BUDGETED MULTIMODAL EVIDENCE SELECTION FOR LONG-VIDEO UNDERSTANDING VIA KNAPSACK OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding long videos with vision–language models (VLMs) requires selecting useful information within limits on the frames and text a model can process. Existing approaches address this by leveraging retrieval augmentation to supplement sampled frames with relevant information, such as speech transcripts and on-screen text, and select the most related evidence items based on similarity thresholds and ranking cutoffs. However, these heuristics do not directly optimize the joint usefulness of heterogeneous, unequal-cost evidence under a fixed model input budget. We make the first attempt in long-video understanding to formulate the final selection of unequal-cost textual evidence as a knapsack-constrained optimization problem, and introduce RCES, a training-free framework that allocates text and frames under separate input budgets without modifying the answering VLM. Specifically, RCES optimizes a monotone submodular objective that balances question relevance, temporal and source coverage, and representativeness of the selected evidence under a knapsack constraint, which can be solved by a density-greedy algorithm with performance guarantees on the objective value. A separate frame selector is further proposed to combine uniformly spaced frames for global context with short frame sequences around question-relevant moments. Across three long-video benchmarks and five VLM backbones, RCES achieves competitive accuracy with substantial reductions in textual evidence cost or latency, compared to related baselines.

1 INTRODUCTION

Long-video understanding matters wherever the recording is long but the question is narrow: finding where a concept is explained in a recorded lecture, establishing when an event occurred in hours of surveillance footage, or recovering what was decided in a long meeting recording. Such videos, spanning tens of minutes to hours, exceed what vision–language models (VLMs) can process at full temporal resolution. The events that answer a question are temporally sparse, while the number of visual tokens a model can process efficiently is bounded, so a fixed frame budget covers the timeline ever more coarsely as duration grows. This difficulty has drawn much attention very recently, and retrieval augmentation has emerged as a central strategy: it supplements the sampled frames with question-relevant information extracted from the video itself, including transcripts from automatic speech recognition (ASR), text from optical character recognition (OCR), and other timestamped evidence, so that a frozen VLM receives both a visual and a textual view of the video without processing it in full. Existing studies differ in the evidence they gather and in how many rounds of retrieval they run. For instance, Video-RAG (Luo et al., 2025) grounds the frozen VLM in a single pass over ASR, OCR, and detection text; TV-RAG (Cao et al., 2025) broadens what is gathered; and VideoDetective (Yang et al., 2026) retrieves iteratively, refining hypotheses across rounds.

However, retrieval creates a second problem: **budget allocation**. As video duration grows, extraction can produce substantially more candidate evidence than the final prompt can accommodate. Including more evidence may improve accuracy, but consumes limited context-window capacity and increases inference cost; consequently, what is retrieved need not be what should ultimately be presented to the VLM. This distinction matters because candidate evidence is heterogeneous: items differ in length, timestamp, source, question relevance, and redundancy. For example, a short OCR string and a

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.

paragraph of transcript consume very different fractions of the same text budget while contributing differently to the answer. Existing auxiliary-text pipelines (Luo et al., 2025; Cao et al., 2025; Yang et al., 2026) commonly control the retained evidence through similarity filtering, ranking cutoffs, or iterative retrieval, but these mechanisms alone do not explicitly optimize which combination of unequal-cost textual items should occupy a fixed final prompt budget. A related allocation problem has already been studied on the visual side, where principled frame and visual-token selection methods optimize under constraints on the number of retained visual units (Sun et al., 2025; Cho et al., 2025; Huang et al., 2026). The textual channel, however, has a different budget structure, because ASR and OCR items vary in length, so the constraint applies to their total cost rather than simply to how many items are selected. This motivates the question we study: *can final textual-evidence selection for long-video understanding be formulated directly as a budgeted optimization problem over unequal-cost items, rather than controlled indirectly through retrieval heuristics?*

We address this question with **RCES**, a training-free framework for budgeted evidence selection in long-video understanding without modifying the answering VLM, and make three contributions.

- **We formulate final textual-evidence selection as a budgeted knapsack optimization problem.** Unlike frame selection, where the budget is typically a count of visual units, retrieved ASR and OCR items have unequal lengths and therefore unequal costs. RCES explicitly selects which textual evidence should occupy a fixed prompt budget while accounting for question relevance, temporal coverage, source coverage, and representativeness.
- **We develop a principled solution to this optimization problem.** We show the resulting objective is monotone submodular under a knapsack constraint, which enables efficient optimization with a constant-factor approximation guarantee. We complement the textual selector with a separate budgeted frame allocator that preserves global temporal coverage while concentrating visual capacity around question-relevant moments.
- **We evaluate the resulting accuracy–efficiency trade-off across benchmarks and VLM backbones.** Experiments on three long-video benchmarks and five frozen VLM backbones show that RCES maintains competitive accuracy while reducing evidence volume or inference latency depending on the evaluation setting, compared with frame-selection and retrieval baselines.

2 Related Work

Auxiliary-text retrieval for long-video understanding. Retrieval augmentation supplements sampled frames with textual evidence extracted from the video. Video-RAG (Luo et al., 2025) uses ASR, OCR, and object-detection outputs to ground a frozen VLM through single-turn retrieval, with similarity thresholds controlling which auxiliary text is retained. TV-RAG (Cao et al., 2025) incorporates temporal weighting into retrieval and entropy-based frame selection. RCES builds on this auxiliary-evidence setting and adds an explicit allocation stage after candidate retrieval: the retrieved ASR/OCR items compete for a fixed budget in the final prompt.

Structured and agentic long-video reasoning. Prior work acquires evidence through iterative search (Wang et al., 2024; Ma et al., 2024) and hypothesis verification (Yang et al., 2026). To preserve relationships across video content, structured approaches organize retrieval through graphs (Shen et al., 2025) or hierarchical audiovisual representations (Yin et al., 2026). Reasoning is further supported by multi-agent collaboration (Yan et al., 2026) and progressive, coarse-to-fine exploration of relevant shots (Wang et al., 2026). RCES instead addresses final textual-evidence allocation under an explicit budget, optimizing a computable set objective without iterative calls to the answering VLM.

Visual evidence selection and compression. Question-conditioned frame selection seeks to retain relevant visual information within a limited input size. RAG-Adapter (Tan et al., 2025) retrieves frames relevant to the question, while Q-Frame (Zhang et al., 2025a) combines query-aware frame selection with resolution adaptation. Other methods account for relationships among selected frames: AKS (Tang et al., 2025) balances question relevance and video coverage, and AdaRD-Key (Zhang et al., 2025b) combines relevance with log-determinant diversity. Earlier video-summarization work uses sequential DPPs to select informative, diverse subsets while accounting for temporal structure (Gong et al., 2014). Adaptive methods also guide the search over the video: FOCUS (Zhu et al., 2025) uses bandit-based exploration to identify informative regions before selecting frames

within them, while HiMu (Ben-Ami et al., 2026) decomposes the question into a hierarchical logic tree and combines visual and audio expert signals to score frames. Selection can operate at both frame and token levels, e.g., APVR (Gao et al., 2025) combines pivot-frame retrieval with query-aware visual-token selection. RCES pairs its text selector with a separate frame allocator that combines uniform anchors and question-conditioned frame packets. It leaves per-frame tokenization unchanged and can be combined with visual-token compression.

Submodular selection for long-video understanding. Submodular objectives provide a way to model diminishing returns when selecting representative evidence. Classical results establish approximation guarantees under cardinality (Nemhauser et al., 1978) and knapsack constraints (Sviridenko, 2004; Khuller et al., 1999), and facility location is an established coverage objective (Lin & Bilmes, 2011). Within video understanding, MDP^3 (Sun et al., 2025) uses determinantal point processes for list-wise frame selection, and Adaptive Greedy Frame Selection (Huang et al., 2026) combines question relevance with facility-location representativeness under a frame-count constraint. FLoC (Cho et al., 2025) applies facility location to visual-token compression under a constraint on the number of retained tokens. RCES differs technically in both selected units and constraint. The visual formulations above select frames or visual tokens under a cardinality budget, whereas retrieved ASR/OCR items have heterogeneous lengths and unequal evidence costs. RCES consequently optimizes a knapsack-constrained objective over timestamped textual items, combining question relevance, temporal and source coverage, and facility-location representativeness. Selection uses marginal objective gain per unit cost; the visual frame budget is allocated separately.

