every settled score increases by 1.8–3.4 points, with episode-block intervals excluding zero. The blinded secondary pool (+2.7 points over 1,678 citation-required answers), scorer and workload perturbations, and InternVL3-38B (Zhu et al., 2025) transfer retain the direction (Appendix D–F).

5.4 SYSTEMS FEASIBILITY

On the common submitted-query population, ServeTrace reaches 70.8 QA F1, within 0.8 points of the 6.2 GB/h Store-all ceiling, at 2.1 GB/h median storage and 82.1% served-state citation correctness. The complete matched-stack comparison and realized cap sweep are in Appendix Tables 6 and 30.

6 DISCUSSION AND CONCLUSION

Settled materialization changes fixed-output citation correctness by +1.8 to +4.4 points across three retrieval organizations, with every stream-block interval above zero. Across methods, 61/72 to 68/72 streams improve, 4.1–6.7% of labels change at 80%, and fail-to-pass transitions outnumber reversals by 11–14 times. The direction persists on TVQA+ and after replacing the answer backbone with InternVL3-38B.

Snapshot binding makes the correction interpretable: settled scoring may follow a committed root through a provenance-authorized refinement but cannot introduce evidence from later retrieval. It therefore measures post-response credit assigned by evidence materialization, not a changed answer, citation, or retrieval output. Decision-time evaluation should retain and score the evidence snapshot that produced each response.

ETHICS STATEMENT

The study uses established datasets under their original licenses or terms and collects no new meeting or crisis recordings. Annotation is limited to QA verification, event boundaries, and minimal evidence localization. The authors assessed whether the annotation protocol required ethics-board review. Because it asks annotators to evaluate existing benchmark content and collects no personal, behavioral, or sensitive attributes about them, the authors' institutional assessment classified it as non-human-subject technical evaluation; ethics-board review was therefore not required. Source media remain

governed by their original terms, and stored evidence is subject to operator-managed access, redaction, deletion, consent, and retention controls.

Reproducibility statement

Sections 3 and 4 specify the audit record, paired intervention, models, hardware, budgets, aggregation units, and uncertainty procedure. Appendix B provides annotation, parity, stream, and perturbation details, while the remaining appendix tables report the complete comparison and robustness results.

AI use statement

Generative language models were used for manuscript preparation and submission-readiness editing, including revisions to wording and organization throughout the paper and its title, abstract, and introduction. The preparation record does not attribute experimental observations or changes to reported measurements to generative AI. The authors are responsible for the submitted manuscript.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction*, volume 3869, pp. 28–39, 2006. doi: 10.1007/11677482_3. URL https://doi.org/10.1007/11677482_3.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, et al. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020. URL https://arxiv.org/abs/2009.09941.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, pp.

1693–1701, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/afdec7005cc9f14302cd0474fd0f3c96-Abstract.html.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, et al. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL https://www.jmlr.org/papers/v24/23-0037.html.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21278–21298, 2025. doi: 10.18653/v1/2025.findings-acl.1096. URL https://aclanthology.org/2025.findings-acl.1096/.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8211–8225, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.730. URL https://aclanthology.org/2020.acl-main.730/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f593c9c251d4d7cf14d4ab9861dfb7eb-Abstract-Conference.html.

Yu A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473. URL https://doi.org/10.1109/TPAMI.2018.2889473.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022. URL https://arxiv.org/abs/2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Engineering Evaluation

Method	QA F1 mean±std	Storage (GB/h) med [p25,p75]	Citation corr. (%) mean±std
Store-all	71.6 ± 1.1	6.2 [5.4, 7.3]	78.7 ± 1.9
Sliding summary	64.3 ± 1.4	1.1 [0.8, 1.3]	58.4 ± 2.7
Summary-only	66.1 ± 1.3	0.8 [0.6, 1.0]	54.2 ± 3.1
Non-event segmentation	68.7 ± 1.1	2.8 [2.1, 3.4]	73.6 ± 2.1
Tier-less indexing	69.4 ± 1.0	2.4 [1.8, 3.1]	76.8 ± 1.8
ServeTrace	70.8 ± 0.9	2.1 [1.6, 2.8]	82.1 ± 1.6

Table 6: Feasibility under shared models, prompts, and scorer settings. QA F1 uses the common submitted-query population with unanswered queries scored zero. Capped methods share the nominal 2.5 GB/h target and 3 s SLA; Store-all is an uncapped, over-budget ceiling.

The complete operating-point table supports the compact feasibility summary in Sec. 5: ServeTrace combines the strongest served citation score among capped systems with 70.8 QA F1 and 2.1 GB/h median storage.

Table 7 presents a controlled burst setting before the backlog bins for the full workload. Provisional service remains available as the minute-scale tail forms, explaining why coverage stays high while citation localization deteriorates. Figure 3 shows the corresponding backlog and latency distributions.

Ingest rate (fps)	Sketch backlog (s) med [p25,p75]	SLA violations (%)	QA F1 drop (pts)
2	0.7 [0.2, 1.3]	0.4 ± 0.2	0.1 ± 0.1
5	2.1 [0.9, 3.8]	1.3 ± 0.4	0.3 ± 0.2
10	5.2 [2.6, 9.7]	3.8 ± 1.1	0.9 ± 0.3
15	11.4 [5.8, 21.3]	7.2 ± 1.8	1.7 ± 0.5

Table 7: Handling bursts as ingest rate rises. Service stays online through the sketch tier, although more intense bursts increase backlog, SLA violations, and QA loss.

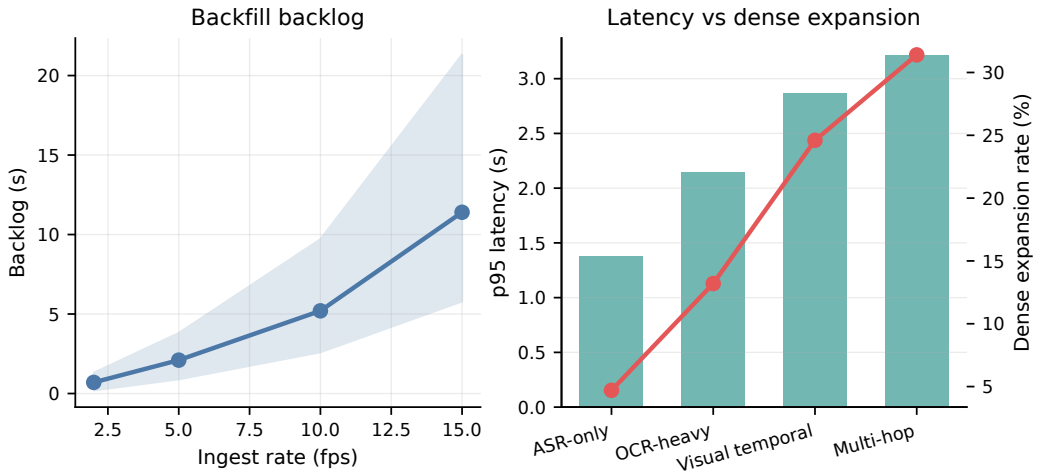

Figure 3: Stress effects on backlog and latency tails. Left: backlog widens with ingest rate. Right: query families are compared by p95 latency, and a descriptive secondary axis overlays the frequency of dense expansion.

