CAPSIEVE: COUPLED MODALITY, EXTENT, AND STOPPING UNDER MULTIMEDIA SERVING BUDGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

A multimedia request can begin with a cheap transcript, expand into dense frames, approach its decode and context limits, and still cross the wall-clock deadline despite every acquisition being plan-feasible when selected; stopping earlier can instead force abstention or an unsupported answer. CapSieve controls these competing exits with one masked policy over modality–extent–stop tuples and a fail-closed answerability gate under latency, decoded-media, and token budgets. An executed-route ledger links every answer or abstention to native task quality, realized SLA compliance, acquisition work, and claim support. Across DocVQA, ChartQA, ActivityNet-QA, and AudioCaps, CapSieve scores 71.2 ± 0.8 at the 4 s tier versus 72.3 ± 0.9 for uncapped fetch-all, while reducing median normalized decode work from 1.00 to 0.39 and passing the SLA on 95.2% of the fixed mixture. It exceeds a matched constrained actor-critic by 0.6 points (95% CI [0.2, 1.0]), improves automatic evidence alignment by 3.5 points, and uses 0.04 less median decode; reward-only training reaches 70.8. A blinded 4 s audit independently yields 86.2% citation-support precision and 79.1% recall, compared with 80.3% and 76.4% for the actor-critic. Route accounting then locates the shared serving pressure: visual-temporal and multi-hop requests occupy 33.1% of the mixture but carry 59.7% of decode work, 66.7% of abstentions, 74.5% of SLA failures, and 74.2% of evidence contradictions—up to $2.25\times$ their prevalence. The pattern persists across cap tiers, policies, and a second VLM backbone, turning executed routes into a concrete guide for where additional serving and modeling capacity has the highest leverage.

1 INTRODUCTION

A transcript-only state leaves most of a request's latency, decode, and token envelope intact. Expanding that route to frames can consume decode headroom; composing separated visual and audio evidence can fill the context; generation may then cross the wall-clock deadline even when the selected acquisition was feasible under conservative estimates. Stopping one step earlier avoids the expansion but may produce an abstention or a claim unsupported by the acquired evidence. This cap-exhaustion trajectory is the unit that an average answer score or average token count cannot represent.

The corresponding control variable is not a scalar fidelity setting or an isolated stop score. It is the joint choice of *which modality to acquire, at what extent, and whether the current evidence is sufficient to stop*. All three coordinates interact with end-to-end latency, decoded media, and context-token budgets. Evaluating the decision also requires keeping planned feasibility separate from realized service compliance: a permissible action can still participate in a late route, and that overrun must remain attached to the policy that produced it.

CapSieve implements this route-level control. Its state combines remaining budgets, evidence coverage, and answerability; a conservative mask removes infeasible modality–extent expansions, measured consumption updates the residual envelope, and wall-clock latency independently determines SLA success. A stop action emits a cited answer only when the answerability gate accepts the accumulated evidence and otherwise terminates fail-closed. Hindsight routes provide an initialization signal, while a reward-only variant tests the same policy without that signal; evaluation always uses only evidence acquired along the executed route.

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 comparison changes acquisition logic while sharing answer models, candidate evidence, OCR/ASR outputs, prompts, extent grids, cap enforcement, held-out IDs, and seeds. It spans fixed-order and uncertainty controls, a calibrated VOILA-style selector (Bhope et al., 2026), a constrained actor-critic, component interventions, three cap tiers, and an InternVL3-8B replication. At 4 s, CapSieve retains 98.5% of uncapped fetch-all quality at 39% of its median normalized decode work and 95.2% SLA pass. It gains 0.6 aggregate points (95% CI [0.2, 1.0]) and 3.5 automatic evidence points over the constrained actor-critic at lower decode; reward-only training reaches 70.8. Human adjudication confirms the evidence advantage, and coupling the three action coordinates produces the largest component drop (0.9 points) when removed.

CapSieve contributes an executable acquisition object and a route-level evaluation object. The acquisition object couples modality, extent, and stopping while separating planned cap enforcement from realized SLA outcomes (Section 3). The evaluation object records native quality, work, abstention, deadline failure, and citation support on the same executed route, then measures how each burden is distributed over exclusive request slices (Sections 3–4). Following these routes through the results—from budget exhaustion, to early stopping and evidence support, to domain and backbone changes, and finally to burden localization—reveals the systems consequence: the 33.1% visual-temporal and multi-hop portion carries 59.7% of decode work and over 74% of both SLA failures and contradictions. Stable localization across caps, policies, and backbones identifies where marginal serving and modeling capacity can address the largest share of residual pressure.

2 Related Work

Stopping and retrieval decisions. Selective prediction controls answer emission (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019), and adaptive RAG controls when or where to retrieve text or knowledge (Jiang et al., 2023; Asai et al., 2024; Jeong et al., 2024). R1-Router extends retrieval to step-wise multimodal knowledge-base selection with Step-GRPO (Peng et al., 2025). These methods select an answer or source; CapSieve instead evaluates every feasible source together with a modality-specific extent and an explicit stop decision, all against the same three-dimensional residual budget.

Fidelity, frames, and model identity. VOILA selects visual fidelity before execution using calibrated value of information (Bhope et al., 2026). VideoRouter divides long-video tokens between temporal coverage and high-resolution fragments (Lin et al., 2026), ReQuest revisits uncertain questions through question-aware frame selection (Kim et al., 2026), and sparse video methods allocate visual observations (Lei et al., 2021; Han et al., 2024; Xu et al., 2026). LatentRouter selects among multimodal models (Cheng & Dong, 2026), while adaptive computation varies network depth or exit time (Graves, 2016; Bolukbasi et al., 2017; Schuster et al., 2022). CapSieve holds model identity and candidate evidence fixed and varies the acquisition trajectory. The matched VOILA-style baseline tests calibrated one-shot candidate value; the actor-critic tests sequential constrained control on the same state and action surface. Neither comparison drops realized overruns, context exhaustion, or fail-closed abstentions from accounting.

