FROM RELEVANCE TO REPRESENTATIVES: EQUAL-RATE EVIDENCE QUANTIZATION FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long video understanding often requires selecting useful visual observations under a limited frame budget. Uniform sampling distributes observations by elapsed time, whereas global relevance ranking can concentrate them in locally redundant high-scoring regions. In this paper, we propose Equal-rate Evidence Quantization ($\mathrm{E}^2\mathrm{Q}$), a training-free frame selector for long video understanding that separates where to allocate the budget from which observation to retain. $\mathrm{E}^2\mathrm{Q}$ uses normalized frame–question relevance to allocate the frame budget, partitioning the ordered candidates into consecutive intervals with approximately equal relevance mass and retaining the highest-scoring frame from each interval. A budget-dependent projection handles overly concentrated candidate weights before partitioning, enabling relevance-adaptive intervals without an additional relevance–coverage coefficient. We characterize this construction and bound frame-level representation error in terms of within-interval variation and interval-mass imbalance. Across four long-video question answering benchmarks and three models, $\mathrm{E}^2\mathrm{Q}$ achieves the highest accuracy among training-free selectors in ten of the twelve model–benchmark settings. Controlled shared-score comparisons further show consistent gains when the relevance signal is fixed and isolate the respective contributions of relevance-mass allocation and within-interval representative selection to these gains.

1 INTRODUCTION

A long video may contain thousands of candidate frames, while a Multimodal Large Language Model (MLLM) (Bai et al., 2025; Zhu et al., 2025; Li et al., 2025) can process only a small fraction of them within a practical compute budget. To cope with this limit, existing methods extend the context window (Zhang et al., 2025b; Chen et al., 2025), compress visual tokens (Li et al., 2024; Shen et al., 2025), encode videos in segments (Weng et al., 2024), select among answers from multiple sampled frame sets (Suo et al., 2025), or search videos with agents (Wang et al., 2024; Yuan et al., 2025). Frame selection complements them by determining which observations are available to the model for downstream reasoning. A brief action, an object in the background, or a change of state can be lost before reasoning begins. Selecting useful observations is therefore an important part of long-video Question Answering (QA) on benchmarks such as Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2025), and LVBench (Wang et al., 2025a).

Two simple strategies expose the allocation problem: 1) uniform sampling, common in video MLLMs (Li et al., 2025; Zhu et al., 2025), distributes frames across the timeline, including regions unrelated to the question; 2) global relevance ranking, a standard baseline (Tang et al., 2025; Liu et al., 2025), instead concentrates the budget on high-scoring frames, but nearby frames often receive similar scores, so the selected set can contain repeated views of the same moment. Figure 1 illustrates this trade-off. For a question asking where a person leaves a key, uniform sampling can miss the brief placement, while global ranking can spend several slots on nearby views of the person holding the key. The resulting problem is one of allocation: how should a small frame budget be distributed toward question-relevant observations while preserving enough temporal coverage to capture distinct evidence without spreading it uniformly over time or collapsing it onto a redundant high-score region?

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.

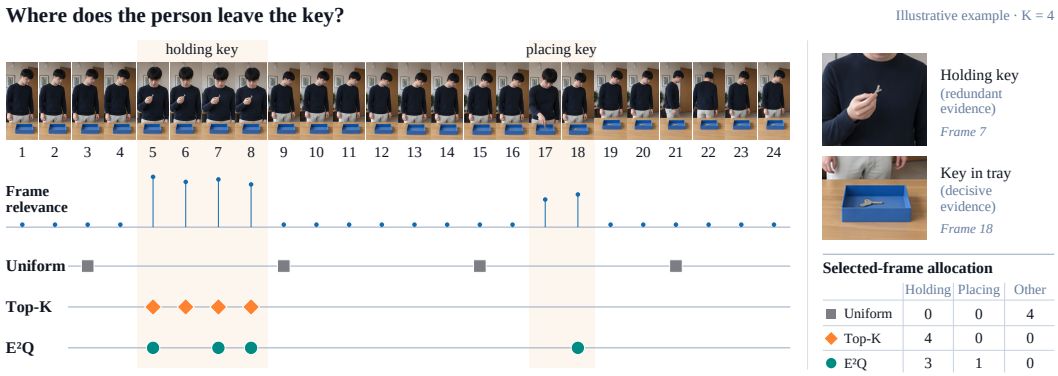


Figure 1: **Illustrative frame allocation with a four-frame budget.** Uniform sampling (Frames 3, 9, 15, and 21) misses the holding and placement events, whereas Top-K spends all four slots on the holding region (Frames 5–8). E^{2}Q distributes the budget across the relevance pattern, retaining representatives from the holding region and the later placement event (Frames 5, 7, 8, and 18).

Existing selectors already combine relevance with coverage, but they make different choices about what to do once frame relevance has been estimated. AKS (Tang et al., 2025) recursively partitions the timeline before selecting frames, while RIDGE (Xu et al., 2026) first interprets the relevance curve through temporal roles and then allocates observations across them. MDP^3 (Sun et al., 2025) instead complements relevance with diversity among the selected frames. These approaches improve coverage under the same fixed observation budget by introducing temporal structure or an additional selection criterion beyond the relevance scores themselves. BOLT (Liu et al., 2025) takes a more direct route. It uses cumulative relevance to map regularly spaced relevance levels to individual frames through inverse transform sampling, showing that the frame budget can be distributed in relevance space rather than only along the time axis. If the relevance distribution already indicates where question-relevant content is concentrated, the budget can in principle be allocated from that distribution directly, without first converting it into an explicit temporal structure. Cumulative relevance provides such an allocation signal, but directly using it for frame selection introduces a separate issue: the frame at each cumulative relevance level is selected because of where it lies in the distribution, not necessarily because it is the highest-scoring frame in that part of the video. Where an observation slot is allocated and which frame fills that slot are therefore related but distinct decisions.

To this end, we introduce Equal-rate Evidence Quantization (E^{2}Q), guided by a simple rule: *allocate approximately equal shares of relevance, then retain the highest-scoring observation within each share*. Under a fixed frame budget, E^{2}Q partitions the ordered candidates into consecutive intervals with approximately equal relevance mass and selects the highest-scoring frame from each interval. This keeps allocation tied to relevance while leaving representative choice strictly local. Because relevance can be distributed unevenly over time, the resulting intervals adapt in temporal extent, concentrating the budget where question-relevant content is dense while preserving broader coverage across the relevance distribution. A budget-dependent projection handles overly concentrated candidate weights before partitioning. Once the candidates and relevance scores are fixed, E^{2}Q requires only the specified frame budget, with no additional relevance–coverage coefficient or model training.

Our contributions are summarized as follows:

- We formulate fixed-budget frame selection as capacity-constrained quantization (Gray & Neuhoff, 1998) of the query-conditioned relevance distribution. E^{2}Q uses approximately equal relevance mass to allocate observation slots and selects the representative frame for each slot separately.
- We characterize the budget-dependent projection and equal-rate interval allocation, and derive a representation bound that separates approximation error into within-interval variation and relevance-mass imbalance, linking the latter directly to the cumulative-boundary objective.
- Across four long-video QA benchmarks and three answering models, E^{2}Q ranks among the strongest training-free selectors. Controlled shared-score comparisons isolate the roles of relevance-mass allocation, the budget-dependent projection, and within-interval representative selection.

2 Related Work

MLLMs for Long Video Understanding. Recent MLLMs such as Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and LLaVA-OneVision (Li et al., 2025) have shown strong video understanding capabilities, and several lines of work extend them to long videos. Long-context approaches transfer the long context of the language backbone to vision without video training (Zhang et al., 2025b) or scale long-context training and inference (Chen et al., 2025). Token-reduction approaches prune visual tokens after an early layer (Chen et al., 2024), represent each frame with two tokens (Li et al., 2024), consolidate tokens in memory (Song et al., 2024), adaptively remove spatiotemporal redundancy (Shen et al., 2025), or merge tokens hierarchically within short segments (Weng et al., 2024). Other approaches spend more computation at inference time, choosing among answers from several sampled frame subsets with a self-reward (Suo et al., 2025) or letting an agent ground sub-questions to video moments and inspect them iteratively (Yuan et al., 2025).

Query-Aware Frame Selection. Pretrained vision–language models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023) provide frame–question relevance signals for training-free selection. Recent methods combine these signals with different forms of coverage. AKS (Tang et al., 2025) balances relevance and coverage through recursive splitting, while MDP^3 (Sun et al., 2025) incorporates visual diversity and sequential allocation. Q-Frame (Zhang et al., 2025c) combines query-aware selection with adaptive image resolution, FOCUS (Zhu et al., 2026) explores promising temporal clips under a limited budget, RIDGE (Xu et al., 2026) allocates frames across temporal roles inferred from the relevance curve, and WFS-SB (Chen et al., 2026) uses semantic boundaries to guide allocation. Other methods search for keyframes iteratively, using a language-model agent (Wang et al., 2024), a query-adaptive tree representation (Wang et al., 2025b), or temporal search recast as spatial search (Ye et al., 2025). Learned selectors instead rely on supervision or downstream model signals, such as keyframe pseudo-labels from an answering model (Yu et al., 2023), learned rankings of frame combinations (Yu et al., 2025), a trained MLLM-based selector (Hu et al., 2025), or reinforcement learning (Qin et al., 2026). A related line uses the relevance distribution itself to guide sampling: BOLT (Liu et al., 2025) and the sampler of KFS-Bench (Li et al., 2026) select frames at evenly spaced points of a cumulative distribution, and KFS-Bench also uses equal-mass bins to analyze temporal spread. $\mathrm{E}^2\mathrm{Q}$ instead uses equal-mass intervals to allocate slots and then fills each slot with its highest-scoring frame.

3 Methodology

3.1 Problem Formulation and Overview

Given a question u with n distinct candidate frames $V = (x_1, \ldots, x_n)$ in chronological order, we select K frames for a video MLLM, for integer K satisfying $1 \leq K \leq n$. A pretrained image–text scorer provides a query-conditioned positive relevance score $r_i = g(x_i, u)$ for each candidate. These scores guide selection while the answering model remains unchanged. We normalize them as:

$$q_i = \frac{r_i}{\sum_{\ell=1}^{n} r_\ell}. \tag{1}$$

These weights define a relevance distribution over the candidates. The sum of weights over a consecutive set of candidates is its *relevance mass*. Cumulative mass provides the allocation signal. It measures accumulated question relevance, not the number of distinct events or elapsed time.

$\mathrm{E}^2\mathrm{Q}$ uses this distribution to separate allocation from representative selection. With K available observation slots, each slot is assigned a target relevance mass of $1/K$, which we refer to as equal-rate allocation. The ordered candidates are then divided into K consecutive intervals whose masses approximate these targets, with one representative selected from each interval. Because relevance may be distributed unevenly over time, the resulting intervals can differ in temporal extent.

Concretely, $\mathrm{E}^2\mathrm{Q}$ first projects the relevance distribution onto a budget-dependent capped simplex so that no candidate receives more than one slot's target mass before interval boundaries are placed. It then places interval boundaries near the cumulative-relevance levels j/K and retains the highest-scoring frame from each resulting interval, as illustrated in Figure 2. The selected frames are finally passed in chronological order, together with the question, to the unchanged downstream model.