Backlog widens superlinearly with ingest rate, and visual-temporal and multi-hop questions occupy the p95 tail because they trigger dense expansion more often. Table 33 and Table 34 separate the query-family and pipeline-stage costs.

A.1 Boundary quality

Boundary detector	F1 mean±std	Over-seg. (%)	Under-seg. (%)	QA F1 impact (Δ)
ASR-only drift	0.712 ± 0.028	18.4	22.7	-1.6
ASR+visual drift	0.763 ± 0.024	14.2	18.1	-0.8
Fixed 5-min windows	0.641 ± 0.031	8.6	31.4	-2.1
ServeTrace (full)	0.806 ± 0.021	11.3	14.7	—

Table 8: Downstream QA and boundary detection quality. Boundary F1 improves with entity churn; the heavy under-segmentation of fixed windows leads to worse QA.

Adding visual drift and entity churn reduces under-segmentation and downstream QA loss relative to ASR-only drift and fixed windows. Accurate boundaries constrain the candidate evidence window, supporting the item localization measured in Tables 22 and 2.

Mixed-query latency is p50 0.92 s, p95 2.58 s, and p99 4.12 s, with 71.3% of queries resolving via ASR/OCR in <1.2 s and 28.7% triggering dense-frame expansion with p95 3.41 s. The aggregate workload stays within the 3 s SLA at p95, with denser visual reasoning driving the tail. Appendix Tables 33 and 34 break that tail down by query family and pipeline stage.

Variant	QA F1 mean±std	p95 latency (s)	Frames/query med [p25,p75]
ServeTrace (lazy expansion)	70.8 ± 0.9	2.58	24 [12, 46]
w/o lazy expansion (always decode)	71.3 ± 0.9	3.84	82 [58, 118]
w/o event boundaries (fixed windows)	68.7 ± 1.1	2.71	28 [14, 52]
w/o citation-aware compression	70.6 ± 0.9	2.52	22 [11, 42]
w/o sketch tier	70.4 ± 1.0	3.12	26 [13, 48]

Table 9: Ablation study. The operating point changes substantially with event boundaries, lazy expansion, and the sketch tier; citation-aware compression has a more modest effect.

Table 9 shows the operational contribution of event alignment, lazy expansion, citation-aware retention, and sketch-tier service. Always decoding gains 0.5 QA points but raises p95 latency by 1.26 s, while fixed windows lose 2.1 QA points and spread evidence across more frames.

A.2 Organization and tier-binding configuration grid

This grid crosses event shards versus fixed windows with whether the answer policy consumes the current tier and resolution state. The audit logger remains active for every row so served/settled scores can be computed. The environmental enrichment schedule, candidate budget, backbone, prompts, storage target, SLA, query schedule, and five seeds are fixed. Because tier-aware serving can change abstention and output content, this is an end-to-end configuration comparison rather than the fixed-output evaluator-state estimand of RQ2.

Organization / tier use	QA F1	Served	Settled	Δ	Backlog drop
Event / used (ServeTrace)	70.8	82.1	83.9	+1.8	6.3
Event / omitted (Tier-less)	69.4	76.8	81.2	+4.4	5.1
Window / used	69.6	79.3	81.7	+2.4	5.8
Window / omitted (Rolling-ANN)	69.8	77.4	80.7	+3.3	5.7

Organization / tier use	Storage (GB/h)	p95 latency (s)	Answered / citation-required
Event / used (ServeTrace)	2.1	2.58	13126/8624
Event / omitted (Tier-less)	2.4	2.49	13016/8624
Window / used	2.3	2.69	12962/8568
Window / omitted (Rolling-ANN)	2.3	2.61	12884/8516

Table 10: Organization-by-tier-use grid. Citation columns are percentages. Query schedules and budgets are shared, while answered and citation-required counts show the output-population changes induced by each end-to-end configuration.

Using current tier state corresponds to higher served citation scores under both organizations: 82.1 versus 76.8 for event shards and 79.3 versus 77.4 for windows. It also corresponds to smaller settlement corrections (1.8 versus 4.4; 2.4 versus 3.3), while event shards retain the strongest served score. The balanced 2×2 summary estimates a 3.6-point binding main effect [2.9, 4.4], a 1.1-point event-organization effect [0.5, 1.8], and a 3.4-point interaction [2.1, 4.7] on served citation correctness. Binding is also associated with a 1.75-point reduction [1.0, 2.5] in the settlement correction. These stream-block intervals support the implementation design; the fixed-output evaluator-state estimand remains RQ2.

B Protocol, Metrics, and Parity Details

The benchmark makes incomplete-backfill behavior visible under substantial QA difficulty. Streams are sampled to preserve domain diversity across news, meetings, and crisis scenarios; questions are admitted only when the answer is supported by in-stream audio-visual evidence; and a citation-required subset is maintained so the paper can evaluate both answer quality and localization quality under the same served tier. Gold event boundaries are annotated on the stratified 30-stream subset to preserve the domain proportions of the full 72-stream pool while keeping manual adjudication tractable.

ID	Goal	Setting	Primary metrics
E1	QA quality vs storage	long streams (1–6 h)	F1/acc, GB/h
E2	Real-time ingest stability	ingest rate sweep (2–15 fps)	max fps, backlog, SLA violations
E3	Query latency	mixed query types	p50/p95/p99 latency, decode cost
E4	Evidence quality	citation-required subset	citation correctness, missing evidence
E5	Boundary detection	labeled boundary subset	boundary F1, downstream QA impact
E6	Burst tier benefits	sketch tier on/off	SLA, quality under bursts
E7	Retention ablations	fixed TTL vs adaptive	storage vs accuracy vs citation corr.
E8	Robustness	OCR/ASR noise, overload	graceful degradation

Table 11: Appendix experiment matrix.

