ONE RANKING, ANY BUDGET: MATRYOSHKA EVIDENCE-TO-CONTEXT FRAME SELECTION FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

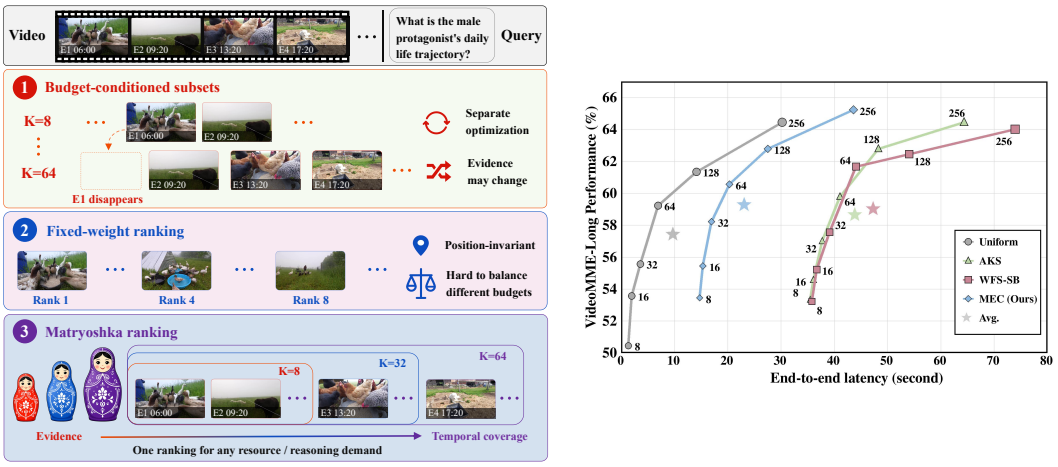
Frame selection is essential for applying Large Multimodal Models (LMMs) to long videos due to severe frame redundancy and limited context windows. Since the appropriate frame budget varies with the downstream LMM, reasoning demands, and latency constraints, a practical selector should serve multiple budgets. However, existing methods typically optimize an isolated frame subset for each predefined budget: when the budget changes, previously selected evidence may be replaced rather than progressively augmented. A fixed-weight ranking allows prefix reuse across budgets but applies the same weighting at every position, overlooking the distinct roles of early and later ranks. We formulate long-video frame selection as a Matryoshka ranking problem: constructing a single priority sequence whose small prefixes concentrate query-conditioned evidence, while progressively larger prefixes preserve this evidence and add broader temporal context. Efficiently constructing such a ranking is itself challenging, as densely sampling long videos and evaluating frame–query relevance incurs substantial overhead. We therefore introduce **M**atryoshka **E**vidence-to-**C**ontext (**MEC**) Frame Selection, a training-free framework that builds a reusable sparse video index, discovers candidates through sparse probing and local zooming, and greedily constructs a position-adaptive ranking—early positions emphasize evidence; later positions progressively favor temporal coverage while preserving visual diversity. A single ranking can thus be truncated to any target budget without rerunning the selector. Across four benchmarks and six frame budgets, **MEC improves average accuracy over uniform sampling by 3.77 points, matches strong state-of-the-art selectors, and reduces end-to-end selection latency by 47.37-51.19%.**

1 INTRODUCTION

Recent advances in Large Multimodal Models (LMMs) (Bai et al., 2025; Wang et al., 2025; Clark et al., 2026; An et al., 2026) have extended multimodal reasoning from static images to video. Applications include video question answering, temporal event localization, dense captioning, and video object tracking (Shaar et al., 2026; Zhang et al., 2026; Fiastre et al., 2026). Yet scaling from short clips to long-form video remains challenging. A long recording may contain tens of thousands of highly redundant frames, while the evidence needed to answer a question may be confined to a fleeting action, a subtle state transition, or several events separated by minutes.

Processing every frame is infeasible, yet indiscriminate temporal subsampling can miss brief, subtle, or temporally distant evidence. Keyframe selection retains a compact set of query-relevant frames for the downstream LMM. Agentic or iterative temporal search can adaptively locate evidence (Zhang et al., 2025d; Zou et al., 2026), but repeated decisions, VLM evaluations, and, where applicable, tool invocations introduce additional search steps and latency. Visual-token compression shortens token sequences through hierarchical aggregation, merging, or pruning (Li et al., 2026b; Fu et al., 2025b; Alvar et al., 2025), but aggressive compression may discard fine-grained, query-critical evidence. Frame selection reduces redundant input while preserving visual evidence at its original fidelity. What should a frame-selection strategy provide for long-video understanding?

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.



(a) Frame-selection paradigms **(b) Accuracy–latency trade-off**

Figure 1: Motivation and empirical trade-off of MEC. (a) Budget-conditioned selection optimizes each subset separately and may replace earlier evidence; fixed-weight ranking reuses prefixes but struggles to balance performance across budgets. MEC instead constructs one Matryoshka ranking whose prefixes preserve early evidence and progressively extend temporal context. (b) On Video-MME Long, each marker is annotated with its frame budget and stars denote six-budget averages. MEC provides a favorable accuracy–latency balance relative to stronger budget-specific selectors.

A selection strategy should remain effective as the frame budget changes. Many existing keyframe methods optimize a separate subset for each predefined budget K. Because the appropriate budget varies with the downstream LMM, reasoning demands, and latency constraints, changing K requires rerunning the selector. Moreover, as illustrated in Figure 1a, evidence selected under a tight budget may disappear as the budget increases. A fixed-weight ranking can serve different budgets through nested prefixes, but it uses the same balance of evidence, temporal coverage, and visual diversity at every rank, limiting its ability to concentrate evidence in short prefixes and expand context in longer ones. Inspired by the Matryoshka principle explored in representation learning and multimodal visual-token modeling (Kusupati et al., 2022; Cai et al., 2025), we formulate long-video frame selection as the *Matryoshka ranking problem*: the task of constructing one budget-independent priority sequence whose short prefixes concentrate query-conditioned evidence, while progressively longer prefixes preserve that evidence and add complementary context.