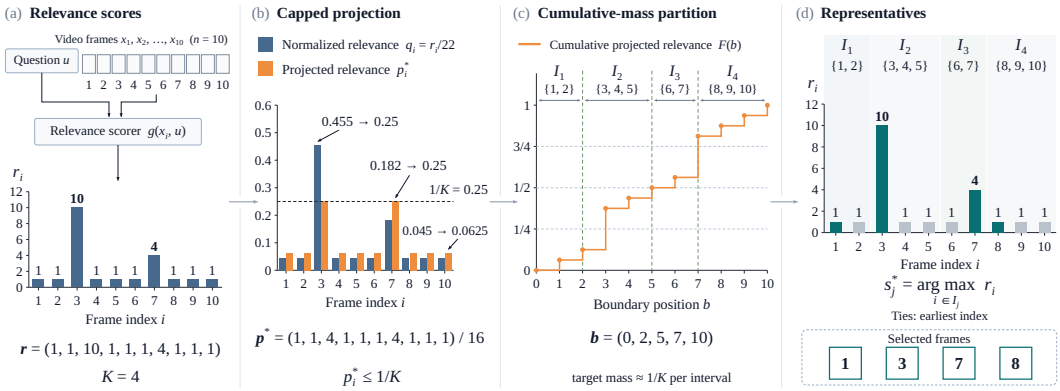


Figure 2: **Overview of** $\mathrm{E}^2\mathrm{Q}$ **with** $K = 4$**.** (a) A question-conditioned scorer produces raw frame relevance scores. (b) The normalized relevance distribution is projected onto the budget-dependent capped simplex. (c) Cumulative projected relevance determines K consecutive allocation intervals. (d) The highest-scoring candidate within each interval is retained as its representative.

3.2 Capping Concentrated Relevance

Each interval targets mass $1/K$. If a candidate has $q_i > 1/K$, a single jump in cumulative relevance can span multiple target levels. Choosing the nearest boundary independently for each level can then map multiple targets to the same cut, leaving an interval empty. $\mathrm{E}^2\mathrm{Q}$ therefore caps each candidate's mass at $1/K$ before partitioning. This keeps each cumulative jump within a single target share.

We define the projected distribution $p^\star$ as the KL projection of q onto the capped simplex $\mathcal{P}_K$:

$$p^\star = \arg\min_{p \in \mathcal{P}_K} D_{\mathrm{KL}}(p \| q), \qquad \mathcal{P}_K = \left\{ p \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1,\ p_i \le \frac{1}{K} \right\}. \tag{2}$$

The unique solution is:

$$p_i^\star = \min\left(c q_i, \frac{1}{K} \right), \qquad \sum_i p_i^\star = 1, \tag{3}$$

where $c > 0$ makes the projected weights sum to one. This water-filling rule (Warmuth & Kuzmin, 2008) redistributes excess mass proportionally among uncapped candidates, preserving their relative weights. If all $q_i \le 1/K$, then $p^\star = q$. Appendix A.1 derives this solution, including the case $n = K$.

3.3 Partitioning by Cumulative Relevance

Let $F(b) = \sum_{i=1}^b p_i^\star$, with $F(0) = 0$, denote the cumulative projected relevance mass through candidate b. The target levels for the internal boundaries are j/K, $j = 1, \ldots, K-1$. Because boundaries can only be placed after discrete candidates, these target levels may not be attained exactly.

$\mathrm{E}^2\mathrm{Q}$ chooses integer boundaries that minimize their total deviation from the target levels:

$$(b_1^\star, \ldots, b_{K-1}^\star) \in \arg\min_{0 = b_0 < b_1 < \cdots < b_K = n} \sum_{j=1}^{K-1} \left| F(b_j) - \frac{j}{K} \right|. \tag{4}$$

The intervals are $I_j = \{b_{j-1}^\star + 1, \ldots, b_j^\star\}$, with $b_0^\star = 0$ and $b_K^\star = n$. They target approximately equal relevance mass rather than equal duration. Regions with higher relevance per candidate generally reach successive target levels over fewer candidates, while low-relevance regions can span more candidates. Thus, the allocation adapts to the question while the frame budget remains fixed.

Our implementation solves Equation 4 using a prefix-minimum dynamic program in $O(Kn)$ time, with ties resolved toward earlier predecessors and an earlier final boundary. Appendix B gives the recurrence and implementation details. Under the cap, the earliest feasible boundaries nearest to their targets form a strictly increasing sequence and jointly attain the same optimum in exact arithmetic, as shown in Appendix A.2. This keeps independently fitted targets compatible with a valid partition.

3.4 Selecting Interval Representatives

With one output slot assigned to each interval, $\mathrm{E}^2\mathrm{Q}$ selects a single representative from that interval. Specifically, it retains the candidate with the highest original relevance score:

$$s_j^\star = \arg\max_{i \in I_j} r_i, \qquad j = 1, \ldots, K, \tag{5}$$

with exact ties resolved toward the earlier candidate. The boundaries constrain where a frame can be selected without requiring the boundary candidate itself to be retained. For the resulting intervals, maximizing $p_i^\star$ selects the same representatives as maximizing r_i in exact arithmetic. Corollary 1 in Appendix A.2 provides the proof. Because the intervals are nonempty and disjoint, the procedure returns exactly K distinct frames in chronological order. The current implementation requires $O(Kn)$ time and memory after scoring. No additional threshold or relevance–coverage coefficient is introduced beyond the frame budget K, so no further hyperparameter tuning is required.

4 Representation Analysis

We analyze the approximation induced by representing each interval with one frame. Let f_i be a bounded scalar quantity associated with candidate i, such as an indicator that a queried object is visible. We compare its relevance-weighted average $\sum_i p_i^\star f_i$ with the equal-weight average $K^{-1}\sum_j f_{s_j^\star}$ over the selected representatives. This surrogate characterizes the resulting finite representation under this fixed frame budget and does not assume equal attention weights in the downstream model.

Write $m_j = \sum_{i \in I_j} p_i^\star$ for the relevance mass of interval j, and let $\mathrm{osc}_{I_j}(f) = \max_{i,\ell \in I_j} |f_i - f_\ell|$ denote the within-interval variation of f. Define the cumulative-boundary error $\Delta_{\mathrm{cut}} = \sum_{j=1}^{K-1} |F(b_j^\star) - j/K|$, the total deviation of the boundary masses from their targets, and $\|f\|_\infty = \max_i |f_i|$.

For the analysis, let $\widetilde{s}_j \in I_j$ denote an arbitrary representative of interval j; choosing $\widetilde{s}_j = s_j^\star$ recovers the representatives selected by $\mathrm{E}^2\mathrm{Q}$.

Theorem 1 (Representation error)**.** *For any bounded frame-level quantity f and any representatives $\widetilde{s}_j \in I_j$, $j = 1, \ldots, K$, we have*

$$\left| \sum_{i=1}^{n} p_i^\star f_i - \frac{1}{K} \sum_{j=1}^{K} f_{\widetilde{s}_j} \right| \le \sum_{j=1}^{K} m_j \,\mathrm{osc}_{I_j}(f) + 2\|f\|_\infty \Delta_{\mathrm{cut}}. \tag{6}$$

The first term bounds the error from replacing the candidates in each interval by one observation. It is small when f varies little within the interval. If f also has a bounded rate of change over time, the interval's duration controls this variation up to that rate constant. Appendix A provides the full derivation. This condition concerns the specific quantity represented by f, not the relevance score.

The second term accounts for the difference between an interval's mass m_j and the equal weight $1/K$ assigned to its representative. Let $e_j = F(b_j^\star) - j/K$, with $e_0 = e_K = 0$. Then $m_j - 1/K = e_j - e_{j-1}$, so $\sum_j |m_j - 1/K| \le 2\Delta_{\mathrm{cut}}$. Equation 4 minimizes Δ_{cut} and thus tightens this upper bound on mass imbalance, while the within-interval variation term depends on the video content.

Comparing the selected average with the original distribution q introduces an additional term bounded by $\|f\|_\infty \|q - p^\star\|_1$. Appendix A gives the full proof and a KL-based bound on this term. The theorem concerns the frame-level representation rather than the downstream model's computation or QA accuracy, and it does not favor a particular representative; Section 5.4 therefore compares the highest-scoring choice with temporal centers under controlled inputs.

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. We evaluate on the full question sets of four long-video QA benchmarks: Video-MME (Fu et al., 2025) (VMM; 2,700 questions), the validation set of LongVideoBench (Wu

et al., 2024) (LVB; 1,337 questions), the development set of MLVU (Zhou et al., 2025) (2,174 questions), and LVBench (Wang et al., 2025a) (1,549 questions). Video-MME is evaluated without subtitles. Unless otherwise stated, results are reported as accuracy (%) under each benchmark's official aggregation; for MLVU, we report the macro-average over its multiple-choice tasks (M-Avg).

Models and frame budget. Unless stated otherwise, all experiments use a frame budget of $K = 32$. The native comparison evaluates three downstream models, Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), and LLaVA-OneVision-7B (Li et al., 2025); controlled analyses use Qwen2.5-VL-7B. $\mathrm{E}^2\mathrm{Q}$ uses BLIP (Li et al., 2022) as its default frame–question relevance scorer.

Baselines. We compare with uniform sampling and six recent training-free frame selectors: AKS (Tang et al., 2025), MDP^3 (Sun et al., 2025), BOLT (Liu et al., 2025), FOCUS (Zhu et al., 2026), RIDGE (Xu et al., 2026), and WFS-SB (Chen et al., 2026). For the native comparison, we rerun their official implementations under a common downstream evaluation pipeline while retaining each selector's default relevance scoring and candidate-preprocessing configuration.

Evaluation protocols. The native comparison fixes the frame budget, downstream model, prompt, decoding, and answer evaluation, while each selector retains its own scoring and candidate preprocessing. The controlled protocol additionally fixes the candidate pool, relevance scores, and visual preprocessing, so that differences more directly reflect the selection rule. Appendix C details the benchmark settings, candidate scoring, inference settings, and controlled selection rules.

5.2 Main Results

Table 1 reports the native comparison. $\mathrm{E}^2\mathrm{Q}$ achieves the highest accuracy in ten of the twelve model–benchmark settings; in the remaining two (MLVU and LVBench with LLaVA-OneVision-7B), it trails the best selector by 0.1 and 0.7 points. It outperforms uniform sampling in all twelve settings, by 1.9–12.8 percentage points, and attains the highest four-benchmark average for all three models, so its gains are not tied to a single downstream model. Its largest margin over the strongest competing selector is 2.6 points (MLVU with Qwen2.5-VL-7B). No single baseline is consistently the strongest: the best competing selector varies among RIDGE, WFS-SB, MDP^3, and FOCUS across settings. For Qwen2.5-VL-7B, gains over WFS-SB are significant on three of the four benchmarks (Appendix D.3).

5.3 Frame-Selection Quality

We evaluate the selected frames directly using the annotations of KFS-Bench (Li et al., 2026), which mark the scenes needed to answer each question and cover 888 Video-MME and 1,291 LongVideoBench questions. At $K = 32$, all selectors, including the controls of Section 5.4, share the same candidate sampling and BLIP scores; scene annotations are used only for evaluation. Table 2 reports Key Frame Rate (KFR), the fraction of selected frames that fall in relevant scenes; Balanced Scene Recall (BSR), scene coverage under the benchmark's sampling thresholds; and the Unified Keyframe Sampling Score (UKSS), which combines these criteria with a distribution-similarity term. Appendix G.1 gives the full metrics, the aggregation rule, and a single- versus multi-scene breakdown.