Table 11 maps every appendix result to a prespecified question about quality, latency, evidence localization, or robustness.

Table 12 fixes the aggregation unit and operating target for QA, coverage, citation quality, staleness, latency, and storage, keeping the outcomes separable in the backlog analysis.

Annotation and scoring are separated. In the first pass, annotators see the question, reference answer, and source stream but no method output. They define the required support slots and enumerate source-native alternatives within each slot. In the second pass, each method's retrieval unit is shown without its method identity; annotators validate only ambiguous unit-to-primitive and provenance links. A third adjudicator resolves disagreements with duplicate-window, boundary, sketch-insufficiency, or OCR/ASR-omission reason codes. Independent boundary marks within 5 s are merged before adjudication so near-coincident transitions do not become separate primitives.

Both ledgers are then frozen. The matcher in Sec. 3 computes exact, over-broad, shifted, missing, and wrong-item outcomes without further human labels. Table 6 aggregates them into per-answer citation F1, while Table 22 exposes item-level precision and recall. A representative equivalence case is a speaker continuing one claim while the video alternates between a podium shot and a slide montage: either the stable speaker frame or an aligned slide crop can satisfy the visual-identity slot, while an unrelated neighboring montage is a wrong-item outcome. For example, if the answer is "the evacuation order was issued by the county emergency office," separate verbal-identity and visual-identity slots require a short ASR span naming the office and an aligned lower-third crop. A 30 s clip of the briefing is over-broad even when it contains the same words. A handle rooted in a different briefing segment is wrong-item; a committed root whose crop has not materialized is missing. These deterministic distinctions make served-tier localization loss measurable under backlog.

A common qualitative pattern is that the system still returns a non-abstained response while the citation becomes less auditable under backlog. For a question such as "Which office issued the evacuation order?," sketch-tier service may return a response supported only by coarse ASR and a broad keyframe window, so the citation points to an over-broad time span or an unresolved visual region. Full backfill can materialize the short ASR span and lower-third crop through that handle's

Metric	Definition	Aggregation	SLA / Budget	Exp. IDs
QA F1	Token overlap F1 vs gold answer; unanswered queries score zero unless labeled answered-only	Submitted-query mean±std over 5 seeds; Table 20 conditions on answered queries	—	E1, E6, E7, E8
Coverage	Fraction of submitted queries returning an answer	Mean±std; per-stream	reported with QA/cit.	E2, E3, E6, E7
Policy abstention	Fraction receiving an explicit no-answer decision	Mean±std; per-stream	policy outcome	E2, E6
Service failure	Fraction ending in timeout or execution error	Mean±std; per-stream	SLA outcome	E2, E3, E6
Staleness	Wall-clock age of oldest cited evidence item	Median [p25,p75]; per-query	$\leq$5 min	E1, E3, E6
Rare-event retention	Fraction of rare-event tagged shards retained	Mean over rare-event subset	$\geq$85%	E7, E8
SLA pass rate	Fraction of queries with latency $\leq$3 s	Mean±std; per-stream	target $\geq$95%	E2, E3, E6
Citation correctness	Prec/Recall of cited evidence items	Per-answer scores; stream-macro aggregate or bin-micro slice	—	E4, E7
Cold-start success	QA F1 within first 60 s of stream	Mean±std; per-stream	$\geq$60% F1	E1, E6
Storage growth	Total indexed data (GB/hour)	Median [p25,p75]; per-stream	$\leq$2.5 GB/h	E1, E7

Table 12: Metric definitions, aggregation units, and budget targets.

committed root; settled-state scoring then verifies the frozen handle without substituting an unrelated neighboring item. This is the concrete protocol trace that Table 2 aggregates.

We evaluate robustness with a controlled drift generator. Topic drift occurs at $t \in \{30, 90, 150\}$ min, modality dropout at $t \in \{60, 120\}$ min with WER+15 / CER+20 for 10 min, and burst spikes at $t = 45$ min with ingest rising from 3 fps to 12 fps for 15 min. Evaluation windows are labeled *none*, *mild*, or *severe*; 58.4% of windows are none, 27.1% mild, and 14.5% severe. Cold-start begins with empty memory, queries start at $t = 15$ s, and the first 60 s are scored while full backfill is still incomplete.

All methods share identical VLM/ASR/OCR versions, prompts, ingest events, asynchronous evidence-revision journal, and evaluation harness. They differ in memory organization, retention, retrieval refresh, and whether the answer policy consumes served-tier state. Matched-budget comparisons among capped methods enforce identical GB/h caps and the same 3 s latency SLA; Store-all is the uncapped ceiling. Table 13 holds models, prompts, retrieval parameters, caching, the enrichment

Parity dimension	Store-all	Sliding sum.	Non-event	Tier-less	ServeTrace
Index type / params	HNSW M=32	HNSW M=32	HNSW M=32	HNSW M=32	HNSW M=32
Embedding model	BGE-M3+CLIP	BGE-M3+CLIP	BGE-M3+CLIP	BGE-M3+CLIP	BGE-M3+CLIP
Candidate pool	all shards	all windows	all windows	all shards	all shards
VLM / prompts	identical	identical	identical	identical	identical
Caching policy	LRU 2 GB	LRU 2 GB	LRU 2 GB	LRU 2 GB	LRU 2 GB
Served-tier semantics	full-only	summary-only	unmarked current	unmarked current	bound sketch/full
Storage cap	uncapped	matched	matched	matched	matched
Latency SLA	3 s cap	3 s cap	3 s cap	3 s cap	3 s cap

Table 13: Implementation factors held constant across baselines.

schedule, and latency constant, and holds the storage target constant among capped methods; Store-all is the stated uncapped ceiling. The row on served-tier semantics records what each answer policy can inspect; all rows retain the common revision journal used by the paired scorer.

The evaluation harness consumes only answer text, returned citations, coverage, and, when applicable, a served-tier tag. It does not inspect boundary scores, shard utility values, retention decisions, or any other internal ServeTrace state while scoring baselines. For non-tiered baselines, the served tier is simply the evidence they actually return at query time. The interface therefore applies uniformly to tiered and non-tiered retrieval systems.

Stream type	Duration	Ingest fps	Query rate	ASR/OCR/ Visual mix
News/meetings	1–6 h	2–5	1/2–1/5 min	48/31/21
Crisis events	1–3 h	5–12	1/1–1/3 min	36/24/40

Table 14: Stream durations, ingest rates, and query mix.

Table 14 keeps queries arriving during ingest, making backlog-conditioned evidence states observable. Query rates are one every 2–5 minutes for news/meetings and one every 1–3 minutes for crisis streams.