Evidence evaluation and burden localization. FActScore, RAGChecker, RAGAS, and ARES evaluate textual claim support or retrieval quality (Min et al., 2023; Ru et al., 2024; Es et al., 2024; Saad-Falcon et al., 2024). CapSieve attaches automatic and blinded human support judgments to the multimedia route and service outcome that produced each claim. Its burden profile measures a different object from worst-slice accuracy: the fraction of total loss, work, abstention, SLA failure, and contradiction mass generated by an exclusive request slice. Keeping those burdens separate connects acquisition design to capacity allocation without collapsing heterogeneous route failures into one score.

3 CapSieve

3.1 Joint fail-closed evidence acquisition

CapSieve starts from media metadata and a short transcript snippet when available. At step t, its state

$$s_t = [e(q), b_t, c_t, u_t]$$

CapSieve inference. (1) Initialize cheap metadata and transcript evidence, budgets b_0, and state s_0. (2) Estimate each acquisition action's latency, decode, and token increments; construct $\mathcal{A}_t^{\text{feas}}$. (3) Select one coupled modality–extent–decision action from the masked policy. (4) For `EXPAND`, acquire the requested evidence, charge measured consumption, and update coverage and uncertainty. (5) For `STOP`, emit an answer and citations when answerability passes; otherwise abstain. (6) If only `STOP` remains, terminate fail-closed and record the realized SLA outcome.

Figure 1: CapSieve's deployable decision loop. Training-only answers and fetch-all evidence never enter the state, action mask, or answerability decision at evaluation time.

contains a 1024-dimensional `bge-m3` query embedding $e(q)$ (Chen et al., 2024), remaining budgets b_t, coverage features c_t, and uncertainty features u_t. Coverage summarizes lexical and entity hits, timestamp overlap, and dispersion over retrieved OCR or transcript evidence. Uncertainty combines self-consistency disagreement with a learned answerability score over the current context. All generations used to estimate uncertainty consume the same latency and token budgets as answer generation.

An action $a_t = (m_t, x_t, d_t)$ jointly chooses modality m_t, extent x_t, and $d_t \in \{\texttt{STOP}, \texttt{EXPAND}\}$, with m_t and x_t null for the terminal `STOP` action. Available evidence includes OCR spans, longer ASR windows, keyframes, dense frames, and audio segments. The modality grids are keyframes $\{2, 4, 8, 12\}$, dense frames $\{32, 64, 128, 256\}$, and audio windows $\{5, 10, 20, 30\}$ seconds; trajectories typically contain 2–6 decisions. Scoring a complete tuple is essential: an extent is useful only for the selected modality, and the alternative to every expansion is a terminal decision on the evidence already paid for.

Two-layer cap semantics. Let $\widehat{L}_t(a)$, $\widehat{D}_t(a)$, and $\widehat{T}_t(a)$ denote conservative incremental latency, decode, and token consumption. A service tier initializes $b_0 = (B_L, B_D, B_T)$; after each executed action, $b_t = (B_t^L, B_t^D, B_t^T)$ is the remaining envelope. The planned-feasible action set is

$$\mathcal{A}_t^{\text{feas}} = \{a : d_t = \texttt{EXPAND},\ \widehat{L}_t(a) \le B_t^L,\ \widehat{D}_t(a) \le B_t^D,\ \widehat{T}_t(a) \le B_t^T\} \cup \{\texttt{STOP}\}.$$

Masking restricts acquisition to $\mathcal{A}_t^{\text{feas}}$ and updates b_{t+1} from measured consumption. Context use enters both the hard feasibility test and the reward, while conservative latency and decode estimates protect quantities whose realized costs can vary. The realized SLA is evaluated separately from request-level wall time $L(\tau)$: a request passes exactly when $L(\tau) \le B_L$. Hardware variation and stochastic generation can therefore produce a realized overrun after a planned-feasible route; such overruns remain SLA failures. If no expansion is feasible, CapSieve answers only when its answerability head accepts the accumulated evidence and otherwise abstains.

3.2 Training and component isolation

Training has three separated stages. First, an offline hindsight oracle receives each training answer and its full fetch-all candidate pool and labels the minimum-cost path whose evidence attains `support` under the evaluation verifier. Second, a two-layer MLP with hidden size 512 scores feasible modality–extent–stop tuples and is warm-started on those routes. Third, the masked policy is tuned with

$$r(\tau) = q(\tau) - \lambda L(\tau) - \mu T(\tau) - \nu D(\tau),$$
$$\max_{\pi_\theta} \mathbb{E}[r(\tau)] - \alpha \mathbb{E}[(L(\tau) - B_L)_+] - \beta \mathbb{E}[(D(\tau) - B_D)_+].$$

Here q, L, T, D are native task quality, end-to-end latency, context tokens, and decode work. The token term $\mu T(\tau)$ and the action mask enforce context cost; the hinge terms address realized latency and decode consumption beyond their estimated envelopes. AdamW uses learning rate 2×10^{-4}, cosine decay, 1k warmup steps, batch size 512, and entropy bonus 0.005. Variants train for 80k–160k steps, requiring 14.2 A100 80GB GPU-hours. Policy and cost-estimator fitting use benchmark training partitions; threshold selection and model selection use validation partitions. The answerability threshold and conservative cost margins are fixed before evaluation and shared across cap tiers. Evaluation IDs enter none of these fits, and evaluation receives neither benchmark answers nor fetch-all labels.