$\mathrm{E}^2\mathrm{Q}$ achieves the highest KFR and UKSS on both subsets (UKSS 0.3129 and 0.2611). Relative to Top-K, it raises BSR by 0.0780 on Video-MME and 0.0369 on LongVideoBench, covering more relevant scenes while also placing more of its frames in them. Candidate-grid uniform has a slightly higher BSR on Video-MME (0.8035 versus 0.7976) but lower KFR and UKSS. $\mathrm{E}^2\mathrm{Q}$ also improves all three metrics over RIDGE and projected inverse-CDF on both subsets. Because projected inverse-CDF uses the same projected distribution, this comparison supports forming relevance-mass intervals and choosing a representative within each, rather than selecting CDF points directly.

5.4 Controlled Analysis of $\mathrm{E}^2\mathrm{Q}$

To isolate the effects of allocation and representative selection, we compare $\mathrm{E}^2\mathrm{Q}$ with allocation controls and component ablations under the controlled protocol (Table 3; exact rules in Appendix C.2).

Allocation controls. Candidate-grid uniform selects evenly spaced positions in the common candidate list, and Top-K selects the globally highest-scoring candidates. Temporal-bin max forms K nonempty intervals by elapsed time and retains the highest-scoring frame in each; relative to $\mathrm{E}^2\mathrm{Q}$, it changes the allocation basis from cumulative relevance to time while keeping the interval-maximum rule.

Table 1: **Native comparison at $K = 32$.** Accuracy (%) of three downstream models on four long-video benchmarks; Avg. is the unweighted mean. Bold and underlined values denote the best and second-best selectors; small blue numbers give the gain of E^2Q over uniform sampling.

	Selector	Video-MME	LongVideoBench	MLVU	LVBench	Avg.
Qwen2.5-VL-7B	Uniform	61.2	58.9	59.7	38.5	54.6
	AKS	63.1	63.2	67.2	43.4	59.2
	MDP^3	63.8	60.0	66.2	41.3	57.8
	BOLT	63.2	60.0	68.4	39.3	57.7
	FOCUS	63.4	61.0	68.2	46.3	59.7
	RIDGE	63.7	<u>65.4</u>	69.6	<u>50.9</u>	<u>62.4</u>
	WFS-SB	<u>64.2</u>	64.4	<u>69.9</u>	48.1	61.7
	E^2Q **(ours)**	**64.9** +3.7	**65.8** +6.9	**72.5** +12.8	**51.2** +12.7	**63.6** +9.0
InternVL3-8B	Uniform	65.6	58.5	68.4	43.5	59.0
	AKS	66.3	61.5	74.2	52.4	63.6
	MDP^3	<u>66.8</u>	60.9	73.9	47.3	62.2
	BOLT	66.1	59.2	70.8	44.2	60.1
	FOCUS	64.3	61.0	72.0	53.2	62.6
	RIDGE	66.6	61.8	<u>74.6</u>	<u>54.5</u>	<u>64.4</u>
	WFS-SB	66.3	<u>62.5</u>	74.5	53.9	64.3
	E^2Q **(ours)**	**67.5** +1.9	**62.8** +4.3	**74.8** +6.4	**55.1** +11.6	**65.1** +6.1
LLaVA-OneVision-7B	Uniform	56.5	54.8	63.1	40.4	53.7
	AKS	58.4	59.3	67.5	47.3	58.1
	MDP^3	58.5	59.0	**68.2**	41.6	56.8
	BOLT	57.9	59.6	66.8	38.2	55.6
	FOCUS	58.3	<u>60.7</u>	67.1	47.5	58.4
	RIDGE	59.6	60.1	68.0	**48.9**	59.2
	WFS-SB	<u>61.9</u>	60.4	67.9	<u>48.4</u>	<u>59.7</u>
	E^2Q **(ours)**	**62.2** +5.7	**61.6** +6.8	<u>68.1</u> +5.0	48.2 +7.8	**60.0** +6.3

Table 2: **Frame-selection quality on KFS-Bench at $K = 32$.** All selectors share the candidates and BLIP scores. Scores lie in $[0, 1]$; bold and underlined values denote the best and second-best results.

	Video-MME			LongVideoBench		
Selector	**KFR ↑**	**BSR ↑**	**UKSS ↑**	**KFR ↑**	**BSR ↑**	**UKSS ↑**
Candidate-grid uniform	0.2267	**0.8035**	<u>0.3053</u>	0.1588	0.7222	0.1975
Top-K	<u>0.2940</u>	0.7196	0.2641	<u>0.1942</u>	0.7432	0.2414
Projected inverse-CDF	0.2568	0.7823	0.2936	0.1879	<u>0.7630</u>	0.2448
RIDGE	0.2580	0.7362	0.2640	0.1918	0.7622	<u>0.2459</u>
E^2Q **(ours)**	**0.3015**	<u>0.7976</u>	**0.3129**	**0.2170**	**0.7801**	**0.2611**

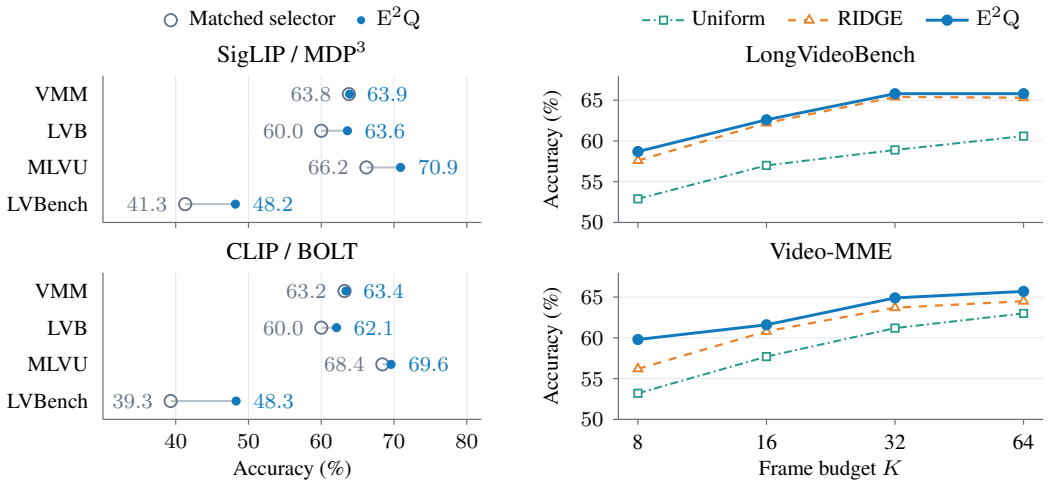
Projected inverse-CDF uses the same $p^\star$ as E^2Q but selects, for each $j = 1, \ldots, K$, the first candidate whose cumulative mass reaches $(j - \frac{1}{2})/K$; that is, like the inverse transform sampling of BOLT (Liu et al., 2025), it selects cumulative-mass points directly rather than forming intervals first.

Component ablations. The no-cap variant applies the same partition objective and interval-maximum rule directly to the normalized relevance distribution q. The time-center variant keeps $p^\star$ and the E^2Q boundaries fixed but selects the candidate nearest each interval's temporal midpoint.

Effects of design choices. E^2Q outperforms candidate-grid uniform by 3.5–12.4 percentage points and Top-K by 1.6–1.8 points on the four benchmarks. With the interval-maximum rule fixed, changing the allocation basis from time to relevance mass improves accuracy by 0.7–2.0 points; with the projected distribution fixed, forming intervals improves over direct inverse-CDF selection by 0.5–1.9 points. Among the ablations, replacing interval maxima with temporal centers lowers accuracy by 1.1–1.9 points, and removing the cap lowers it by 0.2–1.5 points. The cap therefore has a smaller effect than the representative rule on three benchmarks but a larger one on LongVideoBench

Table 3: **Controlled comparison at $K = 32$ with BLIP relevance scores and Qwen2.5-VL-7B.** Accuracy (%). All selectors share the candidate pool, relevance scores, downstream model, and visual preprocessing. Bold and underlined values denote the best and second-best results, respectively.

Selection design				Accuracy (%)			
Selector	**Allocation basis**	**Cap**	**Frame choice**	**VMM**	**LVB**	**MLVU**	**LVBench**
Allocation controls							
Candidate-grid uniform	Candidate index	–	Grid point	61.4	59.1	60.2	38.8
Top-K	Global rank	–	Top scores	63.1	64.0	70.8	49.6
Temporal-bin max	Time	–	Interval max	64.2	63.8	71.7	50.5
Projected inverse-CDF	Relevance mass	✓	CDF point	64.4	63.9	71.9	50.6
Component ablations							
E^{2}Q (no cap)	Relevance mass	×	Interval max	<u>64.6</u>	64.3	<u>72.0</u>	<u>51.0</u>
E^{2}Q (time center)	Relevance mass	✓	Time center	63.4	<u>64.7</u>	70.6	49.7
E^{2}Q **(full)**	Relevance mass	✓	Interval max	**64.9**	**65.8**	**72.5**	**51.2**



(a) Scorer-matched comparison (b) Effect of frame budget

Figure 3: **Scorer and frame-budget analyses with Qwen2.5-VL-7B.** (a) Accuracy at $K = 32$ with matched candidates and scores: hollow markers denote MDP^3 (SigLIP) or BOLT (CLIP); filled markers denote E^{2}Q. Exact accuracies are labeled. (b) Accuracy on the full Video-MME and LongVideoBench question sets; RIDGE and E^{2}Q share the BLIP-scored candidate pool, and Uniform is the uniform-sampling baseline of Table 1. Appendices D.1 and F list the exact values.

(1.5 versus 1.1 points). Because the no-cap variant selects the same frames whenever the cap is inactive, these differences arise only from cap-active questions, which account for 8.5–50.3% of the questions (Appendix E). Together, these results support both relevance-mass allocation and within-interval representative selection. Appendix D.2 reports three further controls: Top-K with temporal non-maximum suppression, random-offset systematic sampling, and an ℓ_2 projection.

5.5 Scorer and Frame-Budget Analysis

Scorer-matched comparison. To separate selector design from scorer choice, Figure 3(a) compares E^{2}Q with MDP^3 under SigLIP scores and with BOLT under CLIP scores; within each pair, both methods share the candidate frames and relevance scores but retain their own score transformations and selection rules. E^{2}Q improves over MDP^3 by 0.1–6.9 percentage points and over BOLT by 0.2–9.0 points on the four benchmarks. The differences are small on Video-MME (0.1 and 0.2 points) but much larger on LVBench (6.9 and 9.0 points). Because the scores are shared, the advantage

cannot be explained by scorer choice alone; Appendix D.1 extends this comparison to AKS and RIDGE under BLIP scores and to WFS-SB under its default BLIP-2 (Li et al., 2023) scores.