C Reproducibility and Asset Notes

All primary experiments use seeds $\{13, 37, 61, 71, 97\}$, CUDA 12.4, PyTorch 2.3.1, and NVIDIA driver 550.54.15 on the six-A100 configuration in Sec. 4. Splits are fixed before evaluation; queries from one stream or TVQA+ episode remain in the same split and bootstrap block. The experiment matrix, metric definitions, parity controls, stream schedule, annotation rubric, and scorer perturbations appear in Tables 11–13 and 19. Table 15 records the licenses and handling boundary for each data and model family.

The five-seed operating point, storage sweep, burst study, boundary study, and six perturbation conditions all use two A100 80 GB GPUs for ingest and four for QA inference. TVQA+ uses the same footprint and software stack. Source identifiers and timestamped derived annotations preserve

Asset family	License / terms boundary and handling notes
CNN/DailyMail news videos	Original publisher terms of service. The study records source URLs, ids, timestamps, QA annotations, and derived evaluation metadata.
AMI Meeting Corpus	CC BY 4.0 corpus license. Session ids and attribution links are preserved in the manifests; redistributed derivatives remain tied to the corpus terms.
YFCC100M crisis-event media	Per-item Creative Commons licenses in the Flickr/YFCC metadata. The study records ids, timestamps, and derived annotations; downstream reuse follows each item's attached license.
TVQA+	Official research-release and source-media terms. Episode ids, questions, temporal annotations, and derived citations follow the dataset terms.
Qwen2.5-VL-32B-Instruct / Qwen2.5-32B-Instruct	Apache 2.0 model-card licenses; configs pin upstream model identifiers and fetch weights from the provider.
InternVL3-38B	MIT-licensed code and provider model-card terms for weights; configs pin the OpenGVLab model identifier.
Video-RAG	Upstream implementation and model-card terms; the evaluated revision and retrieval configuration are version-pinned.
Whisper-large-v3 / CLIP ViT-L/14 / BGE-M3	MIT upstream licenses; version-pinned configs reference upstream repositories or model cards.
PaddleOCR v2.7.3	Apache 2.0 upstream license; the OCR component is version-pinned in configs and fetched from upstream.
Derived manifests, adjudication exports, and evaluation tables	Project-derived split manifests, gold citations, reason codes, and figure-ready aggregates governed by the corresponding source-data terms.

Table 15: License and handling boundaries for media, models, and derived metadata.

evaluation traceability while raw publisher media and model weights remain governed by their original terms.

D Evidence-State Sensitivity Package

All analyses in this section keep the same citation-required subset, returned answers and citations, five seeds, and Qwen2.5-VL / Whisper / PaddleOCR / CLIP / BGE-M3 backbone stack from Sec. 4 unless explicitly noted otherwise. They measure within-output corrections, threshold sensitivity, stream-aware uncertainty in the primary deltas, and the relation between one illustrative reporting line and alternative service policies.

The stream-level distribution supplies a population-stability check without comparing method-conditional ranks: 61/72 ServeTrace streams, 66/72 Rolling-ANN streams, and 68/72 Tier-less streams have positive settlement corrections. The dominant fail-to-pass counts and positive paired medians show the same fixed-output shift at answer level.

Figure 2 makes the population structure visible rather than redrawing score-minus-threshold lines. Median per-stream corrections increase from 1.5 points for ServeTrace to 3.0 for Rolling-ANN and 4.2 for Tier-less indexing; their interquartile ranges remain positive, and 61, 66, and 68 of 72 streams, respectively, move upward. Table 16 separately maps these continuous shifts to all five reporting cutoffs.

Stream-block resampling keeps the citation and QA backlog-stratum changes positive and, critically, every paired served-to-settled correction above zero.

Table 17 shows positive low-to-high backlog changes for the separately conditioned citation and QA outcomes and positive paired corrections for all three methods. Table 18 places the 80% sensitivity line beside explicit service semantics. Strict mode improves citation correctness through greater abstention, best-effort mode raises coverage through relaxed citation requirements, and tier-aware mode lies between them.

Method	Pass/pass	Fail/fail	Fail→pass	Pass→fail	Increase / tie / decrease
ServeTrace	6894	1378	329	23	3821/4142/661
Rolling-ANN	6421	1648	412	35	4510/3167/839
Tier-less indexing	6293	1754	528	49	4986/2674/964

Method	Changed label [95% CI]	Median paired Δ [95% CI]	75–85 integrated [95% CI]
ServeTrace	4.1 [3.5, 4.7]	1.4 [1.1, 1.7]	3.6 [3.0, 4.2]
Rolling-ANN	5.2 [4.5, 6.0]	2.7 [2.3, 3.1]	4.8 [4.1, 5.6]
Tier-less indexing	6.7 [5.9, 7.5]	3.8 [3.3, 4.3]	6.1 [5.3, 7.0]

Method	75	77.5	80	82.5	85
ServeTrace	P/P	P/P	P/P	F/P	F/F
Rolling-ANN	P/P	F/P	F/P	F/F	F/F
Tier-less indexing	P/P	F/P	F/P	F/F	F/F

Method	Per-stream median Δ [p25,p75]	Streams with positive Δ
ServeTrace	1.5 [0.8, 2.4]	61/72
Rolling-ANN	3.0 [1.9, 4.1]	66/72
Tier-less indexing	4.2 [2.9, 5.5]	68/72

Table 16: Paired label transitions at 80%, threshold-integrated summaries, all five prespecified reporting cutoffs, and per-stream score-change distributions. P/F pairs denote served/settled outcomes. For integer counts, each unique query is averaged across its five replay realizations and thresholded once; counts therefore sum to the displayed citation-required query denominator.

Quantity	Point estimate (pts)	95% CI
Citation-correctness drop, 0–1 → 5–15 min backlog	6.3	[5.4, 7.1]
QA F1 drop, 0–1 → 5–15 min backlog	2.9	[2.1, 3.8]
Citation-drop minus QA-drop gap	3.4	[2.4, 4.6]
ServeTrace served → settled delta	+1.8	[0.9, 2.7]
Rolling-ANN served → settled delta	+3.3	[2.2, 4.4]
Tier-less served → settled delta	+4.4	[3.2, 5.5]

Table 17: Uncertainty for primary deltas. Confidence intervals use a stream-block bootstrap that resamples the 72 streams with replacement while preserving the reported backlog bins and scorer interface.