Three matched ablations identify the source of the gain. *No hindsight* trains the same masked policy from task and cost rewards alone. *Decoupled heads* selects modality, extent, and stop independently at matched parameter count. *No cost tuning* retains the supervised warm-start but omits the third stage. At 4 s, their aggregates are 70.8, 70.3, and 70.7, respectively, versus 71.2 for CapSieve. These interventions attribute 0.4 points to hindsight initialization, 0.9 to coupled decisions, and a 0.08 median decode-cost reduction to cost-sensitive tuning.

3.3 Route-linked burden concentration

CapSieve summarizes system pressure with a burden concentration profile. Let the held-out requests be partitioned into exclusive slices S_j, let ω_i be the fixed-mixture weight of request i, let $y_i \in [0,1]$ be its native score, and let w_{ik} be its contribution to burden k. We use normalized native-metric loss, decode work, abstention and SLA-failure indicators, and the number of `contradict` evidence events:

$$w_{ik} \in \{1 - y_i,\ D_i,\ \mathbf{1}[\text{abstain}], \mathbf{1}[L_i > B_L],\ n_i^{\text{contradict}}\}.$$

For each burden, its concentration and amplification in slice S_j are

$$C_k(S_j) = \frac{\sum_{i \in S_j} \omega_i w_{ik}}{\sum_i \omega_i w_{ik}}, \qquad A_k(S_j) = \frac{C_k(S_j)}{\sum_{i \in S_j} \omega_i / \sum_i \omega_i}.$$

The vector of C_k values preserves the distinct failure modes rather than collapsing them into a composite score; $A_k > 1$ identifies a slice carrying more burden than its fixed-mixture prevalence. Because every quantity is attached to the executed route, the same ledger supports policy comparison and bottleneck localization: quality gains describe the acquisition controller, while C_k and A_k identify where its residual pressure remains.

4 Experimental Protocol

Models, hardware, and data. The primary stack uses Qwen2.5-VL-32B-Instruct (Bai et al., 2025), Whisper-large-v3 (Radford et al., 2022), PaddleOCR v2.7.3 (Du et al., 2020), and `bge-m3`. Experiments ran on $4\times$ NVIDIA A100 80GB GPUs with CUDA 12.4, PyTorch 2.3.1, and driver 550.54.15; policy inference uses one GPU and VLM generation uses four-way tensor parallelism. The replication substitutes InternVL3-8B (Zhu et al., 2025) while retaining ASR, OCR, embeddings, evidence pools, prompts, and accounting. We report seeds $\{13, 37, 61, 71, 97\}$.

The independent evaluation units are the 5,349 questions in DocVQA v1.0 validation (Mathew et al., 2021), 2,500 ChartQA test questions (Masry et al., 2022), 8,000 ActivityNet-QA test questions (Yu et al., 2019), and 975 AudioCaps test clips (Kim et al., 2019); each AudioCaps clip's five captions form one reference set rather than five requests. Native metrics are ANLS, relaxed accuracy, accuracy, and reference-set token F1. The reported service aggregate is the fixed-mixture index $\sum_b \pi_b \bar{y}_b$, with benchmark weights $\pi = (0.4537, 0.0888, 0.2843, 0.1732)$ in the order above and each $\bar{y}_b$ expressed as a percentage. This vector declares the target service population and is fixed independently of method outcomes. A request from benchmark b has weight $\omega_i = \pi_b / n_b$, where n_b is its number of independent evaluation units. Table 4 retains every native benchmark metric, so the aggregate cannot conceal a domain-specific reversal.

Budgets and accounting. The three reported tiers pair latency caps of 1 s, 2 s, and 4 s with token caps of 4,096, 8,192, and 16,384, respectively; each also enforces its per-request decoded-media budget through the feasible-action mask and the extent grids in Section 3. Each tier is therefore a joint service contract, and cross-tier comparisons trace operating regimes rather than isolate one cap's causal effect. End-to-end wall time begins at policy invocation and ends at answer, abstention, or hard timeout; retrieval, reranking, verification, uncertainty estimation, and generation are included. Policy abstentions and hard timeouts earn zero native task credit. A completed answer beyond B_L retains its native score but is recorded as an SLA failure, so quality and compliance remain separate coordinates. For the modalities present in request i, normalized decode is

$$D_i = \frac{1}{|\mathcal{M}_i|} \sum_{m \in \mathcal{M}_i} \frac{d_{im}}{d_{im}^{\text{all}}},$$

where $\mathcal{M}_i \subseteq \{\text{frames}, \text{audio}\}$ contains only available media, d_{im} is decoded frames or audio seconds, and d_{im}^{all} is the request's fetch-all amount. Thus missing modalities never create a zero denominator and fetch-all is 1.00 for every request. Primary timing uses warm model caches, local media, and one query at a time; Section 6 reports cold-cache and four-request concurrency stressors.

Matched policies. All methods share candidate evidence pools, prompts, OCR/ASR outputs, extent grids, cap masks, answer stack, evaluation IDs, and seeds. Fetch-all measures uncapped headroom. Heuristic uses lexical triggers and fixed extents, Progressive fixes a cheap-to-expensive order, and Uncertainty-only matches the CapSieve MLP and parameter count while removing coverage and evidence-quality features. VOILA-style preserves calibrated value-of-information selection with a gradient-boosted predictor and isotonic calibration over the common candidates. Constrained AC provides the strongest sequential challenge: an oracle-free actor-critic with a Step-GRPO-style step reward over the identical state and coupled action space. Every learned policy uses the same training records, candidate-action traces, seeds, and cap validation; routing, calibration, uncertainty, verification, and generation overheads are charged.