Effect of frame budget. We vary the frame budget over $K \in \{8, 16, 32, 64\}$ on the full Video-MME and LongVideoBench question sets. Because K determines both the target relevance mass $1/K$ and the per-candidate cap, each budget induces its own $\mathrm{E}^2\mathrm{Q}$ partition. Figure 3(b) compares $\mathrm{E}^2\mathrm{Q}$ with uniform sampling and with RIDGE using the same BLIP-scored candidates; Appendix F lists the exact values. $\mathrm{E}^2\mathrm{Q}$ outperforms uniform sampling at every budget on both benchmarks, by 5.2–6.9 points on LongVideoBench and 2.7–6.6 points on Video-MME. It also outperforms scorer-matched RIDGE in all eight budget–benchmark comparisons, by 0.4–1.1 points on LongVideoBench and 0.8–3.6 points on Video-MME, with the largest margins at $K = 8$ (1.1 and 3.6 points), where the budget is most constrained. The advantage is therefore not confined to the default 32-frame setting.

5.6 Case Study

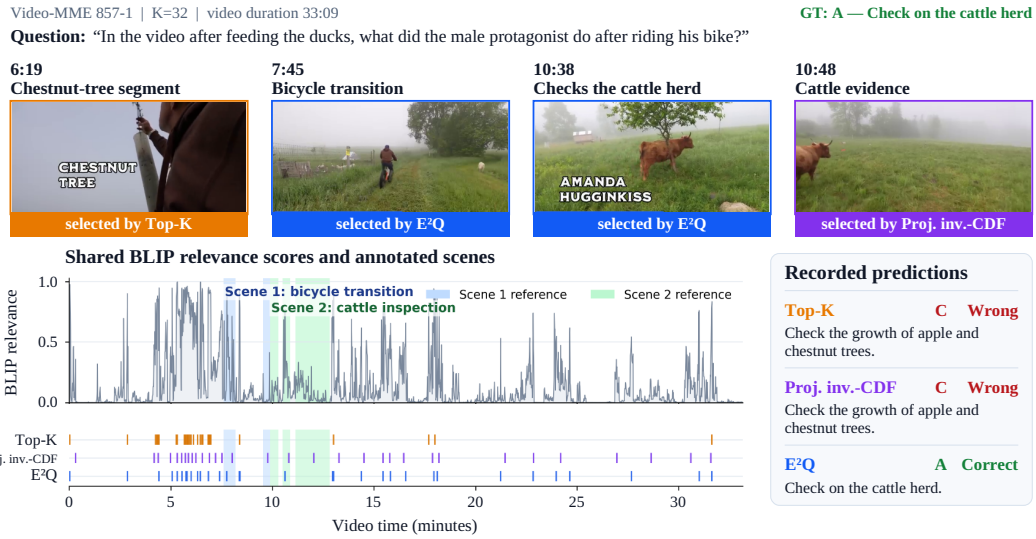


Figure 4: **Video-MME case study at $K = 32$.** Example frames, BLIP scores, selection timelines, and Qwen2.5-VL-7B answers; all selectors share candidates, scores, and inference settings. Blue and green shading mark the annotated bicycle-transition and cattle-inspection intervals, respectively.

The question in Figure 4 asks what the protagonist did after riding his bike. Top-K concentrates on earlier high-scoring frames, including the chestnut-tree segment, and selects no frames from the annotated scenes; the model answers C. Projected inverse-CDF, which selects from exactly the same distribution as $\mathrm{E}^2\mathrm{Q}$ (here $p^\star = q$ because the cap is inactive), reaches both scenes but still yields C. $\mathrm{E}^2\mathrm{Q}$ retains the bicycle transition at 7:45 and the cattle inspection at 10:38, and the model answers correctly (A). Reaching both relevant scenes is thus insufficient for a correct answer; the particular observations retained also matter. Appendix G.3 lists the options and reference intervals, and Appendix I adds six further cases, including one failure. Appendix H discusses limitations.

6 Conclusion

We presented $\mathrm{E}^2\mathrm{Q}$, a training-free frame selector that separates budget allocation from representative selection: it partitions the candidate sequence into consecutive intervals of approximately equal relevance mass and retains the highest-scoring frame in each. Our analysis bounds the error of approximating relevance-weighted frame-level averages with these representatives. Across four long-video QA benchmarks and three downstream models, $\mathrm{E}^2\mathrm{Q}$ achieves the highest accuracy in ten of the twelve settings against six recent training-free selectors, and controlled comparisons support both relevance-mass allocation and within-interval representative selection. $\mathrm{E}^2\mathrm{Q}$ also achieves the highest UKSS among the compared selectors on both KFS-Bench subsets. These results highlight the benefit of distributing the frame budget according to relevance before choosing individual frames.

Ethics Statement

This work complies with the ICLR Code of Ethics. We evaluate our training-free frame selector on publicly available video benchmarks under their original licenses and terms of use; we collect no new data and involve no human subjects. Our method inherits potential biases and privacy risks from the source videos and pretrained models, and efficient long-video analysis may enable misuse such as unauthorized surveillance. Moreover, because frame selection retains only a small fraction of the input, selected frames should not be treated as a complete account of a video when omitted context could cause harm.

Reproducibility Statement

Section 3 defines the selector, Appendix A provides the proofs, and Appendix B gives the algorithm, implementation details, and a worked example. Appendix C specifies the benchmarks, candidate sampling, relevance scoring, inference settings, and controlled selection rules, and Appendices D–G report the numerical results underlying the figures and the supplementary analyses. To facilitate reproduction and future comparisons, the code is included in the submitted supplementary material.

AI Use Statement

We used generative AI tools only to polish the language of the manuscript and to assist in the visualization of some figures. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the entire content of this paper, including its text, claims, and figures.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of ECCV*, pp. 19–35, 2024.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *Proceedings of CVPR*, pp. 24052–24061, 2026.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. LongVILA: Scaling long-context visual language models for long videos. In *Proceedings of ICLR*, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Jack Edmonds. Submodular functions, matroids, and certain polyhedra. In Richard Guy, Haim Hanani, Norbert Sauer, and J. Schönheim (eds.), *Combinatorial Structures and Their Applications*, pp. 69–87. Gordon and Breach, New York, 1970.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of CVPR*, pp. 24108–24118, 2025.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44 (6):2325–2383, 1998.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-LLM based video frame selection for efficient video understanding. In *Proceedings of CVPR*, pp. 13702–13712, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of ICML*, pp. 12888–12900, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of ICML*, pp. 19730–19742, 2023.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *Proceedings of ECCV*, pp. 323–340, 2024.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. KFS-Bench: Comprehensive evaluation of key frame sampling in long video understanding. In *Proceedings of WACV*, pp. 5643–5652, 2026.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *Proceedings of CVPR*, pp. 3318–3327, 2025.

William G. Madow and Lillian H. Madow. On the theory of systematic sampling, I. *The Annals of Mathematical Statistics*, 15(1):1–24, 1944.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of CVPR*, pp. 16944–16953, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pp. 8748–8763, 2021.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of ICML*, pp. 54582–54599, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of CVPR*, pp. 18221–18232, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A training-free approach for list-wise frame selection in video-LLMs. In *Proceedings of ICCV*, pp. 24090–24101, 2025.

Yucheng Suo, Fan Ma, Linchao Zhu, Tianyi Wang, Fengyun Rao, and Yi Yang. From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. In *Proceedings of ICCV*, pp. 23243–23255, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of CVPR*, pp. 29118–29128, 2025.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. LVBench: An extreme long video understanding benchmark. In *Proceedings of ICCV*, pp. 22958–22967, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Proceedings of ECCV*, pp. 58–76, 2024.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of CVPR*, pp. 3272–3283, 2025b.

Manfred K. Warmuth and Dima Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9:2287–2320, 2008.

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. LongVLM: Efficient long video understanding via large language models. In *Proceedings of ECCV*, pp. 453–470, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Proceedings of NeurIPS*, volume 37, pp. 28828–28857, 2024.

Shanqing Xu, Meng Luo, Mengchen Qian, Yuhui Gao, Siyue Peng, Xiaohan Zhong, Xiaojin Zhang, Zhongyu Wei, Wei Chen, and Xiang Bai. RIDGE: Region-informed derivative-guided evidence selection for long video understanding. *arXiv preprint arXiv:2608.29958*, 2026.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of CVPR*, pp. 8579–8591, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Proceedings of NeurIPS*, volume 36, 2023.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. Frame-Voyager: Learning to query frames for video large language models. In *Proceedings of ICLR*, 2025.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. VideoExplorer: Think with videos for agentic long-video understanding. *arXiv preprint arXiv:2506.10821*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of ICCV*, pp. 11975–11986, 2023.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of NAACL*, pp. 881–916, 2025a.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Transactions on Machine Learning Research*, 2025b.

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 ICCV*, pp. 22056–22065, 2025c.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of CVPR*, pp. 13691–13701, 2025.

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.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *Proceedings of ICLR*, 2026.

A Proofs and Additional Analysis

This appendix establishes the projection and boundary properties used by $\mathrm{E}^2\mathrm{Q}$, proves the representation bound in Theorem 1, and gives a geometric interpretation of the budget-dependent cap. Unless stated otherwise, the arguments assume exact arithmetic; implementation tolerances and numerical conventions are given in Appendix B, together with the boundary dynamic program.

A.1 Capped KL Projection

Proposition 1 (Water-filling projection)**.** *For an integer $1 \le K \le n$ and a distribution q with $q_i > 0$ and $\sum_i q_i = 1$, the KL projection in Equation 2 is unique and is given by*

$$p_i^\star = \min\left(cq_i, \frac{1}{K}\right),$$

where $c > 0$ is chosen so that $\sum_i p_i^\star = 1$. For $n > K$, this normalizing factor is unique. For $n = K$, the solution is the uniform distribution, $p_i^\star = 1/K$ for every i.

This closed form also underlies the capping algorithm of Warmuth & Kuzmin (2008).

Proof. The set $\mathcal{P}_K$ is nonempty (the uniform distribution is feasible), compact, and convex. Since $D_{\mathrm{KL}}(p\|q)$ is strictly convex in p, its minimizer is unique. The minimizer is strictly positive. If $p_i = 0$, transferring an infinitesimal amount of mass from any positive coordinate to i is feasible and has directional derivative $-\infty$ for $D_{\mathrm{KL}}(p\|q)$, so it strictly decreases the objective.

Let λ be the multiplier for normalization and $\alpha_i \ge 0$ the multiplier for the upper bound. The Karush–Kuhn–Tucker stationarity condition (Boyd & Vandenberghe, 2004) gives

$$\log(p_i/q_i) + 1 + \lambda + \alpha_i = 0.$$

At an uncapped coordinate, complementary slackness yields $p_i = cq_i$, where $c = \exp(-1 - \lambda)$. At a capped coordinate, $p_i = 1/K$ and, since $\alpha_i \ge 0$, $cq_i \ge 1/K$. Hence every coordinate satisfies

$$p_i = \min\left(cq_i, \frac{1}{K}\right).$$

Define

$$g(c) = \sum_i \min\left(cq_i, \frac{1}{K}\right).$$

The function g is continuous and strictly increasing until all coordinates are capped. Moreover, $g(0) = 0$, while $\lim_{c\to\infty} g(c) = n/K > 1$ when $n > K$. Therefore there is a unique c such that $g(c) = 1$. When $n = K$, feasibility itself forces $p_i = 1/K$ for every i. □

A useful invariance follows immediately: multiplying all raw relevance scores by the same positive constant leaves q, $p^\star$, and the selected frames unchanged in exact arithmetic.