Policy	ServeTrace cit. corr. (%)	ServeTrace coverage (%)	ServeTrace abstain (%)	Tier-less cit. corr. (%)	Tier-less coverage (%)	Tier-less abstain (%)
Strict	84.6	86.2	9.4	79.4	84.8	10.7
Tier-aware	82.1	91.3	4.2	76.8	90.5	5.0
Best-effort	78.7	95.1	1.1	73.2	94.4	1.7

Table 18: Policy context for the 80% reporting line. Strict improves citation correctness by abstaining more, best-effort raises coverage by relaxing citation discipline, and tier-aware mode balances both. Coverage, explicit abstention, and residual timeout/error partition all submitted queries; the residual rate equals 100 − coverage − abstention.

Tol.	Merge	Agg.	ServeTrace (%)	Tier-less (%)	Gap (pts)	Backlog penalty
±1 s	on	maj.	79.6	73.5	6.1	6.8
±2 s	on	maj.	82.1	76.8	5.3	6.3
±3 s	on	maj.	83.5	78.2	5.3	5.9
±2 s	off	maj.	81.2	75.5	5.7	6.4
±2 s	on	uni.	80.8	75.0	5.8	6.1

Table 19: Sensitivity of served-tier citation scoring to tolerance, duplicate-window merging, and the acceptance rule for human-validated mapping links. Absolute values move, but the comparative direction and backlog-conditioned penalty remain stable.

Table 19 perturbs one scorer choice at a time while keeping the returned outputs and workload fixed. The tighter ±1 s tolerance predictably lowers absolute citation scores, while the looser ±3 s tolerance raises them; duplicate-window splitting and unanimous acceptance of ambiguous mapping links act as milder downward shifts. Across all settings, ServeTrace remains ahead of Tier-less indexing and the low-backlog-to-high-backlog penalty stays positive, demonstrating stability to scorer tolerance, duplicate merging, and aggregation. The movement is therefore systematic across scorer choices: ServeTrace ranges from 79.6 to 83.5, Tier-less indexing ranges from 73.5 to 78.2, and the cross-method gap retains its sign. This scorer-level stability complements the paired correction intervals in Table 17 and the cross-dataset and cross-backbone evidence in Tables 5 and 28.

E Comparator and Validation Corroboration

We test the evaluator-state effect beyond the ServeTrace/Tier-less pair through same-bin answer quality, raw-count support, nearest-comparator localization, a matched-budget external comparator, workload slices, and a held-out adjudication pool. Unless explicitly noted otherwise, these analyses keep the same 72-stream benchmark, the same five seeds, the same Qwen2.5-VL/Qwen2.5-32B/Whisper/PaddleOCR/CLIP/BGE-M3 stack, the same nominal 2.5 GB/h target, the same 3 s SLA, and the same served-tier scorer. Together they test the decision-time audit object across answer quality, support size, localization, comparator architecture, workload composition, and adjudication.

Backlog (min)	QA F1 mean±std	Queries	Answered	Sketch-only answered
0–1	72.0 ± 0.6	4931	4620	384
1–5	70.7 ± 0.8	5286	4831	899
5–15	69.1 ± 1.1	3017	2661	862
High–low Δ	-2.9	—	—	—

Table 20: Backlog-aligned QA F1 by served-tier backlog. The table uses the same backlog bins as Table 2, the same tier-aware policy, the same five seeds, the same workload split, and the same 6×A100 hardware stack from Sec. 4. Uncertainty is run-to-run standard deviation.

On the same backlog partition, QA F1 decreases by 2.9 points while citation correctness decreases by 6.3 points. The interval estimate in Table 17 confirms that this difference persists under stream-aware resampling.

Backlog (min)	Citation-required answers	Contributing streams
0–1	3226	54
1–5	3411	61
5–15	1987	43

Table 21: Backlog-bin support counts for the served-tier citation analysis. These counts align with Table 2 and use the same 0–1 / 1–5 / 5–15 minute bins as Table 20.

Each backlog bin contains 1,987–3,411 citation-required answers from 43–61 streams. These denominators complement the effect estimates in Tables 2 and 17.

Method	Prec.	Recall	κ
Store-all	76.4	73.1	0.72
Sliding summary	52.8	48.6	0.68
Non-event seg.	71.3	68.7	0.73
Tier-less indexing	75.2	71.8	0.74
ServeTrace	80.7	77.4	0.76

Table 22: Item-level citation precision and recall, with inter-annotator agreement for ambiguous retrieval-unit-to-primitive mapping decisions on the citation-required subset.

The item-level view preserves the same ordering as per-answer citation F1 and directly includes the nearest tier-semantics comparator. Table 23 decomposes ServeTrace and Tier-less outcomes into exact, over-broad, shifted, and unverifiable citation items; the last category combines missing roots and active wrong items.

Method	Prec.	Recall	κ	Answers	Cited items
ServeTrace	80.7	77.4	0.76	8624	10836
Tier-less indexing	75.2	71.8	0.74	8624	10792

Method	Exact	Over-broad	Shifted	Unverifiable (missing/wrong)
ServeTrace	8745	942	636	513
Tier-less indexing	8114	1217	896	565

Table 23: Item-level citation decomposition on the citation-required subset. The final column aggregates missing and wrong-item false positives under the deterministic hierarchy in Sec. 3. Raw-count columns sum to returned cited items; recall uses the frozen gold-slot denominator from Table 22.

Tier-less indexing produces 275 more over-broad and 260 more shifted items than ServeTrace. Its committed handles therefore expose more version-dependent localization detail under settled materialization, which is consistent with its larger correction.

Baseline	Retrieval unit / refresh	QA F1	Cit. corr.	Served→settled
Rolling-ANN	30 s windows / 15 s hop / 2 s refresh	69.8 ± 1.1	77.4 ± 1.7	+3.3

Baseline	Storage median [p25,p75]	Backlog drop	Answered / citation-required
Rolling-ANN	2.3 [1.9, 2.9]	5.7	12884/8516

Table 24: External matched-budget baseline under shared served-tier scoring. The rolling-ANN comparator uses fixed 30 s windows, 15 s hop, continuously updated HNSW/FAISS retrieval, no event boundaries, no sketch/full served-tier tags, the same 72-stream benchmark, the same five seeds, the same backbones, the same nominal 2.5 GB/h target, and the same 3 s SLA as the main paper.

The fixed-window rolling index remains in the same QA/storage regime as the event-sharded systems while exhibiting a 5.7-point backlog loss and a +3.3-point settled correction.

ServeTrace leads Tier-less indexing in both workload families, and each slice retains a positive backlog penalty. Together with Table 19, this establishes stability across scorer perturbations and workload composition.