Uncertainty and tests. Quality and automatic evidence scores are five-seed mean and standard deviation. Decode cost is the request median with IQR; latency tails and SLA intervals are request-bootstrapped. Method differences use identical request IDs and percentage-normalized native scores. Resampling is stratified by benchmark and recombined with the fixed π, while AudioCaps resamples clips rather than captions. A paired permutation test swaps method labels within request pairs, asking whether the fixed-mixture difference can arise under exchangeable method assignments; the CapSieve–Uncertainty-only result is $p = 0.006$. Paired bootstrap intervals use 10,000 stratified resamples.

Evidence labels. Claims receive `support`, `partial`, or `contradict`; `partial` denotes relevant but insufficient, span-inexact, or temporally mislocalized evidence. The automatic verifier uses OCR span matching, timestamp alignment within ± 2 seconds, and CLIP similarity > 0.72 for visual claims (Radford et al., 2021). At its held-out calibration operating point it achieves 91.3% precision and 84.6% recall, and its full-set score maps support/partial/contradict to 1/0.5/0. The independent audit in RQ3 instead measures whether an emitted answer cites evidence from which a human can recover the benchmark answer; it reports binary route-support precision and recall after excluding partial and obtains Cohen's $\kappa = 0.78$. Abstentions and timeouts remain in task and SLA outcomes but produce no claim–evidence pair for this conditional citation audit.

5 RESULTS

An executed route can hit four observable boundaries: acquisition work, realized deadline, answerability, and citation support. The results follow those boundaries as budgets tighten, isolate the decision structure that prevents premature stopping, adjudicate claim support, test benchmark and backbone changes, and finally locate the requests carrying each residual burden.

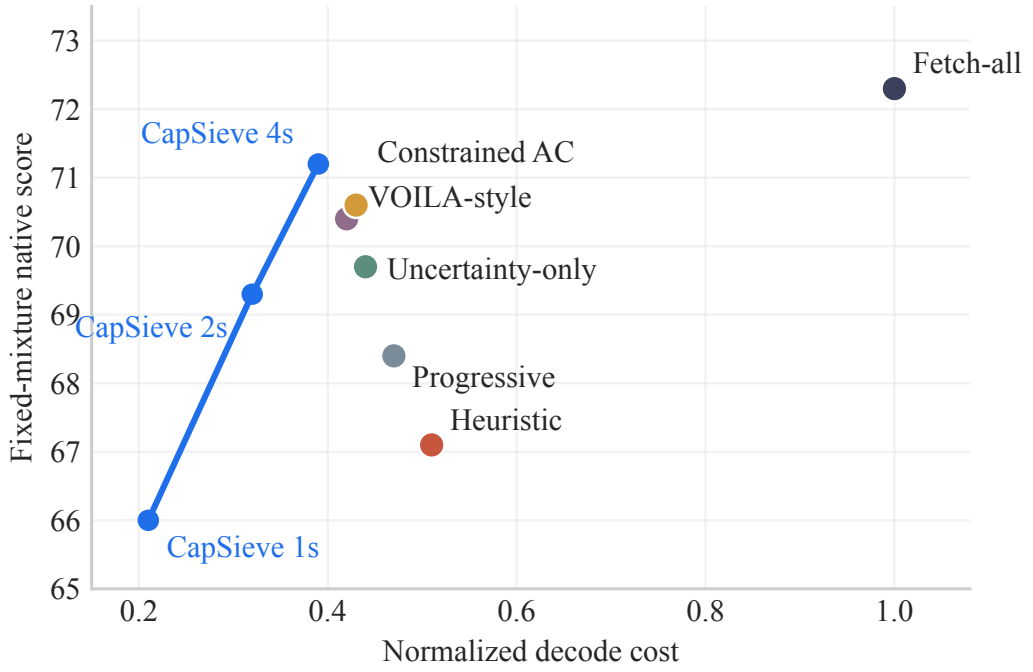

Figure 2: Cap-conditioned operating surface on the fixed held-out mixture. The curve connects CapSieve at 1/2/4 s: (66.0,0.21), (69.3,0.32), and (71.2,0.39) in aggregate/decode coordinates; four-second comparator bars show five-seed SD. Every method charges routing and generation overhead, and abstentions count as zero task credit.

5.1 RQ1: Budget exhaustion—quality, work, and deadlines

Table 1: Matched 4 s operating points on the fixed evaluation mixture. Aggregate and evidence are five-seed mean±SD; decode is median [IQR]. SLA and abstention are mixture-weighted request rates, and abstentions count as task errors.