3 Problem Formulation

We study long-video understanding through question answering. Given a video V of duration T and a question q, a frozen VLM M must produce an answer from visual and textual evidence extracted from V under separate evidence budgets. Retrieval identifies potentially useful evidence; the selection problem studied here determines which of that evidence is presented to the answering model.

Textual and visual evidence. Following the auxiliary-evidence construction of Video-RAG (Luo et al., 2025), we represent information extracted from the video as timestamped textual items. An *ASR* item is a time-aligned chunk of the transcript obtained from the video's audio. An *OCR* item contains on-screen text recognized in a sampled frame and inherits that frame's timestamp. *DET* items provide textual descriptions of detected objects and their spatial information, extracted from query-relevant sampled frames. Thus, ASR is audio-derived, whereas OCR and DET are frame-derived; all three enter the textual channel.

Let $\mathcal{E}_T = \{e_1, \ldots, e_n\}$ denote the textual candidate pool after retrieval and near-duplicate distillation (§4.2). Each item e contains text $\mathrm{text}(e)$, a source $\mathrm{src}(e) \in \{\mathrm{ASR}, \mathrm{OCR}, \mathrm{DET}\}$, and a timestamp $\tau(e)$. We use the midpoint of the associated interval as the timestamp for ASR and the source frame's timestamp for OCR and DET. Each item also carries a normalized question-relevance score $r(e) \in [0, 1]$; its computation is described in §4.2.

Visual evidence is represented separately by a candidate frame bank $\mathcal{E}_V$ sampled from V. A textual item can be retained even when its source frame is not selected as visual input. Given selected subsets $S_T \subseteq \mathcal{E}_T$ and $S_V \subseteq \mathcal{E}_V$, the frozen model produces an answer $\hat{a} = M(q, S_V, S_T)$.

Evidence budgets. Textual items vary in length and therefore consume different amounts of context. We assign each item a positive, backbone-independent evidence cost

$$c_T(e) = \max\big(1,\ \lfloor |\mathrm{text}(e)|/4 \rfloor\big), \tag{1}$$

where $|\mathrm{text}(e)|$ is the number of characters. Because tokenizers differ across backbones, this shared unit supports comparisons of evidence volume across models and benchmarks without depending on a particular tokenizer. This common cost unit tracks tokenizer counts empirically (§5.1); the budget is enforced in these defined units. The textual and visual subsets satisfy

$$\sum_{e \in S_T} c_T(e) \le B_T, \qquad |S_V| \le B_V.$$

Here B_T constrains selected evidence text, excluding the question and prompt instructions, while B_V limits the number of selected frames. The optimization below concerns textual selection; the separate allocation of the frame budget is described in §4.2.

Budgeted textual-evidence selection. Given the candidate pool, selection must account for which items are useful together: individually relevant items may overlap, while items from different sources or moments may provide complementary information. Their unequal costs also determine which combinations fit within the evidence budget. Let $U_q(S_T; S_V)$ denote the usefulness of textual subset S_T for answering q correctly, with the video V, visual subset S_V, and model M held fixed.

Definition 1 (Budgeted textual-evidence selection)**.** *Given a textual candidate pool $\mathcal{E}_T$, a fixed visual subset S_V, a question-conditioned utility U_q, and a text evidence budget B_T, select*

$$S_T^\star \in \arg\max_{S_T \subseteq \mathcal{E}_T} U_q(S_T; S_V) \quad \text{s.t.} \quad \sum_{e \in S_T} c_T(e) \le B_T.$$

This problem presents two challenges. First, the ideal utility U_q is not directly observable at inference time because the correct answer is unknown. Calling the VLM on alternative subsets incurs additional computation without establishing whether its answers are correct. Second, subset selection is combinatorial: unequal costs and overlapping or complementary content make the value of an item depend on the other selected evidence. Ranking items by individual relevance does not account for these interactions, while exhaustive subset evaluation is impractical for large pools. These challenges motivate a computable selection objective and an efficient budget-aware algorithm.

4 METHOD: RCES

RCES replaces the latent utility U_q of §3 with a computable set objective optimized under B_T. We present the text selector and optimizer, then the training-free pipeline for candidate retrieval and distillation, separate frame allocation under B_V, and answering with a frozen VLM.

4.1 BUDGETED TEXTUAL-EVIDENCE SELECTION

To instantiate the latent utility U_q from §3, we seek a surrogate that can be evaluated directly from the retrieved evidence without calling the answering VLM. For clarity, in this subsection we write $\mathcal{E} = \mathcal{E}_T$ and $S = S_T$. A selected subset should primarily contain evidence relevant to the question, while avoiding unnecessary concentration on repeated content, a single temporal region, or a single evidence source. We encode these properties through four set functions: question relevance, temporal coverage, source coverage, and representativeness. The individual components are standard; their role here is to define the utility of a *combination* of unequal-cost textual items under the evidence budget introduced in §3.

Question relevance. Each retrieved item e is associated with its retrieval similarity to the question. Because similarity scales may differ across ASR, OCR, and DET evidence, we normalize retrieval scores within each source to obtain $r(e) \in [0, 1]$. We reward the cumulative relevance of a subset by

$$\mathrm{Rel}(S) = \sum_{e \in S} r(e). \tag{2}$$

This term is modular: the relevance contribution of an item does not depend on which other items have already been selected.

Temporal coverage. Long-video evidence may be concentrated around only a small portion of the timeline even when several moments are potentially useful. We therefore use a lightweight temporal-spread reward. A video of duration T is divided into n_b equal-duration bins, and

$$\mathrm{bin}(e) = \min\left(\left\lfloor n_b \frac{\tau(e)}{T} \right\rfloor, n_b - 1\right). \tag{3}$$

Here, $\mathrm{bin}(e)$ is the index of the equal-duration time interval containing the timestamp $\tau(e)$. The temporal coverage $\mathrm{CovTime}(S)$ counts how many distinct intervals contain selected evidence:

$$\mathrm{CovTime}(S) = \left|\{\mathrm{bin}(e) : e \in S\}\right|. \tag{4}$$

This term encourages temporal spread; it should not be interpreted as a model of temporal reasoning.

Source coverage. ASR, OCR, and DET encode different forms of auxiliary textual evidence. Here, S denotes the selected set of textual evidence items, while $\mathrm{src}(e)$ identifies the source of item e (ASR, OCR, or DET). Source coverage counts the distinct sources represented in S:

$$\mathrm{CovSrc}(S) = \left|\{\mathrm{src}(e) : e \in S\}\right|. \tag{5}$$

This is a lightweight coverage prior rather than a requirement that every source be represented.

Representativeness. Relevance and coarse coverage alone do not account for semantic redundancy among candidate items. We therefore reward subsets that represent the retrieved pool in embedding space using facility location:

$$\mathrm{FL}(S) = \frac{1}{|\mathcal{E}|}\sum_{x\in\mathcal{E}} \max_{s\in S} \mathrm{sim}^{+}(x,s), \qquad \mathrm{sim}^{+}(x,s) = \max\big(0, \mathrm{sim}(x,s)\big), \tag{6}$$

where sim is cosine similarity between Contriever embeddings and the empty-set value is defined as zero. Once a region of the candidate pool is already represented, selecting another highly similar item provides little additional facility-location gain. The nonnegative clamp preserves monotonicity even when raw cosine similarities are negative.