The secondary pool contains 2,143 answered queries, including 1,678 that require citations. It reproduces both directional measurements: citation correctness differs by 5.2 points between observed

Workload slice	ServeTrace cit. corr. (%)	Tier-less cit. corr. (%)	Gap (pts)	Backlog penalty
News/meetings	82.6	77.2	5.4	5.8
Crisis	80.9	74.5	6.4	7.2

Table 25: Cross-workload scorer robustness on the paper's two stream families. The same served-tier scorer and adjudication pool are applied after slicing the citation-required subset by workload family.

Validation unit	Served-tier cit. corr. (%)	Backlog drop (pts)	Served→settled Δ (pts)	Answered	Cit.-req. n
Held-out adjudication pool 2	80.6	5.2	+2.7	2143	1678

Table 26: Secondary evaluator-state check. The record-disjoint pool is stratified by source family, backlog bin, and required modality; method identities and primary-pool outcomes remain hidden during annotation.

backlog endpoints and rises descriptively by 2.7 points under paired settled replay. The primary uncertainty claims remain those in Table 17.

F Independent and Cross-Backbone Transfer

F.1 Complete TVQA+ transfer results

The conversion preserves the official validation episodes and annotations while exposing their source elements incrementally. A citation adapter maps frame ids, subtitle spans, and object-region ids to primitive roots; unresolved mappings remain in the denominator. Official temporal moments and object regions instantiate the gold support slots, while the table's QA F1 is the shared open-answer metric used by all three replayed systems.

Method	QA F1	Coverage	Served	Settled	Citation backlog drop	QA backlog drop
ServeTrace	71.5	95.3	81.0	82.8	5.8	2.6
Rolling-ANN	70.2	94.2	76.9	80.1	5.4	2.8
Video-RAG	72.3	95.8	78.0	81.4	5.9	3.0

Method	Descriptive citation–QA drop difference [95% CI]	Storage (GB/h)	p95 latency (s)	Answered / citation-required
ServeTrace	3.2 [2.0, 4.4]	2.2	2.66	2876/2449
Rolling-ANN	2.6 [1.5, 3.8]	2.4	2.54	2842/2391
Video-RAG	2.9 [1.7, 4.1]	2.5	2.91	2891/2463

Method	77.5	80	82.5
ServeTrace	pass/pass	pass/pass	fail/pass
Rolling-ANN	fail/pass	fail/pass	fail/fail
Video-RAG	pass/pass	fail/pass	fail/fail

Table 27: Complete TVQA+ transfer package. The citation–QA drop difference contrasts separately conditioned descriptive endpoints; paired served/settled scores are the evaluator-state estimand. All results use the official validation split and five-seed episode-block protocol.

All 3 methods change their own label on the transfer grid. More importantly, each method has a positive paired settlement correction, including the released Video-RAG implementation.

F.2 Cross-backbone evaluator-state transfer

We replace Qwen2.5-VL-32B-Instruct with InternVL3-38B while holding ASR/OCR, retrieval, prompts, budgets, policy, splits, scorer, and seeds fixed.

Backbone / method	QA F1	Served	Settled	Backlog drop	80%	Answered / citation-required
Qwen2.5-VL / ServeTrace	70.8	82.1	83.9	6.3	pass/pass	13126/8624
Qwen2.5-VL / Rolling-ANN	69.8	77.4	80.7	5.7	fail/pass	12884/8516
Qwen2.5-VL / Tier-less	69.4	76.8	81.2	5.1	fail/pass	13016/8624
InternVL3 / ServeTrace	71.4	81.5	83.5	6.0	pass/pass	13138/8736
InternVL3 / Rolling-ANN	70.1	77.0	80.5	5.9	fail/pass	12942/8532
InternVL3 / Tier-less	69.9	76.2	80.8	5.4	fail/pass	12987/8611

Table 28: Cross-backbone transfer on the original 72-stream protocol. The scorer-state direction persists after changing the answer-generation VLM family.

The InternVL3 served-to-settled deltas are +2.0 points for ServeTrace (95% CI [1.1, 2.9]), +3.5 [2.4, 4.6] for Rolling-ANN, and +4.6 [3.4, 5.8] for Tier-less indexing. Changing the VLM moves absolute QA and citation scores while preserving a positive settlement correction for every retrieval organization.

G Auditability Evidence Matrix

The matrix makes the enforcement boundary inspectable: the scorer trusts the logged snapshot and evaluates only the committed answer, citation items, backlog, tier, and policy state. Its evidence references point directly to the corresponding manuscript result.

H Additional Operating-Point and Policy Details

The storage sweep places the 2.5 GB/h target at the knee of the quality–cost curve: tighter caps reduce citation completeness, whereas larger budgets add little QA or citation benefit. The 2.5 GB/h value is the controller target used for the default operating point, whereas Table 6 reports realized per-stream storage as median [p25,p75]; the upper quartile can therefore exceed the target without implying that the policy was configured at a different nominal cap.

I Backfill Policy and Deployment Notes

During bursts, the sketch tier may answer queries before full backfill completes. ServeTrace defines explicit semantics: answers indicate whether evidence is *complete* or *sketch-only*, and citation correctness is evaluated under the evidence tier actually served. Abstentions count as QA failures, but coverage is reported separately so quality is not inflated by simply refusing to answer. Table 31 makes abstention, sketch-tier service, and citation requirements explicit, allowing strict, tier-aware, and best-effort outcomes to be compared without policy drift.

The managed deployment boundary supplies authentication, encryption at rest, redaction, deletion, consent tracking, and access control, exposing stored shards only to the service operator and autho-

Protected object	Capability and trust	Property and enforcement boundary
Answer + citations	Backlog delays OCR/crop enrichment; immutable answer-time records and honest tier ids are trusted.	Minimal-item auditability on the served snapshot; the scorer consumes only Sec. 3 record fields.
Evidence snapshot	The version journal records refinements and retractions; immutable snapshot binding is trusted.	Answer-time evidence remains reconstructable; binding precedes asynchronous enrichment.
Retrieval stack	Dataset and retrieval policy change; the frozen TVQA+ split and exposed Video-RAG spans are trusted.	The scorer-state effect transfers across dataset and implementation boundaries.
Stored shards	Burst load and rare-event pressure; authorized operator-controlled storage is trusted.	Citation-preserving retention under storage/SLA controls; the controller precedes full backfill.