Method	Aggregate	Decode	SLA (%)	Abst. (%)	Evid.
Fetch-all (uncapped)	72.3±0.9	1.00 [1.00,1.00]	—	—	—
Heuristic	67.1±1.3	0.51 [0.39,0.64]	94.9	8.8	72.8±2.2
Progressive	68.4±1.1	0.47 [0.34,0.61]	94.1	8.1	74.1±1.9
Uncertainty-only	69.7±1.0	0.44 [0.31,0.58]	94.5	7.5	76.3±1.8
VOILA-style	70.4±0.9	0.42 [0.30,0.55]	94.6	7.4	79.4±1.6
Constrained AC	70.6±0.9	0.43 [0.30,0.57]	94.3	6.9	80.2±1.5
CapSieve, no hindsight	70.8±0.9	0.41 [0.29,0.54]	95.0	6.3	82.2±1.5
CapSieve, decoupled	70.3±1.0	0.42 [0.30,0.56]	94.7	6.7	81.1±1.7
CapSieve, no cost tuning	70.7±0.9	0.47 [0.34,0.62]	93.6	5.7	82.5±1.5
CapSieve	71.2±0.8	0.39 [0.27,0.52]	95.2	5.4	83.7±1.4

At 4 s, CapSieve retains 98.5% of the uncapped fixed-mixture score (71.2 versus 72.3) at 39% of median normalized decode work, with 95.2% SLA pass. It improves on Uncertainty-only by 1.5 points ($p = 0.006$) and on Constrained AC by 0.6 points (95% CI [0.2,1.0], $p = 0.014$). It also raises automatic evidence alignment by 3.5 points while using 0.04 less median decode. Without hindsight labels, CapSieve reaches 70.8 at 0.41 decode, preserving the displayed ordering over Constrained AC (70.6/0.43) under reward-only training.

Figure 2 exposes diminishing acquisition returns: 1 s to 2 s gains 3.3 points for 0.11 decode cost, whereas 2 s to 4 s gains 1.9 points for 0.07. At the common 4 s operating point, CapSieve attains

Table 2: Tight-cap challenge at 1 s. All methods use identical held-out IDs and cap masks; aggregate includes abstentions as errors. Decode is request median and SLA is five-seed mean±SD.

Method	Aggregate	Decode	SLA (%)	Evid.
Heuristic	60.7±1.3	0.26	92.4±1.8	70.2±2.4
Uncertainty-only	63.4±1.1	0.25	91.6±2.0	73.8±2.0
VOILA-style	64.4±1.2	0.23	91.4±1.9	76.1±1.9
Constrained AC	64.9±1.2	0.24	90.8±2.1	76.9±1.8
CapSieve, no hindsight	65.3±1.1	0.22	91.0±2.0	79.0±1.7
CapSieve	66.0±1.1	0.21	91.1±1.9	80.6±1.7

the strongest aggregate and evidence alignment at the lowest learned-policy decode cost among the methods in Table 1.

At 1 s, CapSieve gains 5.3 points over Heuristic, 2.6 over Uncertainty-only, and 1.1 over Constrained AC while maintaining comparable realized SLA. Fail-closed scoring charges every policy abstention and hard timeout as zero task credit; completed late answers remain visible as SLA failures.

5.2 RQ2: Early stopping—which components preserve quality?

The interventions in Table 1 separate initialization, action structure, and optimization. Removing hindsight initialization reduces the aggregate by 0.4 points and evidence alignment by 1.5, quantifying the incremental benefit of the warm start. Decoupling modality, extent, and stopping causes the largest quality loss, 0.9 points, while increasing abstentions by 1.3 points. Omitting cost-sensitive tuning retains 70.7 quality but raises median decode from 0.39 to 0.47. The component evidence therefore identifies coupled action selection as the main quality mechanism and cost tuning as the main decode mechanism; reward-only training preserves the nominal learned-policy ordering.

5.3 RQ3: Citation support—does human adjudication preserve the ordering?

Table 3: Human evidence audit. Panel A pools 1,200 method-stratified claim–evidence pairs across 1/2/4 s. Panel B is a dedicated 4 s stratified audit; n is the binary denominator after excluding `partial`, followed by exact TP/FP/FN/TN and stratified-bootstrap 95% CIs. Cohen's $\kappa = 0.78$ measures pre-adjudication annotator agreement.

Method	n	Precision	Recall	TP/FP/FN/TN
Panel A: cap-pooled audit				
Heuristic	358	68.4	62.1	—
Uncertainty-only	361	73.2	67.8	—
CapSieve	367	82.6	79.3	—
Panel B: dedicated 4 s audit				
Uncertainty-only	181	73.3 [65.3,80.5]	68.1 [60.3,75.3]	96/35/45/5
VOILA-style	184	78.0 [70.4,84.4]	71.0 [63.4,77.9]	103/29/42/10
Constrained AC	183	80.3 [73.0,86.3]	76.4 [69.3,82.4]	110/27/34/12
CapSieve	188	86.2 [79.7,91.1]	79.1 [72.3,84.9]	125/20/33/10

After excluding partial labels, the pooled effective binary denominator is 1086. In the dedicated 4 s audit, CapSieve improves precision over Constrained AC by 5.9 points and recall by 2.7, preserving the automatic-audit ordering with exact counts and resampling uncertainty. The automatic three-level score measures graded support over emitted claims, while human binary precision and recall provide an independently adjudicated view; task quality continues to charge abstentions and timeouts separately.

5.4 RQ4: BENCHMARK AND BACKBONE SHIFT—DOES THE ROUTE ADVANTAGE PERSIST?

Table 4: Native metrics at 4 s: DocVQA ANLS, ChartQA relaxed accuracy, ActivityNet-QA accuracy, AudioCaps reference-set token F1, and the fixed-mixture index defined in Section 4. Entries are five-seed mean (SD).