Combining these components gives the RCES textual objective

$$\boxed{F(S) = \lambda_{\mathrm{Rel}}\mathrm{Rel}(S) + \lambda_T \mathrm{CovTime}(S) + \lambda_{\mathrm{src}}\mathrm{CovSrc}(S) + \lambda_{\mathrm{FL}}\mathrm{FL}(S)} \tag{7}$$

with nonnegative weights. RCES instantiates the budgeted problem of §3 as

$$\max_{S\subseteq\mathcal{E}_T} F(S) \qquad \text{s.t.} \qquad \sum_{e\in S} c_T(e) \le B_T. \tag{8}$$

Unlike a similarity threshold or a fixed top-k rule, Eq. (8) scores evidence jointly while explicitly accounting for the total cost of the selected subset. Alg. 1 returns an approximate solution S_T to this problem.

Structure of the objective. The objective has a useful diminishing-returns structure.

Proposition 1 (Monotone submodularity)**.** *Assume* $r(e) \ge 0$, $\lambda_{\mathrm{Rel}}, \lambda_T, \lambda_{\mathrm{src}}, \lambda_{\mathrm{FL}} \ge 0$, *and* $\mathrm{sim}^{+} \ge 0$ *as defined in Eq.* (6)*. Then* F *in Eq.* (7) *is normalized, monotone nondecreasing, and submodular.*

Relevance is modular; temporal and source coverage are standard coverage functions; and facility location is a nonnegative sum of maximum-similarity functions. Each component is therefore monotone submodular, and nonnegative linear combinations preserve both properties. The complete proof is given in App. E.

Budget-aware optimization. Budgeted submodular text selection has been studied in summarization (Lin & Bilmes, 2010). Monotone submodular maximization under one knapsack constraint admits a $(1 - \exp(-1))$ approximation using partial enumeration followed by greedy completion (Sviridenko, 2004). Instead, RCES uses the substantially lighter modified density-greedy rule. As shown in Alg. 1, at each step, it selects the item with the largest marginal objective gain per unit evidence cost,

$$\rho(e \mid S) = \frac{F(S\cup\{e\}) - F(S)}{c_T(e)}, \tag{9}$$

subject to the remaining budget. The final greedy set is compared with the best single feasible item. This modified greedy procedure has a constant-factor worst-case guarantee for monotone submodular maximization under a knapsack constraint (Tang et al., 2021). Importantly, this guarantee concerns the surrogate objective value F, not downstream question-answering accuracy.

4.2 RCES Pipeline

Figure 1 shows how the selector above is embedded in the full training-free pipeline. For each video–question pair, RCES proceeds from candidate construction to final answer generation without updating the parameters of the answering VLM.

Algorithm 1 Budgeted density greedy for textual evidence selection

Require: pool $\mathcal{E}$, budget B_T, objective F, costs c_T; density ρ from Eq. (9)

$S \leftarrow \emptyset$, $R \leftarrow \mathcal{E}$
if no item satisfies $c_T(e) \le B_T$ **then return** $\emptyset$
while $R \neq \emptyset$ **do**
  $e^\star \leftarrow \arg\max_{e \in R} \rho(e \mid S)$
  if $\sum_{s \in S} c_T(s) + c_T(e^\star) \le B_T$ **then** $S \leftarrow S \cup \{e^\star\}$
  $R \leftarrow R \setminus \{e^\star\}$
end while
$e^\dagger \leftarrow \arg\max_{e \in \mathcal{E}:\, c_T(e) \le B_T} F(\{e\})$
return S if $F(S) \ge F(\{e^\dagger\})$; otherwise $\{e^\dagger\}$

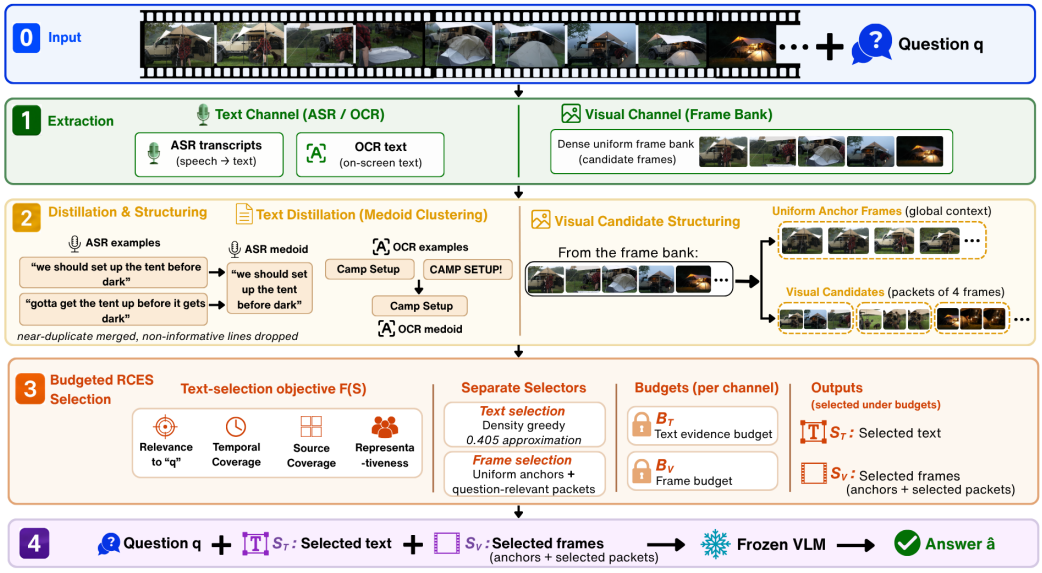


Figure 1: **RCES overview.** The video is converted into retrieved timestamped textual candidates and a candidate frame bank. Text candidates are distilled and selected under evidence budget B_T using the budgeted objective of §4.1. Frames are allocated separately under B_V by combining uniformly spaced global anchors with short packets around question-relevant moments. The selected text S_T and frames S_V are supplied with the question to a frozen VLM.

Evidence extraction and retrieval. Using the evidence construction described in §3, RCES extracts timestamped ASR and OCR evidence, with optional detection text when enabled. Textual items are embedded and retrieved by similarity to the question, yielding $\mathcal{E}_T^{\text{raw}}$ for subsequent distillation and budgeted selection. In parallel, RCES samples a dense candidate frame bank $\mathcal{E}_V$ for visual allocation. Textual evidence extracted from a frame can be retained without selecting that frame as visual input.

Candidate distillation. Retrieved ASR and OCR streams often contain repeated or near-duplicate content, such as overlapping transcript chunks or persistent on-screen text. Before budgeted selection, RCES removes non-informative candidates and groups the remaining items by temporal bin and evidence source. Within each group, items whose Contriever embeddings exceed a similarity threshold θ_{distill} are clustered, and the medoid of each cluster is retained as its textual representative. The representative inherits the largest normalized relevance score among its cluster members. This stage reduces repeated candidates before optimization; detailed filtering rules and the experimental threshold are given in App. A.3. The resulting distilled pool is the $\mathcal{E}_T$ used in Eq. (8). RCES then applies Alg. 1 to obtain S_T satisfying the evidence budget B_T.

Visual allocation: global anchors and local packets. The visual channel is allocated separately because its budget is a frame count rather than an unequal-cost text budget. RCES retains uniformly spaced *anchor* frames for global temporal coverage and allocates the remaining capacity to short

packets around question-relevant moments. Packet centers are ranked by visual–question similarity and temporally separated to reduce temporal overlap; overlapping frames are counted once. This allocator is separate from the textual knapsack objective: frames have uniform unit cost, and Eq. (7) is not applied to the visual channel. Implementation details are in App. A.4.

Answer generation. RCES orders the selected text S_T chronologically and supplies it with frames S_V and question q to the frozen VLM in a single call.