A.2 Nearest Cumulative Boundaries

For a positive capped distribution p, let $F(b) = \sum_{i=1}^{b} p_i$. For internal boundary j, define the feasible set $\mathcal{B}_j = \{j, \dots, n - K + j\}$ and choose

$$\widehat{b}_j = \min\left(\arg\min_{b\in\mathcal{B}_j} \left|F(b) - \frac{j}{K}\right|\right). \tag{7}$$

The outer minimum implements the earlier-boundary tie rule.

Proposition 2 (Distinct nearest boundaries)**.** *If $p_i > 0$, $\sum_i p_i = 1$, and $p_i \le 1/K$, the boundaries in Equation 7 are strictly increasing. They jointly minimize Equation 4 and can be found by a single forward scan in $O(n + K)$ time in exact arithmetic, without dynamic programming.*

Proof. First, the selected boundaries cannot decrease as the target increases. Suppose $\widehat{b}_{j+1} < \widehat{b}_j$. Under this assumption, both positions lie in $\mathcal{B}_j \cap \mathcal{B}_{j+1}$: $\widehat{b}_{j+1} \geq j+1$ and $\widehat{b}_j \leq n-K+j$. Write $x = j/K < y = (j+1)/K$ and $u = F(\widehat{b}_{j+1}) < v = F(\widehat{b}_j)$. Optimality for the two targets gives $|x-v| \leq |x-u|$ and $|y-u| \leq |y-v|$, hence $x \geq (u+v)/2 \geq y$, a contradiction.

It remains to rule out equal adjacent boundaries. Suppose both targets choose b. Then $b-1$ is feasible for target j/K, and $b+1$ is feasible for target $(j+1)/K$. Let $F_- = F(b-1)$, $F_0 = F(b)$, and $F_+ = F(b+1)$. Because an exact tie for the first target would be resolved in favor of the earlier boundary $b-1$, choosing b requires $j/K > (F_- + F_0)/2$. Choosing b over the later boundary $b+1$ requires $(j+1)/K \leq (F_0 + F_+)/2$. Subtracting yields

$$\frac{1}{K} < \frac{F_+ - F_-}{2} = \frac{p_b + p_{b+1}}{2} \leq \frac{1}{K},$$

a contradiction.

Thus the coordinate-wise earliest minimizers are strictly ordered and form a feasible partition. Because each coordinate attains its individual minimum, the resulting sequence also minimizes the sum in Equation 4. Since both the targets and cumulative masses are ordered, a forward pointer visits each candidate boundary at most once, giving $O(n+K)$ time. □

Relation to the implementation. The implementation uses the dynamic program in Appendix B to solve Equation 4 directly. Proposition 2 shows that, under the cap and in exact arithmetic, the coordinate-wise earliest minimizers already form a feasible global optimum. With the same earlier-predecessor and earlier-final-state tie rules, the dynamic program therefore recovers these boundaries; in particular, $\widehat{b}_j = b_j^\star$ for every internal boundary j, as used in Corollary 1 below.

Corollary 1 (Interval maxima and saturated candidates)**.** *For* $\mathrm{E}^2\mathrm{Q}$*'s projected distribution and earliest-nearest boundaries, each interval contains at most one saturated candidate, i.e., a candidate with* $p_i^\star = 1/K$*. In exact arithmetic, the earliest maximizers of* r_i *and* $p_i^\star$ *in each interval coincide, and every saturated candidate is retained as the representative of its interval.*

Proof. Let $h = 1/K$, let $F(b) = \sum_{i=1}^{b} p_i^\star$, and define $e_j = F(b_j^\star) - jh$, with $e_0 = e_K = 0$. Because each of the first j weights is at most h, $F(j) \leq jh$. Similarly, after position $n-K+j$ at most $K-j$ weights remain, so $jh \leq F(n-K+j)$. Hence

$$F(j) \leq jh \leq F(n-K+j),$$

and the feasible range brackets each internal target. Consecutive cumulative values differ by at most h, so a nearest feasible boundary lies within $h/2$ of its target. A boundary exactly $h/2$ above the target cannot be selected, because its predecessor is at least as close and is earlier. Therefore

$$-\frac{h}{2} \leq e_j < \frac{h}{2}.$$

Let $m_j = \sum_{i \in I_j} p_i^\star$ be the mass of interval j. Then

$$m_j = F(b_j^\star) - F(b_{j-1}^\star) = h + e_j - e_{j-1},$$

so $0 < m_j < 2h$. Two saturated candidates would already contribute mass $2h$, so each interval contains at most one saturated candidate, which proves the first claim.

If an interval contains no saturated candidate, then $p_i^\star = cq_i$ throughout that interval, preserving the ordering and exact ties of the raw relevance scores r_i. If it contains one saturated candidate, that candidate satisfies $cq_i \geq h$, whereas every other candidate in the interval satisfies $cq_i < h$; it is therefore the unique maximum under both r_i and $p_i^\star$. The representative rules coincide, and every saturated candidate is retained. For $K = 1$, the same mass argument applies with no internal boundaries, and the single interval contains all candidates. □

A.3 Proof of Theorem 1

To make the representative choice explicit, let $\widetilde{s}_j \in I_j$ be arbitrary. Setting $\widetilde{s}_j = s_j^\star$ recovers the representatives selected by $\mathrm{E}^2\mathrm{Q}$. Add and subtract the mass-weighted representative average:

$$\left|\sum_i p_i^\star f_i - \frac{1}{K}\sum_j f_{\widetilde{s}_j}\right| \le \left|\sum_j \sum_{i\in I_j} p_i^\star (f_i - f_{\widetilde{s}_j})\right| + \left|\sum_j \left(m_j - \frac{1}{K}\right) f_{\widetilde{s}_j}\right|$$
$$\le \sum_j m_j \operatorname{osc}_{I_j}(f) + \|f\|_\infty \sum_j \left|m_j - \frac{1}{K}\right|.$$

Let $e_j = F(b_j^\star) - j/K$, with $e_0 = e_K = 0$. Since $m_j - 1/K = e_j - e_{j-1}$,

$$\sum_j \left|m_j - \frac{1}{K}\right| \le \sum_j \big(|e_j| + |e_{j-1}|\big) = 2\Delta_{\mathrm{cut}}.$$

Combining the two displays proves Equation 6.

For comparison with the original relevance distribution q, the triangle inequality gives

$$\left|\sum_i q_i f_i - \frac{1}{K}\sum_j f_{\widetilde{s}_j}\right| \le \|f\|_\infty \|q - p^\star\|_1 + \sum_j m_j \operatorname{osc}_{I_j}(f) + 2\|f\|_\infty \Delta_{\mathrm{cut}}.$$

By Pinsker's inequality (Cover & Thomas, 2006), the additional projection term is at most $\|f\|_\infty \sqrt{2 D_{\mathrm{KL}}(p^\star \| q)}$, so the extra error from comparing against q is controlled by the KL projection cost. For normalized timestamps $\tau_i \in [0,1]$, if $|f_i - f_\ell| \le L|\tau_i - \tau_\ell|$, then

$$\operatorname{osc}_{I_j}(f) \le L\left(\max_{i\in I_j} \tau_i - \min_{i\in I_j} \tau_i\right).$$

Thus temporally compact intervals make the first term small whenever f varies smoothly in time. The theorem applies to any representative within each interval; it bounds the approximation error in terms of the variation of f alone, independently of the representative-selection rule.

A.4 Geometric Interpretation of the Cap

The following characterization, which follows from classical polyhedral results (Edmonds, 1970) and appears in Warmuth & Kuzmin (2008), explains why the cap $1/K$ is natural for a budget of K equal-weight output slots. For a K-element subset S, define $v^S = K^{-1}\mathbf{1}_S$ and $\mathcal{U}_K = \{v^S : |S| = K\}$.

Proposition 3 (Feasible equal-weight marginals)**.** *The capped simplex satisfies*

$$\mathcal{P}_K = \operatorname{conv}(\mathcal{U}_K). \tag{8}$$

Proof. Scaling $\mathcal{P}_K$ by K gives the polytope

$$\left\{z \in [0,1]^n : \sum_i z_i = K\right\}.$$

A nonbinary vector in this set has at least two fractional coordinates, because its coordinate sum is the integer K. Moving a sufficiently small amount between two such coordinates in either direction produces two distinct feasible vectors whose average is the original one, so it cannot be a vertex. Every vertex is thus an indicator of a K-element subset. Since a bounded polytope is the convex hull of its vertices, scaling back by K yields $\mathcal{P}_K = \operatorname{conv}(\mathcal{U}_K)$. □

For a randomized K-element subset, the expected equal-weight vector has entries $\Pr(i \in S)/K$. Thus the identity characterizes the feasible average allocations without specifying a unique distribution over subsets. Random-offset systematic sampling (Appendix C.2) preserves $\Pr(i \in S) = K p_i^\star$; $\mathrm{E}^2\mathrm{Q}$ instead uses deterministic interval representatives and therefore involves no sampling randomness.

B Algorithm and Worked Example

Algorithm 1 summarizes $\mathrm{E}^2\mathrm{Q}$. The paragraphs below detail the projection, the boundary dynamic program, and the numerical conventions, followed by a worked example.

Algorithm 1 Equal-rate Evidence Quantization ($\mathrm{E}^2\mathrm{Q}$)

Require: n chronological candidates with positive scores $(r_i)_{i=1}^n$; integer budget $1 \le K \le n$
Ensure: K selected frames in chronological order
1: $q_i \leftarrow r_i / \sum_{\ell=1}^n r_\ell$ for $i = 1, \dots, n$
2: Compute $p^\star$ by water-filling: cap weights above $1/K$ and redistribute the remaining mass proportionally until no uncapped weight exceeds the cap
3: $F(b) \leftarrow \sum_{i=1}^b p_i^\star$ for $b = 0, \dots, n$
4: Minimize Equation 4 with the prefix-minimum recurrence below and backtrack the internal boundaries $b_1^\star, \dots, b_{K-1}^\star$ from the stored predecessors
5: **for** $j = 1, \dots, K$ **do**
6: $\quad s_j^\star \leftarrow \arg\max_{i \in I_j} r_i$, preferring the earliest exact maximizer
7: **end for**
8: **return** the frames $s_1^\star < \cdots < s_K^\star$ in chronological order

Projection implementation. For $n > K$, let A contain the candidates already fixed at the cap, initially $A = \varnothing$. At each step, set

$$c_A = \frac{1 - |A|/K}{\sum_{i \notin A} q_i}$$

and fix every remaining candidate with $c_A q_i > 1/K$ at the cap. Recompute c_A after removing those candidates from the free set. When no free weight exceeds the cap, assign $p_i^\star = c_A q_i$ to the remaining candidates. For $n = K$, the projection is trivial, and we set $p_i^\star = 1/K$ directly.