Method	DocVQA	ChartQA	ActNet-QA	AudioCaps	Aggregate
Uncertainty-only	72.2 (0.8)	78.0 (1.1)	64.1 (1.0)	68.0 (1.4)	69.7 (1.0)
VOILA-style	72.7 (0.8)	78.8 (1.0)	65.0 (0.9)	68.8 (1.3)	70.4 (0.9)
Constrained AC	72.9 (0.8)	79.0 (1.0)	65.3 (0.9)	69.0 (1.3)	70.6 (0.9)
CapSieve, no hindsight	73.1 (0.8)	79.2 (1.0)	65.6 (0.9)	69.2 (1.2)	70.8 (0.9)
CapSieve	73.4 (0.7)	79.6 (0.9)	66.0 (0.8)	69.4 (1.2)	71.2 (0.8)

Relative to Constrained AC, CapSieve gains 0.5 DocVQA points (95% CI [0.1,0.9]), 0.6 on ChartQA (CI [0.0,1.2]), 0.7 on ActivityNet-QA (CI [0.2,1.2]), and 0.4 AudioCaps F1 (CI [-0.1,0.9]). The fixed-mixture advantage is thus supported by improvement or statistical parity on every benchmark, with the clearest gain where sequential visual acquisition is active. Appendix A reports the second-backbone native metrics.

Cross-backbone replication. Table 5 replaces Qwen2.5-VL-32B-Instruct with InternVL3-8B while preserving evidence candidates, ASR/OCR, prompts, splits, and cap accounting.

Table 5: InternVL3-8B replication at 4 s. Decode is normalized to this stack's own fetch-all run; p95 is in seconds and SLA is a percentage. Tail columns report the combined visual-temporal plus multi-hop share of each burden.

Method	Aggregate	Decode	p95/SLA	Evid.	Tail decode	Tail contrad.
Fetch-all	69.5±1.0	1.00	3.62/96.8	—	54.9	—
Constrained AC	67.8±1.0	0.41	2.08/96.4	79.1±1.7	59.2	70.5
CapSieve, no hindsight	68.0±0.9	0.40	2.02/96.7	80.4±1.6	58.9	71.4
CapSieve	68.5±0.9	0.38	1.96/97.0	82.3±1.5	58.4	72.2

With the smaller backbone, CapSieve retains a 0.7 point advantage over Constrained AC at 0.03 lower decode cost. The visual-temporal and multi-hop portion carries 58.4% of decode, 64.8% of abstentions, 70.5% of SLA failures, and 72.2% of contradictions. Thus the backbone change preserves both the acquisition ordering and the location of residual system pressure when absolute quality and timing change.

5.5 RQ5: RESIDUAL PRESSURE—WHICH REQUESTS CROSS EACH FAILURE BOUNDARY?

Table 6: CapSieve's fixed-mixture burden concentration profile at 4 s. Every entry is the slice's share of the corresponding mixture-wide burden. Abstentions count as task failures, and evidence events are automatic `contradict` labels.

Slice	Loss (%)	Decode (%)	Abst. (%)	SLA fail (%)	Contrad. (%)
ASR-only	24.0	15.3	13.8	9.5	8.1
OCR-heavy	22.6	25.0	19.5	16.0	17.7
Visual-temporal	31.0	36.4	37.8	42.7	41.5
Multi-hop	22.4	23.3	28.9	31.8	32.7

The head of the workload creates CapSieve's resource margin: ASR-only and OCR-heavy requests occupy 66.9% of the fixed mixture but consume only 40.3% of decode work. The visual-temporal and multi-hop portion occupies 33.1% yet carries 53.4% of task loss, 59.7% of decode, 66.7% of abstentions, 74.5% of SLA failures, and 74.2% of contradictions.

Relative to fixed-mixture prevalence, these loads are amplified by 1.61, 1.80, 2.02, 2.25, and 2.24, respectively. Deadline failure is therefore the maximally amplified burden, closely followed by evidence contradiction. Appendix B shows the same ordering under Constrained AC and at both 1 s and 4 s, identifying a shared workload bottleneck rather than a policy-specific artifact.

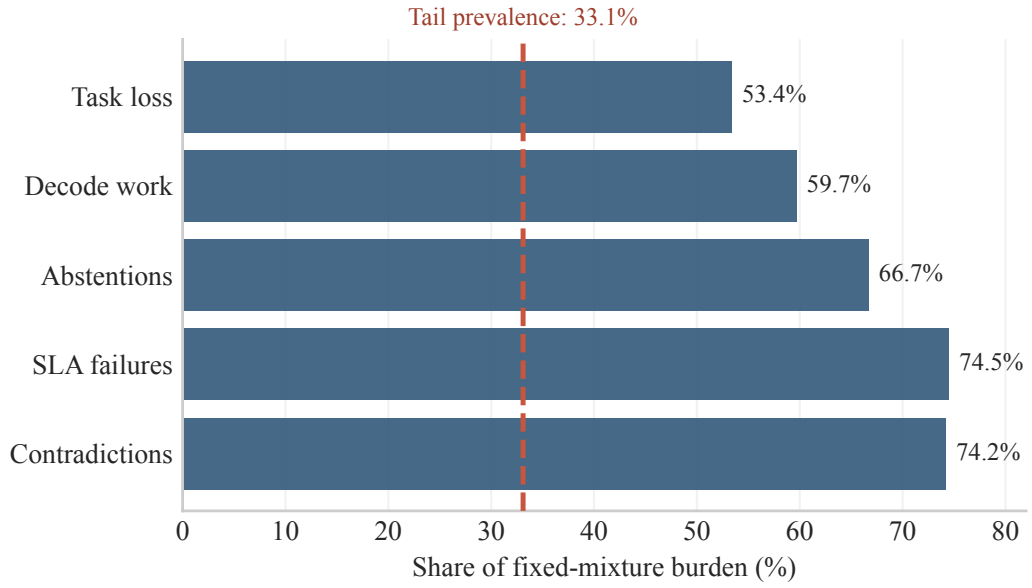

Figure 3: Burden amplification for the 33.1% visual-temporal and multi-hop portion at 4 s. Bars report its share of each fixed-mixture burden; the dashed line is its mixture prevalence. Every burden exceeds prevalence, with the largest amplification in SLA failures and evidence contradictions.