5 EXPERIMENTS

5.1 SETUP

Benchmarks and backbones. We evaluate on three long-video benchmarks through multiple-choice question answering. **VideoMME** (Fu et al., 2025) contains 2,700 questions evenly divided across short, medium, and long videos. We evaluate without dataset-provided subtitles; ASR transcripts are extracted from the video audio. **LongVideoBench** (Wu et al., 2024) uses its validation split of 1,337 questions; **MLVU** (Zhou et al., 2025) contributes 2,174 multiple-choice questions over seven tasks. The primary controlled comparison uses Qwen2.5-VL-7B-Instruct (Qwen Team, 2025); we additionally evaluate LLaVA-Video-7B (Zhang et al., 2024), InternVideo2.5-8B (Wang et al., 2025b), InternVL2.5-8B (Chen et al., 2024), and LongLLaVA-7B (Wang et al., 2025a).

Baselines and evaluation regimes. Baselines include uniform sampling, CLIP Top-K, Q-Frame (Zhang et al., 2025a), FOCUS (Zhu et al., 2025), AKS (Tang et al., 2025), HiMu (Ben-Ami et al., 2026), and the retrieval systems Video-RAG (Luo et al., 2025), TV-RAG (Cao et al., 2025), and VideoDetective (Yang et al., 2026). Table 1 compares complete systems on VideoMME under a matched text budget, $B_T = 800$. Table 2 fixes LLaVA-Video-7B across three benchmarks: competitors use their native text footprints, while RCES retains $B_T = 800$. Component effects are examined in §5.3; reproduction details are in App. A.

Metrics and implementation. We report accuracy, mean textual evidence cost from Eq. (1), and mean wall-clock latency in seconds per question. Evidence cost uses backbone-independent units, rather than tokenizer-specific counts. MLVU uses M-Avg, the unweighted mean of seven task accuracies. Exact two-sided McNemar tests use paired question-level correctness; on MLVU they assess micro accuracy, not M-Avg (App. D). Evidence-cost validation is in App. A. Unless stated otherwise, RCES uses $B_T = 800$, $B_V = 64$, and no detection. We use $\lambda_{\mathrm{Rel}} = 1.0$, $\lambda_T = 0.2$, $\lambda_{\mathrm{src}} = 0.2$, $\lambda_{\mathrm{FL}} = 0.3$, and $n_b = 8$ across benchmarks and backbones. Weights were fixed before the sensitivity analysis (App. C); implementation settings are in App. A.

5.2 MAIN RESULTS

RCES achieves competitive accuracy with substantial reductions in evidence cost or latency in several settings. The trade-offs differ between matched text budgets and native retrieval footprints.

Matched-budget comparison. Under the common $B_T = 800$ cap, RCES has the lowest latency among retrieval methods on all three backbones (Table 1). Compared to TV-RAG, it uses 23.5–24.5% less evidence and reduces latency by 59–72%. Accuracy is 65.33 versus 64.04 on Qwen2.5-VL, 68.30 versus 67.48 on LLaVA-Video, and 67.90 versus 67.30 on InternVideo2.5; the Qwen advantage is not statistically significant (App. D). Compared with Video-RAG, RCES has higher measured accuracy and lower latency on all three backbones, with similar evidence cost. Frame-only methods provide a separate reference: on Qwen, AKS reaches 63.20 accuracy but is much slower, while FOCUS is faster than RCES (15.6 versus 17.3 seconds) but with a much lower accuracy.

Native-footprint comparison across benchmarks. Table 2 compares retrieval systems using LLaVA-Video-7B: competitors retain their native text footprints, while RCES uses $B_T = 800$. On VideoMME, RCES exceeds Video-RAG's accuracy (68.30 versus 67.52) with 72.3% lower evidence cost and 47% lower latency. TV-RAG achieves higher accuracy (69.48), but has higher evidence cost and latency than RCES. On LongVideoBench, RCES reaches 59.39 accuracy versus 59.91 for Video-RAG and 59.16 for TV-RAG, with 70.5% and 39.2% lower evidence cost, respectively. Its latency (22.9 s) exceeds Video-RAG's (18.0 s) and is close to TV-RAG's (23.6 s). On MLVU, RCES has the highest measured M-Avg (68.19) and lowest latency (13.8 s) among these systems, despite

Table 1: **Matched-budget comparison on VideoMME** ($B_V = 64$). Retrieval methods use $B_T = 800$; frame-only methods receive no retrieved text. Ev. denotes the evidence used: frames (F), ASR (A), and OCR (O). Cost follows Eq. (1); latency is mean seconds per question. Bold marks best recorded values within each method block and backbone. — denotes inapplicable text cost. RCES duration breakdowns are in App. B.

		Qwen2.5-VL-7B			**LLaVA-Video-7B**			**InternVideo2.5-8B**		
Method	Ev.	Acc. ↑	Cost ↓	Lat. ↓	Acc. ↑	Cost ↓	Lat. ↓	Acc. ↑	Cost ↓	Lat. ↓
Frame selection (B_T not applicable)										
Uniform sampling	F	61.15	—	40.2	64.15	—	**9.7**	60.89	—	**7.7**
Top-K (CLIP)	F	60.80	—	35.5	63.89	—	22.6	61.52	—	15.3
Q-Frame	F	61.30	—	69.4	64.9	—	58.2	61.56	—	41.7
HiMu	F	58.74	—	52.9	62.4	—	45.7	**62.33**	—	38.1
FOCUS	F	61.11	—	**15.6**	61.11	—	19.7	60.33	—	8.4
AKS	F	**63.20**	—	182.3	**65.11**	—	180.1	62.11	—	19.2
Budgeted retrieval (B_T=800 on every method)										
Video-RAG	F+A+O	61.00	**451**	25.1	63.78	472	43.3	66.33	461	40.8
TV-RAG	F+A+O	64.04	616	42.4	67.48	596	40.6	67.30	595	35.1
VideoDetective	F+A+O	54.22	709	284.3	59.63	721	273.6	57.15	729	291.9
RCES	F+A+O	**65.33**	467	**17.3**	**68.30**	**450**	**15.2**	**67.90**	**455**	**9.7**

higher evidence cost than Video-RAG and TV-RAG (396 versus 385 and 299 units). Most distilled pools (88.6%) already fit within B_T, limiting further text reduction. Paired correctness tests find no significant differences between RCES and Video-RAG or TV-RAG on LongVideoBench or MLVU (App. D); the MLVU tests assess micro accuracy, not M-Avg.

Table 2: **Native-footprint comparison across three benchmarks** (LLaVA-Video-7B, $B_V = 64$). Competitors use their native text footprints without the common cap; RCES uses $B_T = 800$. Cost follows Eq. (1); latency is mean seconds per question. Bold marks the best recorded value per column.

	VideoMME			LongVideoBench			MLVU		
Method	Acc. ↑	Cost ↓	Lat. ↓	Acc. ↑	Cost ↓	Lat. ↓	M-Avg ↑	Cost ↓	Lat. ↓
Video-RAG	67.52	1623	28.7	**59.91**	1907	**18.0**	67.44	385	34.6
TV-RAG	**69.48**	1481	20.3	59.16	924	23.6	67.23	**299**	17.7
VideoDetective	58.33	2451	314.5	47.20	1136	144.3	59.96	2568	244.9
RCES (B_T=800)	68.30	**450**	**15.2**	59.39	**562**	22.9	**68.19**	396	**13.8**

Additional backbones. Across five backbones and three benchmarks, RCES increases measured accuracy over the corresponding uniform-frame baseline without retrieved text in all 15 pairs (Table 4). InternVL2.5-8B improves from 61.20 to 63.13 on LongVideoBench (+1.93 points). LongVideoBench supplies subtitles as a source for textual retrieval; their availability does not mean that the baseline receives them in its prompt. These comparisons measure the combined effects of retrieval, text selection, and frame allocation with unchanged RCES settings.

5.3 Ablation Studies

The ablations distinguish the effect of textual selection under a fixed budget from the effects of the set objective and visual allocator.