Boundary dynamic program. We solve Equation 4 directly with dynamic programming. Proposition 2 shows that, in exact arithmetic, the resulting boundaries coincide with those obtained by the faster nearest-boundary scan. Using the feasible sets $\mathcal{B}_j$ from Appendix A, let $D_j(b)$ be the minimum cost of the first j internal boundaries when boundary j is placed at b, and initialize

$$D_1(b) = \left| F(b) - \frac{1}{K} \right|, \qquad b \in \mathcal{B}_1.$$

For $j = 2, \dots, K-1$,

$$D_j(b) = \left| F(b) - \frac{j}{K} \right| + \min_{\substack{a < b \\ a \in \mathcal{B}_{j-1}}} D_{j-1}(a), \qquad b \in \mathcal{B}_j.$$

A running prefix minimum computes each layer in $O(n)$ time. The implementation keeps the earlier predecessor when costs are exactly equal, selects the earliest minimum-cost final boundary, and then backtracks. For $K = 1$, the single interval contains all candidates and the earliest global score maximum is selected. For $n = K$, each interval is a singleton and every candidate is selected. The dynamic program takes $O(Kn)$ time and $O(Kn)$ memory including backpointers. The current active-set projection takes at most $O(Kn)$ time, and representative selection takes $O(n)$.

Input assumptions and numerical conventions. The candidate and score arrays have equal length and are aligned in chronological order. Scores are finite and strictly positive, and K is an integer satisfying $1 \le K \le n$. The proofs assume exact arithmetic and the tie rules above; the implementation uses float64 arithmetic and an absolute tolerance of 10^{-12} for the capacity check.

Worked example with $K = 4$. Figure 2 uses ten candidates with

$$r = (1, 1, 10, 1, 1, 1, 4, 1, 1, 1), \qquad K = 4.$$

Normalization divides by 22. Since $q_3 = 10/22 > 1/4$, the cap is active. The water-filling projection uses $c = 22/16$ on the uncapped coordinates and gives

$$p^\star = (1, 1, 4, 1, 1, 1, 4, 1, 1, 1)/16.$$

The three internal cumulative-mass targets are $4/16$, $8/16$, and $12/16$. For the first target,

$$F(2) = \frac{2}{16}, \qquad F(3) = \frac{6}{16},$$

so boundaries 2 and 3 are both $2/16$ from the target; the earlier-tie rule gives $b_1^\star = 2$. The second target is attained exactly at $F(5) = 8/16$, hence $b_2^\star = 5$. For the third target, $F(6) = 9/16$ and $F(7) = 13/16$, whose distances from $12/16$ are $3/16$ and $1/16$, respectively, so $b_3^\star = 7$.

Thus the complete boundary vector is

$$(b_0^\star, \ldots, b_4^\star) = (0, 2, 5, 7, 10),$$

which gives intervals $\{1, 2\}$, $\{3, 4, 5\}$, $\{6, 7\}$, and $\{8, 9, 10\}$. Taking the earliest raw-score maximizer in each interval yields selected indices $(1, 3, 7, 8)$. The interval masses are $2/16$, $6/16$, $5/16$, and $3/16$; they need not equal $1/4$ exactly because cuts are restricted to discrete candidate boundaries.

C Evaluation Protocol

C.1 Data, Scores, and Model Inputs

Table 4 summarizes the evaluation settings of the four benchmarks; all QA accuracy comparisons use their full question sets, while the annotation-based analyses in Appendix G use the annotated subsets described there. Controlled experiments use Qwen2.5-VL-7B with a 32-frame budget unless stated otherwise. Within each controlled comparison, all selectors receive the same question IDs, chronological candidate frames, timestamps, and cached relevance scores; visual preprocessing, prompts, model weights, and decoding settings are also held fixed across selectors.

Table 4: **Evaluation settings of the four benchmarks.** QA accuracy comparisons use the full question set of each benchmark; Video-MME is evaluated without subtitles or audio.

Benchmark	Split	Questions	Metric	Evaluator	Max new tokens
Video-MME	Full	2,700	Accuracy	`lmms-eval` fork	16
LongVideoBench	Validation	1,337	Accuracy	`lmms-eval` fork	32
MLVU	Development	2,174	M-Avg	`lmms-eval` fork	16
LVBench	Full	1,549	Accuracy	Selected-image	128

Candidate frames. For a video with T native frames and measured frame rate ν, we sample candidate frames at the native indices

$$\mathcal{C} = \{\ell\delta : \ell \in \mathbb{Z}_{\geq 0},\ \ell\delta < T\}, \qquad \delta = \max\{\lfloor \nu \rfloor, 1\}.$$

This gives approximately one candidate per second, starting at frame zero without forcing the final frame. There is no fixed candidate count or maximum-candidate truncation. Timestamps are native indices divided by ν, and the budget K is applied only after all candidates are scored.

Relevance scores. We use the image–text matching (ITM) head of BLIP-1 (Li et al., 2022) with the pretrained checkpoint `Salesforce/blip-itm-base-coco`, without fine-tuning. The text input is the question followed by all answer options in their original order, without the ground-truth answer. When the stored query already contains the options, it is used directly. The processor applies padding and truncation with its default length settings. With the ITM head enabled, the relevance of frame x_i to question u is computed as

$$r_i = \left[\mathrm{softmax}(\mathrm{ITM}(x_i, u))\right]_1,$$

the positive matching-class probability. Scores are computed on GPU in FP16 and cached with the query and native frame indices. $\mathrm{E}^2\mathrm{Q}$ uses the cached positive scores directly in Equation 1, without ranking, temperature calibration, or time-gap reweighting. Neither downstream predictions nor ground-truth labels enter scoring. Multiplying all scores by a common positive constant leaves the selection unchanged, whereas a general monotone transformation need not.

Downstream inference. Controlled experiments use Qwen2.5-VL-7B-Instruct with batch size one. Selected frames are decoded with Decord and supplied as an image list in chronological order, without audio or external subtitles. The Qwen adapter uses `use_keyframe=True`, `max_num_frames=32`, `max_pixels=12845056`, and `interleave_visuals=False` at the default budget. A fixed, dataset-specific zero-shot prompt asks for only the option letter. Video-MME, MLVU, and LongVideoBench are evaluated with a modified fork of `lmms-eval` (Zhang et al., 2025a) and its Qwen2.5-VL keyframe adapter; LVBench uses a separate selected-image evaluator. All evaluations use greedy decoding (`do_sample=False`, `temperature=0`, `top_p=1`, `num_beams=1`), with the maximum number of new tokens listed in Table 4. In the native comparison, InternVL3-8B and LLaVA-OneVision-7B are run with their official default inference configurations.

C.2 Controlled Selection Rules

Candidate-grid uniform and Top-K. Candidate-grid uniform selects the candidates at K equally spaced positions of the common candidate list, rounding each position to the nearest index. Top-K sorts candidates by the common relevance score, breaks exact ties toward the earlier candidate, keeps the K highest-scoring candidates, and returns them in chronological order.

Temporal-bin max. With normalized timestamps $\tau_i \in [0, 1]$, temporal-bin max chooses strictly increasing boundaries minimizing $\sum_{j=1}^{K-1} |\tau_{b_j} - j/K|$ and retains the highest-scoring candidate in each interval. Here b_j is the index of the final candidate in interval j, so τ_{b_j} is the timestamp immediately before the cut. Equal-cost boundary choices follow the dynamic program's tie rules in Appendix B. It shares the nonempty-interval and maximum-representative rules of $\mathrm{E}^2\mathrm{Q}$ and differs only in using elapsed time, rather than cumulative relevance, as the allocation axis.

Projected inverse-CDF. This control computes the same $p^\star$ as $\mathrm{E}^2\mathrm{Q}$ and selects the smallest index i satisfying $F(i) \geq (j - 1/2)/K$ for $j = 1, \ldots, K$. An exact CDF crossing is assigned to the smallest such index. Under $p_i^\star \leq 1/K$, the selected indices are distinct in exact arithmetic.

No-cap and representative variants. $\mathrm{E}^2\mathrm{Q}$ (no cap) partitions the original normalized distribution q using the same boundary objective (Equation 4) and interval maxima as $\mathrm{E}^2\mathrm{Q}$. $\mathrm{E}^2\mathrm{Q}$ (time center) retains $p^\star$ and the $\mathrm{E}^2\mathrm{Q}$ boundaries but selects the candidate nearest each interval's temporal midpoint, with earlier candidates breaking exact ties; each variant thus changes a single component of $\mathrm{E}^2\mathrm{Q}$.

Top-K with temporal NMS. This control applies temporal non-maximum suppression (NMS) to Top-K: it traverses the common score ranking and suppresses candidates within normalized temporal distance $1/(2K)$ of a selected frame. If fewer than K survive, the remaining positions are filled from unused candidates in score order, so exactly K frames are always returned.

Random-offset systematic sampling. Following classical systematic sampling (Madow & Madow, 1944), we draw $\xi \sim \mathrm{Unif}[0, 1)$ and invert the projected CDF at $(j - 1 + \xi)/K$, $j = 1, \ldots, K$. Since $p_i^\star \leq 1/K$, the sampled indices are distinct almost surely and candidate i has inclusion probability $Kp_i^\star$. We report the mean accuracy over three runs with PCG64 seeds 17, 29, and 43.

ℓ_2 projection. The alternative projection solves

$$\min_{p \in \mathcal{P}_K} \|p - q\|_2^2.$$

Its solution has the form

$$p_i = \min\left\{\max(q_i - \lambda, 0), \frac{1}{K}\right\},$$

where λ is chosen to enforce normalization. The partitioning and representative-selection steps are otherwise identical to those of $\mathrm{E}^2\mathrm{Q}$, so only the projection differs.

D Additional Comparisons and Controls

D.1 Scorer-Matched Comparisons

Table 5 lists the exact accuracies plotted in Figure 3(a) and adds two scorer blocks: a BLIP block with AKS and RIDGE, and a BLIP-2 block with WFS-SB, whose default scorer is the ITM head of BLIP-2 (Li et al., 2023). Candidate frames and relevance scores are shared within each scorer block, while method-specific score transformations and selection rules are retained; all rows use the full question sets and the same model, frame budget, and visual preprocessing. Within the BLIP block, E^2Q exceeds RIDGE by 0.3–2.9 points and AKS by 1.8–7.8 points on the four benchmarks. Within the BLIP-2 block, it exceeds WFS-SB on every benchmark, by 0.4–2.1 points, so its lead over WFS-SB in Table 1 persists when both selectors use WFS-SB's default BLIP-2 scores.

Table 5: **Expanded scorer-matched comparison at $K = 32$ with Qwen2.5-VL-7B.** Accuracy (%); Avg. is the unweighted mean. Selectors within each scorer block share candidate frames and relevance scores; BLIP-2 is the default scorer of WFS-SB. Bold marks the best result in each block, and small blue numbers give the gain of E^2Q over the block's strongest baseline.