6 Discussion

Cold-cache and four-request concurrency stressors reduce but do not reverse CapSieve's serving margin: its p95/SLA is 4.27 s/91.8% versus 4.43 s/90.5% for Constrained AC under cold caches, and 4.71 s/89.7% versus 4.92 s/88.2% at four concurrent requests. Across the evaluated stacks, caps, and fixed mixture, route accounting therefore retains both the policy comparison and the location of residual work, abstention, deadline failure, and contradiction pressure.

7 Conclusion

CapSieve couples modality, extent, and stopping under one service budget and evaluates the route that actually executes. At 4 s it retains 98.5% of uncapped fixed-mixture quality at 39% median decode work, while improving on the matched learned controls. The executed-route ledger further shows that the 33.1% visual-temporal and multi-hop portion carries most residual decode, abstention, SLA-failure, and contradiction burden across caps, policies, and VLM backbones.

AI Use Statement

Generative AI tools were used in documented manuscript preparation for formatting checks, citation-metadata verification, and localized language and presentation edits. This disclosure covers that documented assistance.

Reproducibility Statement

The state and action definitions, cap semantics, training stages, model and hardware configuration, independent evaluation units, fixed-mixture aggregation, uncertainty conventions, and matched-policy controls appear in Sections 3–4. The appendices specify benchmark-level outcomes, slice assignment, latency tails, evidence-audit sampling and adjudication, and the cap-stressor matrix. The experiments use official releases of the four benchmarks and the published model or package terms for Qwen2.5-VL, InternVL3, Whisper, PaddleOCR, CLIP, and `bge-m3`; they introduce no new dataset asset.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct. Model release used: Qwen/Qwen2.5-VL-32B-Instruct.

Rahul Atul Bhope, K. R. Jayaram, Vinod Muthusamy, Ritesh Kumar, Vatche Isahagian, and Nalini Venkatasubramanian. VOILA: Value-of-information guided fidelity selection for cost-aware multimodal question answering. *arXiv preprint arXiv:2602.03007*, 2026. doi: 10.48550/arXiv.2602.03007. URL https://arxiv.org/abs/2602.03007.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2017. URL https://proceedings.mlr.press/v70/bolukbasi17a.html.

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, 2024. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Xueqi Cheng and Yushun Dong. LatentRouter: Can we choose the right multimodal model before seeing its answer? *arXiv preprint arXiv:2605.11301*, 2026. doi: 10.48550/arXiv.2605.11301. URL https://arxiv.org/abs/2605.11301.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020. doi: 10.48550/arXiv.2009.09941. URL https://pypi.org/project/paddleocr/2.7.3/. Software release used: PaddleOCR 2.7.3.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016. doi: 10.48550/arXiv.1603.08983. URL https://arxiv.org/abs/1603.08983.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Self-adaptive sampling for accurate video question answering on image text models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2522–2534. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.162. URL https://aclanthology.org/2024.findings-naacl.162/.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Minkuk Kim, Suyong Yun, Young Tae Kim, Jinyoung Moon, Jinwoo Choi, and Seong Tae Kim. ReQuest: Rethinking-based question-aware frame selection for long-form video QA. *arXiv preprint arXiv:2607.01737*, 2026. doi: 10.48550/arXiv.2607.01737. URL https://arxiv.org/abs/2607.01737.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: ClipBERT for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7331–7341, 2021. doi: 10.1109/CVPR46437.2021.00725.

Kuanwei Lin, Wenhao Zhang, and Ge Li. VideoRouter: Query-adaptive dual routing for efficient long-video understanding. *arXiv preprint arXiv:2605.05848*, 2026. doi: 10.48550/arXiv.2605.05848. URL https://arxiv.org/abs/2605.05848.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. doi: 10.1109/WACV48630.2021.00225.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei 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, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Chunyi Peng, Zhipeng Xu, Zhenghao Liu, Yishan Li, Yukun Yan, Shuo Wang, Zhiyuan Liu, Yu Gu, Minghe Yu, Ge Yu, and Maosong Sun. Learning to route queries across knowledge bases for step-wise retrieval-augmented reasoning. *arXiv preprint arXiv:2505.22095*, 2025. doi: 10.48550/arXiv.2505.22095. URL https://arxiv.org/abs/2505.22095.

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. *arXiv preprint arXiv:2212.04356*, 2022. doi: 10.48550/arXiv.2212.04356. URL https://huggingface.co/openai/whisper-large-v3. Model release used: openai/whisper-large-v3.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21999–22027, 2024. doi: 10.52202/079017-0692.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17456–17472, 2022. doi: 10.52202/068431-1269.

Chenwei Xu, Zhen Ye, Shang Wu, Weijian Li, Zihan Wang, Zhuofan Xia, Lie Lu, Pranav Maneriker, Fan Du, Manling Li, and Han Liu. Towards sparse video understanding and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11357–11368, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Xu_Towards_Sparse_Video_Understanding_and_Reasoning_CVPR_2026_paper.pdf.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019. doi: 10.1609/aaai.v33i01.33019127. URL https://ojs.aaai.org/index.php/AAAI/article/view/4946.

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. doi: 10.48550/arXiv.2504.10479. URL https://huggingface.co/OpenGVLab/InternVL3-8B. Model release used: OpenGVLab/InternVL3-8B.

A Per-Benchmark and Cross-Stack Detail

Table 7: InternVL3-8B native metrics at 4 s, using the same independent evaluation units and fixed-mixture index as Table 4.