Textual allocation. With the same candidate pool and uniform frames at $B_T = 800$, the full objective reaches 63.74 overall and 54.44 long-split accuracy (Table 3). It exceeds relevance-only selection by 1.22 overall points, cost-aware relevance by 1.52, and temporal truncation by 1.81. The corresponding long-split gains are 1.11, 1.44, and 2.00 points. Cost-aware relevance uses the least text (445 evidence units), compared with 463 for relevance-only selection and 467 for the full objective, illustrating the accuracy–evidence-cost trade-off.

Removing one objective term at a time lowers overall accuracy by 0.56–1.33 points and long-split accuracy by 0.67–1.78 points (App. C.1). Removing relevance produces the largest drop, followed by representativeness, temporal coverage, and source coverage. These results support retaining all four terms in this configuration. The approximation guarantee concerns the surrogate objective value, rather than answer accuracy (Remark 1).

Visual allocation. With text selection fixed, anchors and packets raise overall accuracy from 63.74 to 65.33, with smaller long-split gains; App. C examines lower frame budgets and other packet layouts.

Text-budget sensitivity. Larger text budgets increase evidence cost, but do not consistently improve overall accuracy. Among the tested budgets, overall accuracy peaks at $B_T = 800$, while long-split accuracy increases throughout the sweep (Table 3). Based on these VideoMME results, we selected $B_T = 800$ and kept it fixed for the remaining benchmarks.

Objective-weight sensitivity. Across ten configurations on a stratified subset of 421 questions, overall accuracy spans 0.48 points (62.23–62.71); varying the representativeness weight yields a 2.13-point long-split spread. Weights were fixed before this analysis and reused across benchmarks and backbones, so the sweep measures sensitivity rather than selecting weights (App. C).

Table 3: **Allocation ablations on VideoMME** (Qwen2.5-VL-7B, $B_V = 64$). Left: fixed pool and uniform frames. Right: full objective with frame packets, varying only B_T. Comparing the full-objective $B_T = 800$ settings isolates frame allocation. Cost is mean evidence cost.

Text selection ($B_T = 800$)

Selection rule	Overall ↑	Long ↑	Cost
Temporal truncation	61.93	52.44	463
Relevance-only	62.52	53.33	463
Cost-aware relevance	62.22	53.00	445
Full RCES objective	**63.74**	**54.44**	467

Text budget (full objective + packets)

B_T	Overall ↑	Long ↑	Cost
500	64.19	54.22	319
800	65.33	54.78	467
1000	65.00	55.67	554

Additional analyses. Detection is disabled because of its latency cost (App. C.7); further component and solver analyses are in App. C.

6 Conclusion and Limitations

We introduced RCES, a training-free framework that formulates final unequal-cost textual-evidence selection as monotone-submodular maximization under a knapsack constraint, with visual evidence allocated separately. The text selector allocates the budget by marginal set utility per unit cost, while uniform anchors and question-relevant frame packets provide global context and local visual detail to a frozen answering VLM. Across three benchmarks and five VLM backbones, RCES maintains competitive accuracy while reducing textual evidence cost, latency, or both in several controlled comparisons. In the controlled VideoMME ablations, the full objective outperforms the three simpler text-selection rules, and removing any single objective term lowers measured accuracy. Larger text budgets do not consistently improve overall accuracy, motivating explicit control over the amount and composition of evidence presented to the model.

Limitations. RCES cannot recover evidence absent from its candidate pool, and budgeted selection offers little scope for text reduction when the pool already fits within B_T. The approximation guarantee applies to the surrogate objective rather than QA accuracy, and text and frames are selected under separate budgets. Our evaluation is limited to multiple-choice QA; improving candidate acquisition and joint multimodal allocation remain future directions.

AI use statement

We used generative AI to identify related work, assist with mathematical formulations, derivations, and proofs, and draft and refine the text. The authors take responsibility for the paper's final content.

Reproducibility statement

RCES is training-free, with deterministic selection given the scored candidate pool and fixed tie-breaking, and greedy VLM decoding. We report single runs; results may vary across hardware and extraction-model versions. Budgets and objective weights are in §5.1; other settings and the full protocol are in App. A. We release our code and per-question outputs; all backbones and benchmarks are public.

REFERENCES

Dan Ben-Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. Himu: Hierarchical multimodal frame selection for long video question answering, 2026. URL https://arxiv.org/abs/2603.18558.

Zongsheng Cao, Yangfan He, Anran Liu, Feng Chen, Zepeng Wang, and Jun Xie. Tv-rag: A temporal-aware and semantic entropy-weighted framework for long video retrieval and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. doi: 10.1145/3746027.3755873. URL https://arxiv.org/abs/2512.23483.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2024. URL https://arxiv.org/abs/2412.05271.

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. Floc: Facility location-based efficient visual token compression for long video understanding, 2025. URL https://arxiv.org/abs/2511.00141.

Chaoyou Fu et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Hong Gao, Yiming Bao, Xuezhan Tu, Bin Zhong, and Minling Zhang. Apvr: Hour-level long video understanding with adaptive pivot visual information retrieval, 2025. URL https://arxiv.org/abs/2506.04953.

Bo Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, 2014.

Yuning Huang, Xiaoyu Ji, Joseph Huang, Yichi Zhang, and Fengqing Zhu. Adaptive greedy frame selection for long video understanding, 2026. URL https://arxiv.org/abs/2603.20180.

Samir Khuller, Anna Moss, and Joseph S. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999. doi: 10.1016/S0020-0190(99)00031-9.

Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 912–920, Los Angeles, California, 2010. Association for Computational Linguistics. URL https://aclanthology.org/N10-1134/.

Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 510–520. Association for Computational Linguistics, 2011. URL https://aclanthology.org/P11-1052/.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, 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*, 2025. URL https://arxiv.org/abs/2411.13093.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding, 2024. URL https://arxiv.org/abs/2406.12846.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978. doi: 10.1007/BF01588971.

Qwen Team. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Xiaoqian Shen, Wenxuan Zhang, Jun Chen, and Mohamed Elhoseiny. Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.14032.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP^3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2501.02885.

Maxim Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004. doi: 10.1016/S0167-6377(03)00062-2.

Xichen Tan, Yunfan Ye, Yuanjing Luo, Qian Wan, Fang Liu, and Zhiping Cai. Rag-adapter: A plug-and-play rag-enhanced framework for long video understanding, 2025. URL https://arxiv.org/abs/2503.08576.

Jing Tang, Xueyan Tang, Andrew Lim, Kai Han, Chongshou Li, and Junsong Yuan. Revisiting modified greedy algorithm for monotone submodular maximization with a knapsack constraint. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(1), 2021. doi: 10.1145/3447386. URL https://doi.org/10.1145/3447386.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025. URL https://arxiv.org/abs/2502.21271.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024. URL https://arxiv.org/abs/2403.10517.

Xidong Wang, Dingjie Song, Shunian Chen, Junying Chen, Zhenyang Cai, Chen Zhang, Lichao Sun, and Benyou Wang. LongLLaVA: Scaling multi-modal LLMs to 1000 images efficiently via a hybrid architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21419–21436. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-emnlp.1168. URL https://aclanthology.org/2025.findings-emnlp.1168/.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling, 2025b. URL https://arxiv.org/abs/2501.12386.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2506.06097.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. Symphony: A cognitively-inspired multi-agent system for long-video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2603.17307.

Ruoliu Yang, Chu Wu, Caifeng Shan, Ran He, and Chaoyou Fu. Videodetective: Clue hunting via both extrinsic query and intrinsic relevance for long video understanding, 2026. URL https://arxiv.org/abs/2603.22285.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2601.13719.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. URL https://arxiv.org/abs/2506.22139.

Xian Zhang, Zexi Wu, Zinuo Li, Hongming Xu, Luqi Gong, Farid Boussaid, Naoufel Werghi, and Mohammed Bennamoun. Adard-key: Adaptive relevance-diversity keyframe sampling for long-form video understanding, 2025b. URL https://arxiv.org/abs/2510.02778.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2024. URL https://arxiv.org/abs/2410.02713.