Protected object	Comparator and held-out setting	Utility, residual error, and evidence
Answer + citations	Tier-less, Rolling-ANN, blinded adjudication, and fixed-output rescoring.	2.58 s p95 and 2.1 GB/h; settled support can mask a served failure. Tables 2, 3, and 16.
Evidence snapshot	Event/window and tier-used/omitted configurations.	Tier-aware rows have higher served scores and smaller corrections in both organizations; the grid is descriptive because output populations differ. Table 10.
Retrieval stack	Released Video-RAG, Rolling-ANN, and 431 held-out TVQA+ episodes.	2.2–2.5 GB/h and 2.54–2.91 s p95; broad evidence spans account for the dominant residual error. Tables 5 and 27.
Stored shards	Store-all and fixed windows; ingest 2/5/10/15 fps and QPS 50/200/1k.	93.2–96.4% cap pass rate; coarse sketch evidence and over-retention account for the residual errors. Tables 7, 30, and 35.

Table 29: Auditability evidence matrix linking each protected object to its capability model, trust boundary, comparator, transfer setting, operational cost, residual error, and paper-internal evidence.

Storage cap (GB/h)	QA F1 mean±std	Citation corr. (%) mean±std	Missing evidence (%) mean±std
1.5	69.6 ± 1.0	80.3 ± 1.7	6.7 ± 1.2
2.0	70.4 ± 0.9	81.6 ± 1.6	5.1 ± 1.0
2.5	70.8 ± 0.9	82.1 ± 1.6	4.4 ± 0.9
3.5	71.1 ± 0.9	82.3 ± 1.5	3.9 ± 0.8

Table 30: Effect of storage budgets on QA and citation metrics.

Policy	Allowed when incomplete	Citation requirement	Trade-off
Strict	no (abstain)	full citations only	best correctness, lower coverage
Tier-aware	yes	sketch citations allowed	higher coverage, tier-dependent quality
Best-effort	yes	optional citations	highest coverage, weakest citation guarantees

Table 31: Policies for answering when full evidence is incomplete.

rized downstream clients. Warm starts reuse ASR/OCR outputs and keyframe embeddings within a stream, while hot-shard caching reduces repeated query-time expansion. When ingest falls behind, the system can either drop backfill to stabilize throughput or abstain when sketch-only evidence cannot satisfy citation requirements; the robustness table reports both trade-offs. ServeTrace's 2.1 GB/h median comprises 0.62 GB/h keyframe embeddings, 0.48 GB/h salient crops, 0.31 GB/h OCR entity indices, 0.27 GB/h ASR span tokens, 0.14 GB/h citation indices, and 0.28 GB/h metadata.

J Secondary Robustness and Drift Results

Condition	QA F1 drop (pts)	Storage change	Citation corr. drop (pts)
OCR noise (blur/low-res, text corruption)	-1.4 ± 0.4	$+8.3\%$	-2.8 ± 0.7
ASR noise (WER+10)	-1.1 ± 0.3	$+4.6\%$	-1.9 ± 0.5
Long stream ($\times 2$ duration)	-0.6 ± 0.2	$+3.2\%$	-0.8 ± 0.3
Long stream ($\times 4$ duration)	-1.3 ± 0.4	$+7.1\%$	-1.6 ± 0.5
Overload (drop backfill)	-0.9 ± 0.3	-12.4%	-3.4 ± 0.9
Overload (abstain on sketch)	-0.2 ± 0.1	$+0.0\%$	-0.4 ± 0.2

Table 32: Controlled perturbations and their effect on QA, storage, and citations.

Across noisy ASR, OCR blur, longer streams, and burst overload, citation quality degrades more sharply than QA F1. The controlled perturbations therefore reproduce the auditability gap through distinct stress paths. The OCR-noise row names the logged perturbation regime rather than a single isolated corruption channel: blurred and low-resolution frames are paired with OCR-text corruption so the effective stress test matches end-to-end OCR failure without reusing the ASR-oriented WER label.

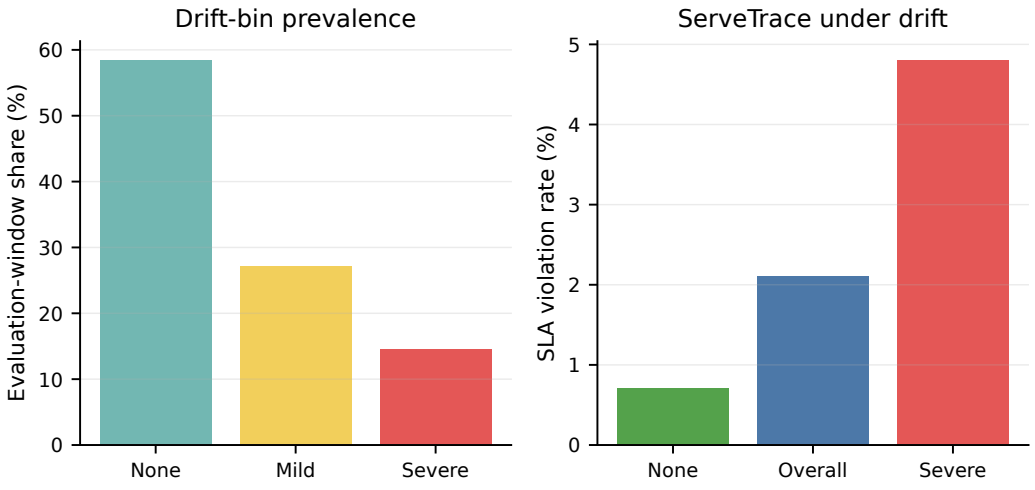

Figure 4: Drift prevalence and SLA failures by severity bin.

Figure 4 separates drift prevalence from its SLA impact: severe windows are uncommon but account for a disproportionate share of service failures.