Selector	VMM	LVB	MLVU	LVBench	Avg.
BLIP scores					
AKS	63.1	63.2	67.2	43.4	59.2
RIDGE	63.7	65.4	69.6	50.9	62.4
E^2Q	**64.9** +1.2	**65.8** +0.4	**72.5** +2.9	**51.2** +0.3	**63.6** +1.2
CLIP scores					
BOLT	63.2	60.0	68.4	39.3	57.7
E^2Q	**63.4** +0.2	**62.1** +2.1	**69.6** +1.2	**48.3** +9.0	**60.9** +3.2
SigLIP scores					
MDP^3	63.8	60.0	66.2	41.3	57.8
E^2Q	**63.9** +0.1	**63.6** +3.6	**70.9** +4.7	**48.2** +6.9	**61.7** +3.9
BLIP-2 scores					
WFS-SB	64.2	64.4	69.9	48.1	61.7
E^2Q	**65.2** +1.0	**64.8** +0.4	**71.3** +1.4	**50.2** +2.1	**62.9** +1.2

D.2 Additional Selection Rules

Table 6 compares three additional controls with the projected inverse-CDF reference and E^2Q under the same controlled inputs as Table 3. The additional controls test temporal suppression, random-offset systematic sampling, and an alternative capped projection. E^2Q attains the highest accuracy on all four benchmarks, tying the ℓ_2-projection variant on LVBench (51.2). Temporal NMS improves Top-K on three benchmarks (by 0.8–1.0 points; Table 3) but remains 0.8–1.9 points below E^2Q. Random-offset systematic sampling is 0.2 points below its deterministic counterpart, projected inverse-CDF, on every benchmark, and replacing the KL projection with the ℓ_2 projection lowers accuracy by up to 1.7 points (on LongVideoBench) without improving it on any benchmark.

Table 6: **Supplementary controlled comparison at $K = 32$.** Accuracy (%). Inputs match Table 3; random-offset systematic sampling reports the mean accuracy over three seeds. Bold values denote the best result in each column, including the tie on LVBench.

Selector	VMM	LVB	MLVU	LVBench
Additional controls				
Top-K + temporal NMS	64.1	63.9	71.6	50.4
Random-offset systematic	64.2	63.7	71.7	50.4
ℓ_2 projection	64.8	64.1	72.3	**51.2**
Reference selectors				
Projected inverse-CDF	64.4	63.9	71.9	50.6
E^2Q	**64.9**	**65.8**	**72.5**	**51.2**

D.3 Paired Significance Tests

Table 7 tests whether the gains of $\mathrm{E}^2\mathrm{Q}$ over WFS-SB with Qwen2.5-VL-7B in Table 1 are statistically significant; in this setting, WFS-SB is the strongest competing selector on Video-MME and MLVU. Both selectors answer the same questions, so the analysis pairs their per-question correctness: p is from a two-sided paired test on these outcomes, and the 95% confidence interval for the accuracy difference is a paired bootstrap interval over questions (Efron & Tibshirani, 1993). The gains are significant at the 5% level on MLVU ($p = 0.004$), LVBench ($p = 0.007$), and LongVideoBench ($p = 0.037$), but not on Video-MME ($p = 0.359$), whose 95% confidence interval includes zero.

Table 7: **Paired significance against WFS-SB with Qwen2.5-VL-7B at $K = 32$.** Accuracy (%) is from Table 1. Δ is $\mathrm{E}^2\mathrm{Q}$ minus WFS-SB in percentage points (pp); the 95% CI is a paired bootstrap interval over questions, and p is from a two-sided paired test on per-question correctness.

Benchmark	WFS-SB	$\mathrm{E}^2\mathrm{Q}$	Δ (pp)	95% CI (pp)	p
Video-MME	64.2	64.9	$+0.7$	$[-0.7, +2.1]$	0.359
LongVideoBench	64.4	65.8	$+1.4$	$[+0.2, +2.7]$	0.037
MLVU	69.9	72.5	$+2.6$	$[+0.8, +4.3]$	0.004
LVBench	48.1	51.2	$+3.1$	$[+0.9, +5.2]$	0.007

E Projection Diagnostics

A question is *cap-active* when $\max_i q_i > 1/K$. Table 8 reports the activation count and rate, projection magnitude, and selected-set change between $\mathrm{E}^2\mathrm{Q}$ and the no-cap variant. Let S and S_0 denote the selections with and without the cap. Over cap-active questions, we report the mean number of replaced frames, $K - |S \cap S_0|$, and the mean Jaccard similarity, $|S \cap S_0|/|S \cup S_0|$. On cap-inactive questions, $p^\star = q$ and both variants select the same frames under identical tie rules, so the cap-active questions are exactly those on which the projection can change the selection.

Table 8: **Projection diagnostics at $K = 32$.** The activation count and rate and the mean $D_{\mathrm{KL}}(p^\star \| q)$ are computed over all questions; the selection change and Jaccard similarity are computed over cap-active questions only. Parentheses give the percentage of all questions.

	All questions		Cap-active questions	
Benchmark	Active, count (%)	Mean D_{KL}	Mean replaced	Mean Jaccard
Video-MME	745 (27.59)	0.0907	6.38	0.683
LongVideoBench	345 (25.80)	0.2127	8.05	0.623
MLVU	1,093 (50.28)	0.1923	7.82	0.631
LVBench	131 (8.46)	0.0064	6.34	0.684

The cap is active on 8.46% (LVBench) to 50.28% (MLVU) of the questions; when active, it replaces 6.34–8.05 of the 32 selected frames on average (mean Jaccard similarity 0.623–0.684). Across benchmarks, the accuracy drop from removing the cap (Table 3) follows the same ordering as the mean projection magnitude, increasing from LVBench (0.2 points; mean D_{KL} 0.0064) to Video-MME (0.3; 0.0907), MLVU (0.5; 0.1923), and LongVideoBench (1.5; 0.2127).

F Additional Frame-Budget Results

Table 9 lists the values plotted in Figure 3(b). Every selector and budget is evaluated on the full Video-MME and LongVideoBench question sets; RIDGE and $\mathrm{E}^2\mathrm{Q}$ share the same BLIP-scored candidate pool, and Uniform is the uniform-sampling baseline of Table 1.

Table 9: **Frame-budget comparison with Qwen2.5-VL-7B.** Accuracy (%). Bold and underlined values denote the best and second-best results for each benchmark and budget; small blue numbers give the gain of E^2Q over scorer-matched RIDGE at the same budget.

	Frame budget K			
Selector	**8**	**16**	**32**	**64**
LongVideoBench				
Uniform	52.9	57.0	58.9	60.6
RIDGE	<u>57.6</u>	<u>62.2</u>	<u>65.4</u>	<u>65.3</u>
E^2Q	**58.7** +1.1	**62.6** +0.4	**65.8** +0.4	**65.8** +0.5
Video-MME				
Uniform	53.2	57.7	61.2	63.0
RIDGE	<u>56.2</u>	<u>60.8</u>	<u>63.7</u>	<u>64.5</u>
E^2Q	**59.8** +3.6	**61.6** +0.8	**64.9** +1.2	**65.7** +1.2

G Direct Selection Analysis

G.1 KFS-Bench

KFS-Bench (Li et al., 2026) provides temporal annotations for scenes needed to answer questions from Video-MME and LongVideoBench. This section gives the full metrics and scene-count breakdown for Section 5.3. We report Key Frame Rate (KFR), Balanced Scene Recall (BSR), and Balanced Distribution Similarity (BDS), which measure the fraction of selected frames in relevant scenes, coverage under scene-specific sampling thresholds, and agreement with a balanced scene-level allocation, respectively. Unified Keyframe Sampling Score (UKSS) combines the three criteria.

We use the official per-question metric definitions. KFR, BSR, and BDS are reported as arithmetic means over questions. For question a, let KFR_a, BSR_a, and BDS_a denote the three per-question scores; with N evaluated questions, UKSS is aggregated as

$$\mathrm{UKSS} = \exp\left[\frac{1}{N}\sum_{a=1}^{N}\log\max\left\{0.01, (\mathrm{KFR}_a\mathrm{BSR}_a\mathrm{BDS}_a)^{1/3}\right\}\right], \tag{9}$$

following the benchmark implementation. All metrics lie in $[0, 1]$. Selectors share the candidate-sampling and BLIP-scoring pipeline and are evaluated on identical annotated question IDs. Scene annotations are used only for evaluation. Table 10 reports all four metrics. E^2Q is best or second-best on every metric; it trails only candidate-grid uniform on Video-MME BSR (0.7976 versus 0.8035) and BDS (0.8489 versus 0.8514), in both cases by small margins.

Table 10: **Direct frame-selection quality on the KFS-Bench annotated subsets at $K = 32$.** Full results for Table 2, including BDS. Higher is better for all metrics; selectors share the candidate pool, BLIP scores, and annotated questions. Bold and underlined values denote the best and second-best results in each column, respectively, among the five selectors.

	Video-MME				LongVideoBench			
Selector	**KFR**	**BSR**	**BDS**	**UKSS**	**KFR**	**BSR**	**BDS**	**UKSS**
Candidate-grid uniform	0.2267	**0.8035**	**0.8514**	<u>0.3053</u>	0.1588	0.7222	0.7691	0.1975
Top-K	<u>0.2940</u>	0.7196	0.7747	0.2641	<u>0.1942</u>	0.7432	0.7898	0.2414
Projected inverse-CDF	0.2568	0.7823	0.8294	0.2936	0.1879	<u>0.7630</u>	<u>0.8101</u>	0.2448
RIDGE	0.2580	0.7362	0.7887	0.2640	0.1918	0.7622	0.8051	<u>0.2459</u>
E^2Q **(ours)**	**0.3015**	<u>0.7976</u>	<u>0.8489</u>	**0.3129**	**0.2170**	**0.7801**	**0.8234**	**0.2611**

Table 11 separates questions by the number of unique annotated `scene_id` values. One scene may contain several disjoint temporal segments; those segments are not counted as different scenes. Video-MME contains 614 single-scene and 274 multi-scene questions; LongVideoBench contains 961 single-scene and 330 multi-scene questions. We use the unmodified official KFS-Bench annotations

at snapshot `2349b9eb`, and UKSS is recomputed within each group. $\mathrm{E}^2\mathrm{Q}$ attains the highest UKSS in all four groups, most clearly on Video-MME multi-scene questions (0.3241 versus 0.2810).

Table 11: **Single- versus multi-scene UKSS on KFS-Bench at $K = 32$.** UKSS is recomputed within each group. Bold and underlined values mark the best and second-best results per column.

	Video-MME		LongVideoBench	
Selector	**Single-scene**	**Multi-scene**	**Single-scene**	**Multi-scene**
Top-K	0.2685	0.2544	0.2307	0.2757
Projected inverse-CDF	<u>0.2994</u>	<u>0.2810</u>	0.2269	<u>0.3052</u>
RIDGE	0.2705	0.2500	<u>0.2339</u>	0.2845
$\mathrm{E}^2\mathrm{Q}$ **(ours)**	**0.3081**	**0.3241**	**0.2446**	**0.3157**

G.2 LVBench Temporal References

For the 1,478 LVBench (Wang et al., 2025a) questions with valid temporal references, a direct hit occurs when a selected timestamp lies inside an annotated interval, including its endpoints. The per-question nearest gap is zero for a hit and otherwise the smallest distance from any selected timestamp to a reference interval; Table 12 reports its mean over all 1,478 questions. A hit measures temporal localization and does not establish that the selected frame contains all information required for the answer. $\mathrm{E}^2\mathrm{Q}$ attains the highest direct-hit rate (58.66%), 4.13 points above the next-best selector (Top-K). Its mean nearest gap (12.42 s) is close to the smallest value, obtained by projected inverse-CDF (11.09 s), and far below those of Top-K (88.49 s) and RIDGE (54.90 s).

Table 12: **Temporal localization on 1,478 LVBench questions with valid reference intervals at $K = 32$.** Bold and underlined values denote the best and second-best results.