Junjie Zhou et al. Mlvu: A comprehensive benchmark for multi-task long video understanding. In *CVPR*, 2025.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding, 2025. URL https://arxiv.org/abs/2510.27280.

A Reproducibility and implementation details

This section expands the setup in §5.1. It gives the evaluation protocol, evidence-cost checks, and implementation details for the text and frame selectors. Extended results follow in App. B, and component analyses in App. C.

A.1 Evaluation protocol

Additional benchmark details. Dataset sizes and scoring are given in §5.1. VideoMME (Fu et al., 2025) uses four answer options per question. LongVideoBench (Wu et al., 2024) covers 753 validation videos in four duration bins from 15 s to one hour; 73% of questions fall in the two longest bins. MLVU (Zhou et al., 2025) uses the development split with 1,122 videos; its two generation tasks are excluded.

Budget and latency accounting. The evidence cost in Eq. (1) is applied to the final prompt. Intermediate calls in an agentic method contribute to latency, but not to this final-evidence cost. Competitor logs do not provide the per-item text needed to recompute tokenizer-specific costs for every method. The matched-budget and native-footprint comparisons therefore use the same defined evidence-cost unit.

Excluded methods. We exclude methods whose published pipelines require substantially more visual input than the controlled setting permits. For example, VGENT (Shen et al., 2025) constructs its graph from approximately 1,280 frames and supplies up to 320 frames to the answering model. A 64-frame comparison would require changes to that pipeline.

Extraction and retrieval models. ASR uses Whisper-large-v3, OCR uses EasyOCR, text retrieval uses Contriever, and frame relevance uses CLIP-ViT-L/14-336. Detection is disabled in the main configuration.

Hardware and analysis seeds. Experiments use two NVIDIA Ada-generation GPUs with $46\,\mathrm{GB}$ of VRAM each. Fixed seeds determine the analysis subsets; this sampling is separate from RCES inference. The paired tests in App. D assess question-level correctness, not variation across repeated runs or hardware.

A.2 Evidence-cost validation

Selected evidence re-tokenized with the LLaVA-Video tokenizer has Pearson correlations of $r = 0.95$ on VideoMME and $r = 0.99$ on MLVU with the cost in Eq. (1). These correlations show an association with token counts; the enforced budget remains in the defined evidence-cost units. On MLVU, 88.6% of distilled pools fit within $B_T = 800$, with mean budget utilization of 0.49.

A.3 Distillation filtering and clustering

This subsection gives the implementation details of §4.2. Filtering removes empty or punctuation-only strings, OCR fragments matching domain names or watermarks, Whisper repetition loops, bare detection strings such as "frame N", and question-independent filler text.

Retrieval scores are min–max normalized within each source. If the score range is below 10^{-9}, all scores for that source are set to one. Within each temporal-bin and source group, candidates are processed in order. An item joins the cluster with the most similar current centroid if its Contriever cosine similarity meets $\theta_{\mathrm{distill}}$; otherwise, it starts a new cluster. On insertion, the previous normalized centroid is weighted by the cluster size, combined with the new embedding, and normalized again.

After clustering, the mean of the member embeddings is recomputed. The member closest to this mean in cosine similarity is retained as the medoid. It inherits the largest normalized relevance score in the cluster. We use $\theta_{\mathrm{distill}} = 0.85$.

Table 4: **System-level gains across five backbones** ($B_T = 800$, $B_V = 64$). Cells show backbone-only → RCES accuracy (M-Avg for MLVU). The baseline uses uniform frames without retrieved text; RCES settings are fixed across benchmarks. Green marks an increase.

Backbone	VideoMME	LongVideoBench	MLVU
Qwen2.5-VL-7B	61.15 → 65.33 (+4.18)	59.24 → 61.33 (+2.09)	63.90 → 64.96 (+1.06)
LLaVA-Video-7B	64.15 → 68.30 (+4.15)	57.07 → 59.39 (+2.32)	66.09 → 68.19 (+2.10)
InternVideo2.5-8B	60.89 → 67.90 (+7.01)	60.81 → 62.83 (+2.02)	68.77 → 69.94 (+1.17)
InternVL2.5-8B	63.67 → 66.67 (+3.00)	61.20 → 63.13 (+1.93)	68.65 → 69.24 (+0.59)
LongLLaVA-7B	53.33 → 59.30 (+5.96)	49.36 → 50.19 (+0.83)	54.20 → 56.21 (+2.01)

A.4 Visual allocation details

The allocator in §4.2 retains the whole frame bank if $|\mathcal{E}_V| \leq B_V$. Otherwise, it selects B_A uniform anchors and ranks packet centers by CLIP similarity to the question. A center must be at least δ bank positions from every previously accepted center. Each packet contains L consecutive frames; packets near a boundary are shifted to stay within the bank.

At most P packets are accepted. Packet insertion stops when fewer than L budget slots remain. Overlapping frames count once, and the highest-scoring unused frames fill any remaining slots. The output is ordered chronologically and contains $\min(B_V, |\mathcal{E}_V|)$ frames. For the main $B_V = 64$ configuration, the candidate bank contains 192 frames, with $B_A = 16$ uniformly spaced anchors, at most $P = 12$ packets of $L = 4$ consecutive frames, and separation $\delta = 6$ bank positions.

B Extended benchmark results

These results expand §5.2. We first compare RCES with the corresponding backbone-only baseline, then report duration-level results and additional MLVU metrics.

B.1 Comparison across five backbones

Table 4 reports the combined effect of retrieval, text selection, and frame allocation. It does not isolate the textual objective; that comparison is in App. C.1.

B.2 Results by video duration

Table 5 breaks down the VideoMME results by duration. Tables 6 and 7 give the corresponding LongVideoBench breakdowns for RCES and its retrieval baselines.

Table 5: VideoMME duration splits for the five backbones of Table 4 (B_T=800, 64 frames). Latency is mean (median) seconds per question. Column-best values are in bold.

Backbone	Overall	short	medium	long	lat. (s)
Qwen2.5-VL-7B	65.33	76.11	65.11	54.78	17.3 (7.5)
LLaVA-Video-7B	**68.30**	**79.11**	66.67	**59.22**	15.2 (4.9)
InternVideo2.5-8B	67.90	77.33	**67.44**	59.00	**9.7 (3.0)**
InternVL2.5-8B	66.67	77.44	64.89	57.89	16.0 (5.6)
LongLLaVA-7B	59.30	66.56	57.78	53.56	17.9 (7.5)

Table 6: LongVideoBench duration bins for the five backbones of Table 4 (validation, B_T=800, 64 frames). Bins are named by their upper bound in seconds. Cost follows Eq. (1); latency is mean (median) seconds per question.

Backbone	Overall	15 s	60 s	600 s	3600 s	Cost	lat. (s)
Qwen2.5-VL-7B	61.33	**75.66**	75.00	61.17	52.48	563	23.4 (20.4)
InternVideo2.5-8B	62.83	74.60	74.42	**64.08**	54.43	563	**19.7** (**18.8**)
InternVL2.5-8B	**63.13**	74.07	**77.33**	63.59	**54.82**	563	21.9 (21.1)
LLaVA-Video-7B	59.39	69.31	74.42	59.71	51.24	562	22.9 (19.5)
LongLLaVA-7B	50.19	61.38	61.63	49.76	43.26	563	24.0 (23.0)

Table 7: LongVideoBench duration bins for the competitors of Table 2 (LLaVA-Video-7B, validation, 64 frames). Competitors run at their native evidence footprint. Cost follows Eq. (1).