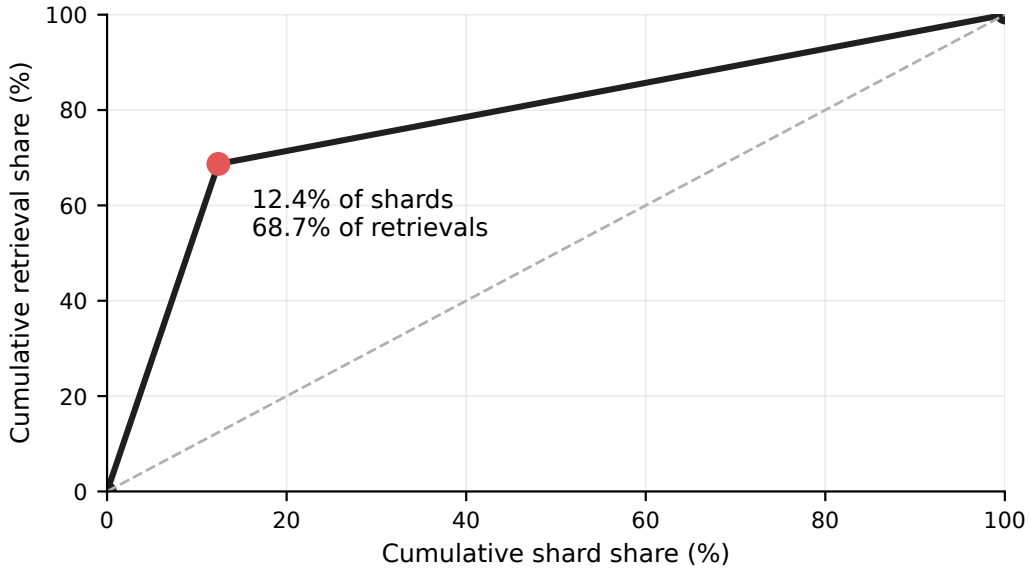

Figure 5: Cumulative retrieval concentration over memory shards.

Figure 5 shows heavy-tailed retrieval demand: frequent shards dominate accesses while rare events still require protection. This concentration motivates citation-aware retention under a fixed storage budget.

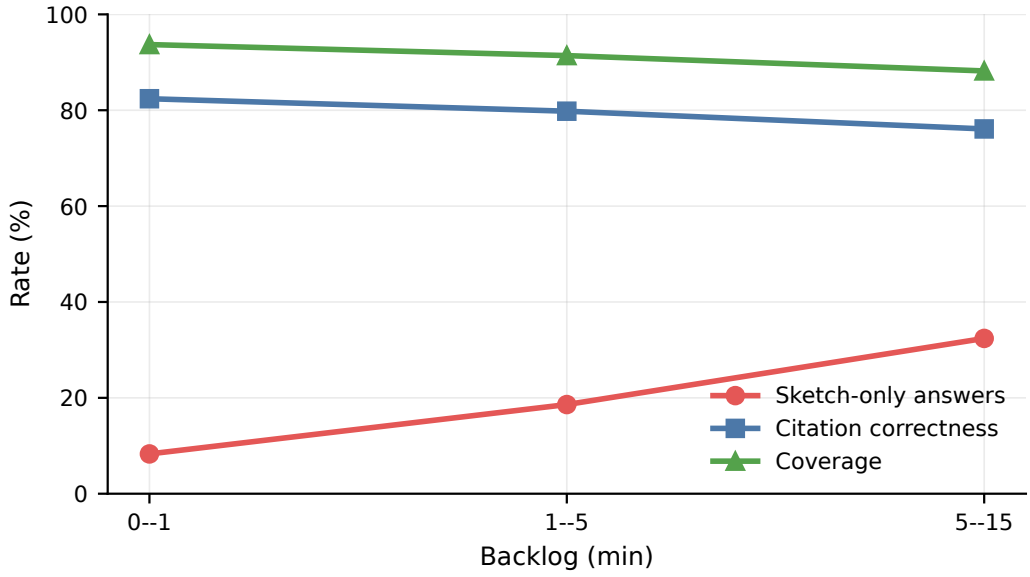

Figure 6: Backlog bins versus sketch-only service, citation correctness, and coverage.

Figure 6 visualizes the separation in Table 2: sketch-only service rises with backlog, citation correctness falls, and coverage remains comparatively high.

Residual errors were dominated by (i) boundary misses that merged distinct events (38.1% of errors), (ii) OCR misses that suppressed correct evidence (27.4%), and (iii) backlog-induced incompleteness where sketch evidence was insufficient (19.2%). The remaining 15.3% were VLM reasoning failures with correct evidence.

K DEPLOYMENT-TIME RUNTIME DETAILS

Query type	F1 mean±std	p95 latency (s)	Dense expansion rate (%)
ASR-only (spoken)	74.2 ± 0.8	1.38	4.7 [2, 8]
OCR-heavy	69.6 ± 1.1	2.14	13.2 [7, 19]
Visual temporal	65.3 ± 1.4	2.87	24.6 [16, 34]
Multi-hop	61.8 ± 1.7	3.21	31.4 [22, 41]

Table 33: Per-type latency breakdown. Spoken-only questions are both cheaper and more accurate, while visual-temporal and multi-hop questions dominate the latency tail.

Spoken-only queries remain on the inexpensive path, while visual-temporal and multi-hop queries trigger dense expansion and dominate the latency tail. Table 34 attributes that tail to its pipeline stages.

Component	Cost p50 (ms)	Cost p95 (ms)	Notes
Boundary detection (per frame)	1.8	4.2	CLIP + ASR drift; batched
Sketch shard creation	12.4	31.7	embedding + entity extraction
OCR refinement (backfill)	38.6	142.3	PaddleOCR; async
Keyframe selection + crop	8.3	24.1	CLIP-based salient frame scoring
Shard query (cheap path)	22.7	68.4	embedding + entity overlap
Dense expansion (frames)	164.2	743.8	drives tail latency; 28.7% of queries
Rerank + generate answer	48.3	186.4	Qwen2.5-VL inference

Table 34: Measured p50/p95 stage costs on the 6×A100 setup.

Table 34 attributes the p95 tail primarily to dense expansion rather than ingest bookkeeping. Stage costs overlap in the asynchronous pipeline and are reported as component measurements. These component p50/p95 values measure overlapping stages: cheap-path retrieval, dense expansion, and generation execute concurrently, and 28.7% of queries enter the dense-expansion branch. Together with the query-family results in Table 33 and Figure 3, they localize the end-to-end tail without summing overlapping costs.

Background retrieval QPS	Ingest CPU util. (%)	Foreground query p95 latency (s)	Cap pass rate (%)
50	34.2	2.21	96.4 ± 0.8
200	47.8	2.64	95.1 ± 1.0
1k	68.3	3.18	93.2 ± 1.4

Table 35: Load sensitivity under background retrieval/control-plane traffic while answer generation follows the fixed benchmark query schedule.

As background candidate-search, policy-check, and snapshot-lookup traffic rises from 50 to 1,000 QPS, foreground p95 answer latency increases from 2.21 to 3.18 s and cap pass rate decreases from 96.4% to 93.2%. Qwen generation remains at the benchmark query schedule; the sweep uses asynchronous batching with up to 256 concurrent retrieval requests, the same lazy-expansion policy, 2 GB hot-shard cache, and four-GPU QA pool.