Selector	**Direct hit (%) ↑**	**Mean nearest gap (s) ↓**
Candidate-grid uniform	35.39	19.57
Top-K	<u>54.53</u>	88.49
Projected inverse-CDF	54.47	**11.09**
RIDGE	52.50	54.90
$\mathrm{E}^2\mathrm{Q}$ **(ours)**	**58.66**	<u>12.42</u>

G.3 Case-Study Details

Figure 4 uses Video-MME question `857-1` from video `25Pt1AZO9EM` (33:09). The answer options are (A) check on the cattle herd, (B) pick up duck eggs, (C) check the growth of apple and chestnut trees, and (D) none of the above; the reference answer is A. The KFS-Bench annotation assigns $[456, 491] \cup [572, 593]$ seconds to the bicycle-transition scene and $[593, 617] \cup [630, 652] \cup [668, 770]$ seconds to the cattle-inspection scene. These intervals include the official evaluator's one-second extension of each annotated end time. Disjoint intervals with the same `scene_id` form one scene.

The displayed frames illustrate the chestnut-tree segment at 6:19, the bicycle transition at 7:45, and cattle evidence at 10:38 and 10:48. The recorded Qwen2.5-VL-7B answers are C for Top-K and projected inverse-CDF, and A for $\mathrm{E}^2\mathrm{Q}$.

H LIMITATIONS

E^2Q depends on the coverage of the candidate pool and on the reliability of frame–question relevance scores. Because it allocates frames by score magnitudes rather than ranks alone, a nonlinear monotone transformation of the scores can change its selection even though it preserves the ranking of candidates. It cannot recover observations absent from the candidates, and because the scorer rates each frame independently, frames whose relevance depends on motion or temporal context may receive low scores and be overlooked. Its relevance-mass intervals need not coincide with semantic events: a sustained event can receive multiple representatives, while complementary observations within one interval may not all be retained. The frame budget K is specified in advance rather than adapted to each question, and selecting fewer frames for the downstream model does not eliminate the cost of scoring the full candidate pool. Our representation analysis bounds a frame-level surrogate rather than downstream QA accuracy. Finally, our experiments cover visual-only multiple-choice long-video QA with 7B–8B open-source models; whether the same gains hold for other task formats, such as open-ended QA, and for larger models remains to be evaluated in future work.

I ADDITIONAL CASE STUDIES

We complement Figure 4 with five successful Video-MME cases (Figures 5–9) and one failure (Figure 10), using shared BLIP-scored candidates, $K = 32$, and Qwen2.5-VL-7B. Shading denotes KFS-Bench scenes; counts use exact timestamps and report in-scene frames and distinct scenes covered. These cases illustrate selection behavior, not how often it occurs.

I.1 SUCCESSFUL CASES

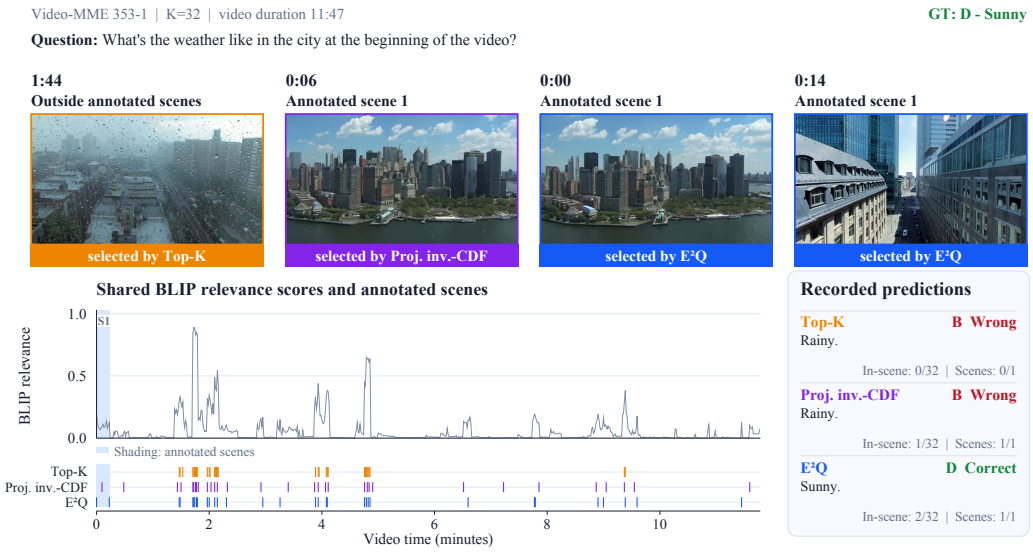


Figure 5: **Weather at the beginning of the video (353-1).** Top-K favors later rainy footage, misses the annotated scene, and yields B (rainy). E^{2}Q retains two opening views and yields D (sunny), the correct answer. Projected inverse-CDF retains one opening view, yet its recorded answer is also B.

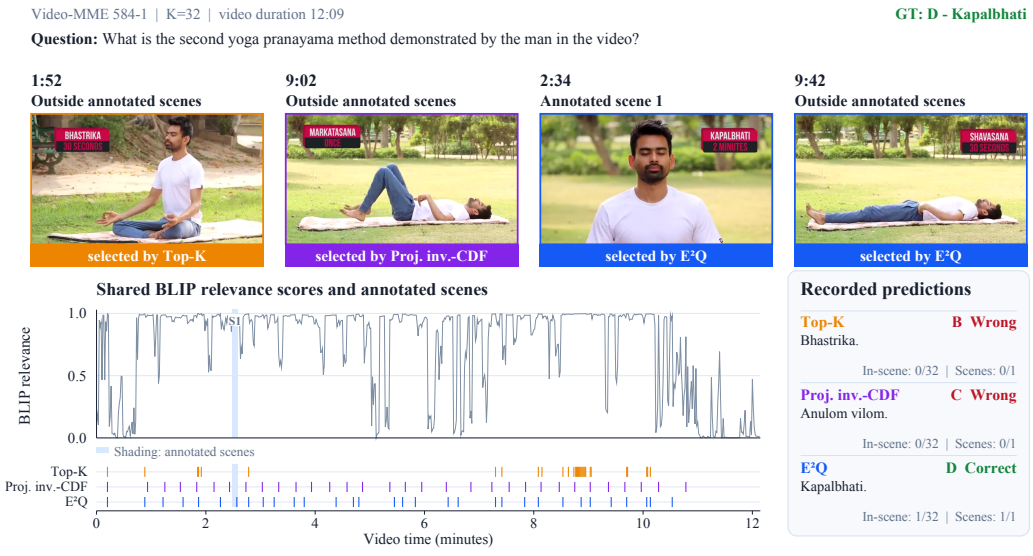


Figure 6: **A brief text cue identifies the requested exercise (584-1).** E^{2}Q selects a frame at 2:34 within the annotated 149–155 s interval, where the on-screen label reads "Kapalbhati," and yields the correct answer D. Neither Top-K nor projected inverse-CDF selects a frame inside that interval; their recorded answers are B (Bhastrika) and C (Anulom vilom), respectively. The example highlights the value of retaining a brief identifying cue among many visually similar exercise frames.



Figure 7: **Recovering the order of five projects (822-1).** Top-K places six frames in only three annotated scenes and yields D, $(b)(a)(d)(c)(e)$. E^{2}Q retains one frame in each of the five scenes and yields the correct sequence A, $(a)(b)(c)(d)(e)$. Projected inverse-CDF misses all five annotated scenes and yields C, $(b)(e)(a)(d)(c)$.

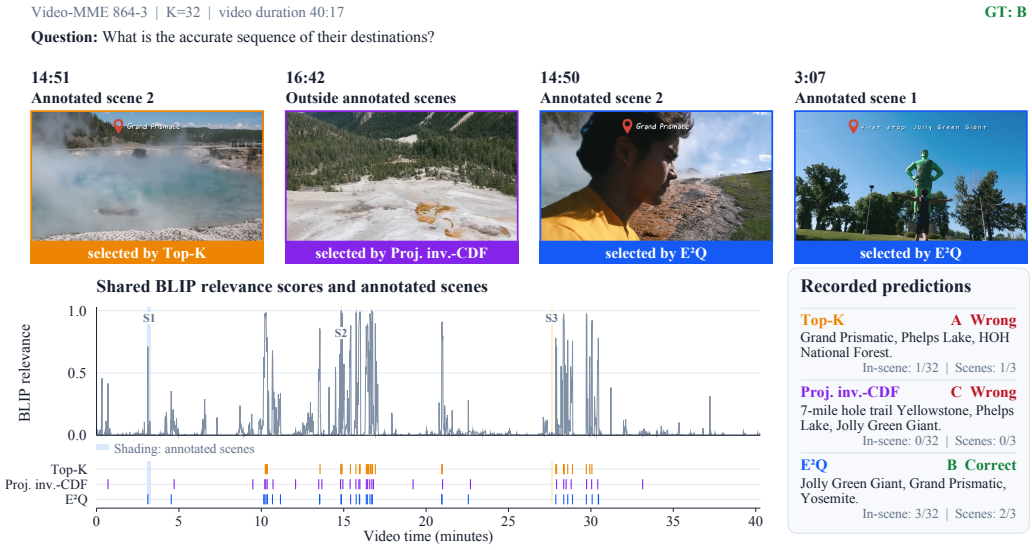


Figure 8: **Ordering travel destinations (864-3).** The reference sequence is B: Jolly Green Giant, Grand Prismatic, then Yosemite. E^{2}Q retains the displayed first-stop and Grand Prismatic cues and yields B, while Top-K and projected inverse-CDF yield A and C. Option A lists Grand Prismatic, Phelps Lake, and HOH National Forest; option C lists the 7-mile hole trail in Yellowstone, Phelps Lake, and Jolly Green Giant. The three in-scene frames selected by E^{2}Q cover two of the three annotated scenes, compared with one scene for Top-K and none for projected inverse-CDF. Here, the correct answer coincides with broader but still incomplete scene coverage: no selected frame falls in the final destination's annotated interval.



Figure 9: **Recognizing a weekly fitness schedule (883-3).** The correct option B specifies glutes and hamstrings on Monday, push on Tuesday, quads and calves on Wednesday, rest with a hair appointment on Thursday, pull on Friday, and cardio and core on Saturday. E^2Q retains the Wednesday and Saturday cues and yields B. Top-K concentrates its three in-scene frames in the Wednesday scene and yields A, swapping Tuesday's and Friday's workouts. Projected inverse-CDF yields C, additionally moving the rest day to Wednesday and the quads-and-calves workout to Thursday. E^2Q covers two of the six annotated scenes. These Wednesday and Saturday cues agree with both A and B, so they alone do not explain the correct prediction from the full 32-frame input.

I.2 Failure Case



Figure 10: **Missing short annotated intervals (362-2).** The answer is A (brush teeth before the hearing). Top-K covers all three annotated scenes and yields A, whereas E^2Q misses their narrow annotated intervals and yields D (eat before or after the hearing). Projected inverse-CDF also has no in-scene frames but yields A: its frame at 174.59 s already shows the eating instruction, just before the annotated interval begins at 175 s. Thus, missing an annotation window does not imply missing all relevant visual content, and temporal coverage alone does not explain the prediction difference.