Method	Overall	15 s	60 s	600 s	3600 s	Cost	lat. (s)
Ours (B_T=800)	59.39	69.31	**74.42**	59.71	51.24	**562**	22.9 (19.5)
Video-RAG	**59.91**	70.90	72.67	**60.44**	**51.95**	1907	**18.0** (**17.5**)
TV-RAG	59.16	**71.96**	73.84	58.50	50.89	924	23.6 (20.5)
VideoDetective	47.20	58.20	62.21	47.82	38.48	1136	144.3 (123.6)

B.3 Additional MLVU metrics

Table 2 gives M-Avg and latency for the retrieval methods. With LLaVA-Video-7B, micro accuracy is 70.65 for RCES, 69.55 for Video-RAG, 69.41 for TV-RAG, and 62.47 for VideoDetective. Micro accuracy weights questions equally, whereas M-Avg weights the seven tasks equally. The paired tests in App. D use micro accuracy.

C Additional ablations and sensitivity analyses

This section expands §5.3. We examine the textual objective, budget and weight sensitivity, visual allocation, candidate processing, and optimization quality. The controlled text-selection comparisons use the facility-location objective and uniform frames. The alternative penalty formulation is discussed separately below.

C.1 Removing individual objective terms

Table 8 evaluates the full objective and four single-term removals on the same 2,700 VideoMME questions, including 900 long-video questions, with Qwen2.5-VL-7B, $B_T = 800$, $B_V = 64$, and uniform frames. Each variant sets only the named term's weight to zero; the candidate pool and other settings remain fixed. The full objective answers 1,721 questions correctly, compared with 1,685 without relevance, 1,695 without representativeness, 1,702 without temporal coverage, and 1,706 without source coverage. These correspond to overall drops of 1.33, 0.96, 0.70, and 0.56 percentage points, respectively. The long-split drops are 1.78, 1.22, 0.89, and 0.67 points. Thus, each removal has a larger measured effect on the long split than overall in this evaluation.

These comparisons support the contribution of each term within the full objective. They do not establish submodularity of answer accuracy: Prop. 1 concerns the surrogate F as a function of the selected evidence set. Removing λ_{Rel} leaves the initial question-conditioned retrieval intact and removes relevance only from final selection.

Table 8: Per-term ablation on VideoMME (Qwen2.5-VL-7B, $B_T = 800$, $B_V = 64$, uniform frames). Each ablated column sets one weight to zero. Accuracy is in percent; Δ is the change from the full objective in percentage points, computed from correct-answer counts before rounding.

	Full objective	$-\lambda_{\text{Rel}}$ (relevance)	$-\lambda_{\text{FL}}$ (repr.)	$-\lambda_T$ (temporal)	$-\lambda_{\text{src}}$ (source)
Correct / 2,700	1,721	1,685	1,695	1,702	1,706
Overall	**63.74**	62.41	62.78	63.04	63.19
Δ	—	−1.33	−0.96	−0.70	−0.56
Correct / 900	490	474	479	482	484
Long	**54.44**	52.67	53.22	53.56	53.78
Δ long	—	−1.78	−1.22	−0.89	−0.67

C.2 Alternative redundancy penalty

An alternative to facility location subtracts a redundancy penalty:

$$F_{\text{pen}}(S) = \lambda_{\text{Rel}}\text{Rel}(S) + \lambda_T\text{CovTime}(S) + \lambda_{\text{src}}\text{CovSrc}(S) - \lambda_R\text{Red}(S),$$
$$\text{Red}(S) = \sum_{e \in S} \max_{e' \in S \setminus \{e\}} \text{sim}(e, e'). \tag{10}$$

Here $\max \emptyset = 0$ and $\lambda_R \geq 0$ controls the redundancy penalty. This objective can decrease when an item is added, so the guarantee in Remark 1 does not apply. An implementation that stops at nonpositive marginal gains may also leave budget unused. Such behavior confounds comparisons with the monotone objective if both the objective and stopping behavior change.

Equivalence at zero weight. Setting $\lambda_R = 0$ in Eq. (10) and $\lambda_{\mathrm{FL}} = 0$ in Eq. (7) gives the same objective: $\lambda_{\mathrm{Rel}}\mathrm{Rel} + \lambda_T \mathrm{CovTime} + \lambda_{\mathrm{src}}\mathrm{CovSrc}$. For the same pool, costs, tie-breaking, and stopping rule, the two selectors therefore return the same subset. This equivalence applies only at zero weight; it does not justify transferring other ablations between the two formulations.

C.3 Objective-weight sensitivity

The sweep varies one weight at a time: each coverage weight takes values in $\{0, 0.1, 0.2, 0.4\}$, and the representativeness weight takes values in $\{0, 0.2, 0.3, 0.5\}$. Together with the shared baseline, these give ten configurations. The stratified analysis subset contains 421 questions from 141 videos, with 47 videos per duration split. It is drawn from the evaluation set, so the sweep is an analysis of sensitivity rather than an independent evaluation on held-out data.

Overall accuracy ranges from 62.23 to 62.71. The source-weight sweep has a 0.00-point range and the temporal-weight sweep a 0.47-point range. The representativeness-weight sweep has a 2.13-point range on long videos. This range is a different comparison from the single-term removal on all 2,700 questions in Table 8.

C.4 Text-budget sensitivity

The detection-off sweep for the final configuration is reported in Table 3. A separate detection-enabled sweep reaches 63.3, 64.8, and 65.1 overall accuracy at $B_T \in \{500, 800, 1000\}$. Detection remains enabled throughout this sweep. We report it separately from the final configuration because the two sweeps use different configurations and do not isolate the effect of detection.

C.5 Frame budgets and packet layouts

At $B_V = 32$, packets reach 64.30 overall accuracy, compared with 63.00 for uniform sampling. The long-split difference is significant ($p = 0.026$), while the overall difference is not ($p = 0.13$). At $B_V = 64$, the main comparison instead shows a significant overall difference and no significant long-split difference (Table 10). These results support a benefit in some settings, but do not establish that it grows as the frame budget shrinks. On MLVU, packets reach 68.19 M-Avg versus 67.67 for uniform frames; the associated question-level test gives $p = 0.407$.

At $B_V = 64$, a separate layout sweep compares 16 anchors and 12 four-frame packets with two alternatives: 16 anchors and 24 two-frame packets, and 32 anchors and eight four-frame packets. The alternatives are 0.27 and 0.42 accuracy points lower within that sweep. The layout experiments were run on all 2,700 VideoMME questions. These figures are reported separately from the main-table results. The $B_V = 32$ configuration uses eight anchors and four six-frame packets.

Question categories and frame packets. In the category breakdown, Object Reasoning has the largest gain among categories with more than 100 questions: $62.33 \rightarrow 66.30$ ($n = 454$). Temporal Perception rises from 54.55 to 67.27, while Spatial Perception falls by 7.41 points. These observations motivate localization as a possible explanation for packet gains; they do not directly test that mechanism.

C.6 Distillation and candidate loss

Distillation reduces the pool searched by the selector using the rules in App. A.3. A medoid may omit an answer-bearing detail present in another cluster member. Pool compression therefore does not by itself establish that all useful information is preserved. Preserving such details remains a limitation of this stage.

C.7 OBJECT-DETECTION LATENCY

Table 9 supports the detection discussion in §5.3. It compares per-question latency with detection enabled and disabled on long videos, broken down by pipeline stage. The reported difference is detection off minus detection on.

Table 9: Per-question latency on the long split, isolating the cost of object detection. Enabling detection adds ∼72 s: detector calls plus the CLIP frame scoring they trigger.

Stage	detection on (B_T=800)	detection off	Δ
total	112.5	40.0	−72.5
preprocess (CLIP frame scoring)	65.7	31.9	−33.8
retrieve & select (detector calls)	40.2	1.5	−38.7
final inference	6.6	6.6	0

C.8 GREEDY VERSUS EXACT OPTIMIZATION

We compare the greedy objective value with an exact branch-and-bound solution of Eq. (7). Both optimize the same objective over the same distilled pool. The analysis includes 220 questions for which the solver establishes an exact optimum over the unrestricted distilled pool. Cases with a restricted pool or an unfinished solver run are excluded.