Method	DocVQA	ChartQA	ActNet-QA	AudioCaps	Aggregate
Fetch-all	71.5	78.4	64.6	67.8	69.5
Constrained AC	69.8	76.0	62.8	66.4	67.8
CapSieve, no hindsight	70.0	76.2	63.0	66.6	68.0
CapSieve	70.5	76.9	63.5	67.2	68.5

B Slice Assignment and Policy-Robust Concentration

Each independent evaluation unit is assigned exactly once from its question, source metadata, and answer-supporting evidence; policy outcomes and burden values are not used for assignment. ASR-only requires that answer-supporting evidence occurs in transcript or audio windows without decisive OCR or frame evidence. OCR-heavy requires localized document or chart text and no temporal composition. Visual-temporal requires a localized frame sequence or audio-visual temporal relation. Multi-hop requires combining at least two separated evidence units, modalities, or retrieval steps. Rules are applied in reverse order of specificity: multi-hop, visual-temporal, OCR-heavy, then ASR-only. Their fixed-mixture prevalences are 38.2%, 28.7%, 21.4%, and 11.7%, respectively.

Table 8: CapSieve slice profile at 4 s. Shares partition the fixed mixture exactly once; evidence is a within-slice five-seed summary and decode is median [IQR].

Request slice	Share	Decode	Evid. (%)
ASR-only	38.2%	0.17 [0.11,0.24]	88.3±1.1
OCR-heavy	28.7%	0.33 [0.22,0.46]	84.1±1.5
Visual-temporal	21.4%	0.57 [0.41,0.79]	78.6±2.0
Multi-hop	11.7%	0.64 [0.46,0.88]	76.2±2.3

Table 9: Combined visual-temporal plus multi-hop burden shares across learned policies and caps. Every column reports a share of the corresponding fixed-mixture total; prevalence is 33.1%.

Policy/cap	Loss	Decode	Abstention	SLA failure	Contradiction
Constrained AC, 1 s	56.1	61.4	68.3	71.2	73.5
CapSieve, 1 s	57.2	62.1	67.5	70.1	76.4
Constrained AC, 4 s	52.5	60.0	66.2	72.6	70.9
CapSieve, 4 s	53.4	59.7	66.7	74.5	74.2

C Request-Level Latency Tails

Table 10: Request-level latency and termination accounting for the strongest external challenge and CapSieve. Percentiles include policy, retrieval, verification, uncertainty, and generation. SLA failure, answer, and abstention are fixed-mixture rates; an SLA failure can occur for either an answer or an abstention.

Policy/cap	p50	p90	p99	SLA fail (%)	Answer (%)	Abstain (%)	Route ms
Constrained AC, 1 s	0.58	0.97	1.54	9.2	84.5	15.5	31
CapSieve, 1 s	0.55	0.94	1.49	8.9	87.1	12.9	28
Constrained AC, 2 s	0.91	1.73	2.83	7.0	89.7	10.3	35
CapSieve, 2 s	0.86	1.65	2.69	6.6	91.0	9.0	31
Constrained AC, 4 s	1.04	2.08	4.51	5.7	93.1	6.9	38
CapSieve, 4 s	0.96	1.91	4.36	4.8	94.6	5.4	34

D Cap Semantics and Stressor Evidence Matrix

Table 11: Enforcement and observed evidence for each request-level service and evidence contract.

Contract	Enforcement and evaluation	Observed operating evidence
Latency	Conservative incremental estimates mask acquisition actions; realized wall time covers routing, retrieval, verification, uncertainty, and generation.	CapSieve uses 34 ms median routing at 4 s and retains its SLA advantage over Constrained AC under cold-cache and four-request concurrency; Tables 1 and 5, Appendix C.
Token cap	Context accounting charges retrieved evidence, uncertainty estimation, and final generation under the shared 4k/8k/16k caps.	The feasible-action mask preserves residual context for answer generation; Section 3.
Decode cap	Measured frame and audio acquisition updates the remaining per-request budget for every expansion.	CapSieve uses 0.39 versus 1.00 fetch-all median normalized work; dense temporal evidence carries the residual burden in Tables 6 and 5.
Abstention	Answerability acts on acquired evidence, and every fail-closed abstention receives zero task quality.	Visual-temporal and multi-hop requests carry 66.7% of 4 s abstention burden under the fixed-mixture weighting; Table 6.
Evidence	Every route citation receives automatic support scoring, complemented by blinded expert adjudication.	CapSieve reaches 83.7 automatic evidence score and leads the dedicated 4 s human audit; Table 3.

E Human Audit Protocol

The cap-pooled audit samples 1,200 claim–evidence pairs, with 544 from DocVQA, 112 from ChartQA, 336 from ActivityNet-QA, and 208 from AudioCaps. It allocates 400 pairs to Heuristic, 400 to Uncertainty-only, and 400 to CapSieve across the three cap tiers. The 4 s audit samples 800 method-stratified pairs across Uncertainty-only, VOILA-style, Constrained AC, and CapSieve.

Three experts independently view each question, answer, cited span or media window, and benchmark answer under blinded method identity. `Support` requires that the citation alone justifies the claim; `partial` marks relevant but insufficient, span-inexact, or temporally mislocalized evidence; `contradict` marks missing, conflicting, or incorrect evidence. Binary route-support precision and recall treat an emitted answer with a support-labeled citation as positive and benchmark-answer recoverability from that evidence as reference-positive; partial cases are excluded. This route-level outcome differs from the verifier's held-out calibration classification in Section 4. Cohen's $\kappa = 0.78$ precedes fourth-expert adjudication. Stratified bootstrap resamples within benchmark and method.