The mean ratio $F(S_{\text{greedy}})/F(S_{\text{exact}})$ is 0.9994. Greedy is exactly optimal on 98.2% of these questions and within 1% on 99.5%; the minimum ratio is 0.8805. This measures optimization quality on those pools and does not measure answer accuracy.

The solver starts from the greedy solution. For a current set S and remaining individually affordable candidates R, monotonicity gives $F(S') \leq F(S \cup R)$ for any feasible completion S' of S. This provides an admissible pruning bound. The bound was checked against exhaustive enumeration on small pools.

D STATISTICAL TESTING DETAILS

This section supports the significance statements in §5.2 and the component comparisons in App. C. Each test pairs predictions for the same question and uses an exact, two-sided McNemar test on correctness. For MLVU, the tested difference is micro accuracy; M-Avg weights tasks equally and is a different quantity.

The upper block of Table 10 covers Qwen2.5-VL-7B on VideoMME. The lower block covers LLaVA-Video-7B: matched-budget VideoMME results and native-footprint LongVideoBench and MLVU results. For the main frame-packets comparison, 147 questions change from incorrect to correct and 104 change from correct to incorrect. These 251 discordant pairs give the reported overall $p = 0.008$. On the long split, the corresponding counts are 38 and 35.

E THEORETICAL PROOFS

The empirical analyses above evaluate accuracy and selection behavior. Here we prove the structural properties used by the optimizer in §4.1. We use the objective and component definitions from Eqs. (2)–(7), without redefining them.

E.1 NOTATION AND ASSUMPTIONS

Let $\mathcal{E} = \mathcal{E}_T$ be a finite, nonempty distilled pool and $S \subseteq \mathcal{E}$. Relevance scores and objective weights are nonnegative, and costs are positive. Similarities use the nonnegative clamp in Eq. (6), with the empty-set maximum defined as zero. If the pool is empty, we define $\mathrm{FL}(\emptyset) = F(\emptyset) = 0$ and return the empty set. For a set function $f : 2^{\mathcal{E}} \to \mathbb{R}$, the marginal gain of e at S is $\Delta_f(e \mid S) = f(S \cup \{e\}) - f(S)$. We call f *monotone* if $\Delta_f(e \mid S) \geq 0$ for every $S \subseteq \mathcal{E}$ and $e \in \mathcal{E} \setminus S$, and *submodular* if $\Delta_f(e \mid A) \geq \Delta_f(e \mid B)$ whenever $A \subseteq B$ and $e \notin B$ (diminishing returns).

Table 10: Exact two-sided McNemar tests on paired correctness. Δ is the first configuration's accuracy minus the second's, in percentage points. *MLVU differences are micro accuracy, not M-Avg.

Qwen2.5-VL-7B: VideoMME

	Overall		Long	
Comparison	Δ	p	Δ	p
Frame packets vs. uniform frames	$+1.59$	0.008	$+0.33$	0.82
Frame packets vs. uniform (B_V=32)	$+1.30$	0.13	$+2.44$	0.026
Ours vs. TV-RAG (budgeted)	$+1.29$	0.109	$+1.11$	0.52

LLaVA-Video-7B: budgeted VideoMME; native LVB/MLVU

	VideoMME		LongVideoBench		MLVU*	
Comparison	Δ	p	Δ	p	Δ	p
Ours vs. Video-RAG	$+4.52$	$< 10^{-4}$	-0.52	0.715	$+1.10$	0.239
Ours vs. TV-RAG	$+0.82$	0.16	$+0.22$	0.900	$+1.24$	0.163
Ours vs. VideoDetective	$+8.67$	$< 10^{-16}$	$+12.19$	$< 10^{-12}$	$+8.19$	$< 10^{-12}$

E.2 Component properties

Relevance.

Lemma 1. Rel *is monotone and modular, hence submodular.*

Proof. For any S and $e \notin S$, $\Delta_{\mathrm{Rel}}(e \mid S) = r(e)$, independent of S. Hence $\Delta_{\mathrm{Rel}}(e \mid A) = \Delta_{\mathrm{Rel}}(e \mid B)$ (modular, so submodular), and $\Delta_{\mathrm{Rel}}(e \mid S) = r(e) \geq 0$ (monotone). □

Temporal and source coverage.

Lemma 2. CovTime *and* CovSrc *are monotone and submodular.*

Proof. Both have the form $g(S) = |\{\phi(e) : e \in S\}|$ for an attribute map ϕ ($\phi = \mathrm{bin}$ for CovTime, $\phi = \mathrm{src}$ for CovSrc). Writing $\Phi(S) = \{\phi(s) : s \in S\}$, the marginal gain is

$$\Delta_g(e \mid S) = \mathbf{1}[\phi(e) \notin \Phi(S)] \in \{0, 1\},$$

so g is monotone. If $A \subseteq B$ then $\Phi(A) \subseteq \Phi(B)$, so an item that adds a new attribute to B also adds one to A; hence $\Delta_g(e \mid A) \geq \Delta_g(e \mid B)$ and g is submodular. □

Facility location.

Lemma 3. FL *is monotone and submodular.*

Proof. Write $\mathrm{FL}(S) = \frac{1}{|\mathcal{E}|} \sum_{x \in \mathcal{E}} f_x(S)$ with $f_x(S) = \max_{s \in S} \mathrm{sim}^+(x, s)$; it suffices that each f_x be monotone submodular. Fix x, let $A \subseteq B$, $e \notin B$, and put $a_x = f_x(A) \leq f_x(B) = b_x$. Since $f_x(S \cup \{e\}) = \max\{f_x(S), \mathrm{sim}^+(x, e)\}$,

$$\Delta_{f_x}(e \mid A) = \max\{0, \mathrm{sim}^+(x, e) - a_x\} \;\geq\; \max\{0, \mathrm{sim}^+(x, e) - b_x\} = \Delta_{f_x}(e \mid B),$$

the inequality using $a_x \leq b_x$; so f_x is submodular, and monotone since $\Delta_{f_x}(e \mid S) \geq 0$. A nonnegative average of monotone submodular functions is monotone submodular, hence so is FL. □

E.3 Proof of the main proposition

Proposition 2. *Under the assumptions of Proposition 1, the objective F in Eq. (7) is normalized, monotone nondecreasing, and submodular.*

Proof. Each of Rel (Lemma 1), CovTime, CovSrc (Lemma 2), and FL (Lemma 3) is monotone submodular and vanishes at $\emptyset$. A nonnegative weighted sum of monotone submodular functions is monotone submodular, and a sum of zeros is zero; hence F is normalized, monotone, and submodular. □

E.4 BUDGETED OPTIMIZATION CONSEQUENCE

Corollary 1. *The selection problem* $\max_S F(S)$ *subject to* $\sum_{e\in S} c_T(e) \le B_T$ *is monotone submodular maximization under a knapsack constraint, and admits a* $(1-\exp(-1))$ *approximation via partial-enumeration greedy (Sviridenko, 2004).*

Here $c_T(S) = \sum_{s\in S} c_T(s)$ denotes the total cost of S.

Remark 1 (Implemented optimizer)**.** *Alg. 1 uses density greedy with a best-singleton safeguard. For a nonempty candidate pool, its objective value has the* 0.405 *guarantee of modified greedy (Tang et al., 2021, Theorem 2.1). Selecting directly among feasible items is equivalent to discarding infeasible items in the cited algorithm, since positive costs make infeasibility permanent as* S *grows. If all remaining feasible marginal gains are zero, monotonicity and submodularity imply that any feasible completion has the same objective value; early stopping therefore preserves the guarantee. An empty pool or the absence of feasible items yields the optimal empty set. This guarantee applies to* F *over the fixed distilled pool, not to the ideal utility* U_q *or answer accuracy.*