Constructing such a ranking efficiently poses a challenge. Densely sampling a long video and evaluating each frame against the query can incur substantial selection cost. We address this challenge with Matryoshka Evidence-to-Context (MEC) Frame Selection, a training-free framework that separates reusable video analysis from query-conditioned discovery. MEC first builds a sparse index of low-resolution appearance, local visual change, and observability. For each query, it scores sparse probes using multiple cues, activates high-value segments, and zooms around anchors to form a compact candidate pool. It then greedily constructs the ranking using a position-dependent marginal surrogate: early positions emphasize query-conditioned evidence, whereas later positions progressively increase temporal coverage while maintaining visual diversity. Any target budget can be served by truncating this ranking and restoring chronological order, without rerunning the selector. Across four long-video benchmarks and six frame budgets, MEC remains competitive with strong methods while offering a favorable accuracy–latency trade-off. As shown in Figure 1b, **MEC's six-budget average on the Long subset of Video-MME exceeds AKS by** 0.63 **accuracy points while reducing cached end-to-end average latency by** 47.37%.

Our contributions are as follows: **1)** We formulate the Matryoshka ranking problem for long-video frame selection: one sequence serves all budgets, concentrating query-conditioned evidence in short prefixes and progressively expanding context in longer ones. **2)** We introduce MEC, a training-free framework that efficiently constructs a position-adaptive Matryoshka ranking to serve multiple frame budgets in a single run. **3)** Extensive experiments across four long-video benchmarks and six frame budgets demonstrate the effectiveness of MEC and its favorable accuracy–latency trade-off.

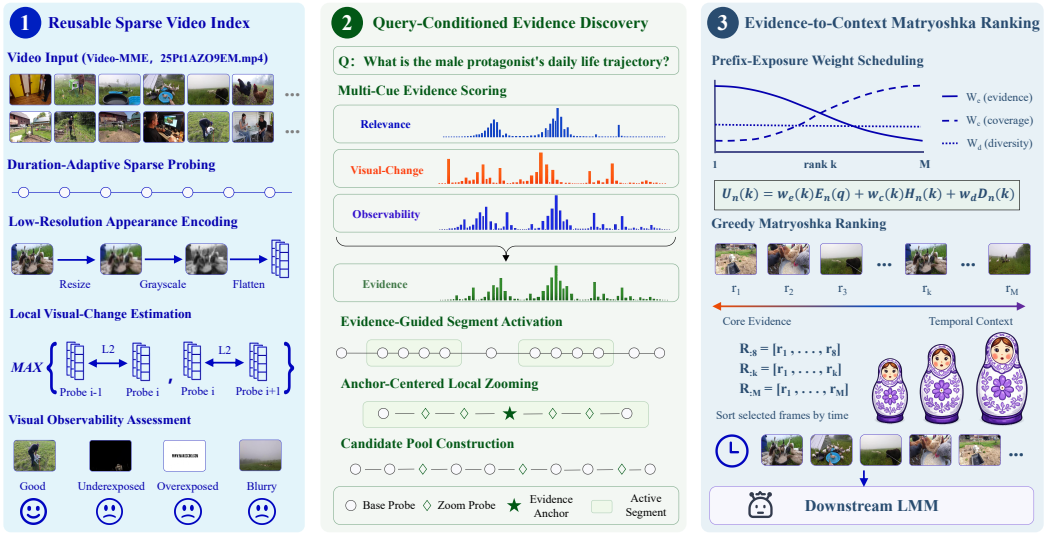


Figure 2: Overview of MEC. A reusable sparse video index caches low-resolution appearance encodings, local visual-change estimates, and observability priors. For each query, MEC scores multi-cue evidence, activates promising segments, and performs local zooming before constructing a position-adaptive Matryoshka ranking whose nested prefixes support any frame budget.

2 Related Work

Frame Selection for Long-Video Understanding. Frame selection fits long videos into the limited context windows of LMMs. Existing methods fall into several overlapping categories. Query-aware sampling prioritizes relevant frames, optionally balancing coverage or diversity, as in AKS, Nar-KFC, BOLT, MDP^3, and AdaRD-Key (Tang et al., 2025; Fang et al., 2026; Liu et al., 2025a; Sun et al., 2025; Zhang et al., 2025c). Structure-aware selection partitions or hierarchically organizes a video into clips, events, or scenes before allocating or refining frames, as in EFS, WFS-SB, and InfoShot (Chen et al., 2026a;b; Zhao et al., 2026); DIG further adapts selection to global or localized queries (Li et al., 2026a). Temporal search probes promising regions or iteratively acquires evidence, as in AKeyS, A.I.R., FOCUS, and FrameThinker (Fan et al., 2025; Zou et al., 2026; Zhu et al., 2026; He et al., 2026). Learned selectors use model-generated supervision or reinforcement learning, including the M-LLM-based selector of Hu et al. (2025), the RL-based selector of Qin et al. (2026), and ReaSon (Zhou et al., 2026). Frame/token co-optimization determines which temporal content to retain and how many visual tokens or what resolution to assign to it, as in Q-Frame, KVTP, and QuoTA (Zhang et al., 2025b; Liu et al., 2025b; Luo et al., 2025). These categories overlap, emphasizing subset quality, search decisions, or token allocation. In contrast, MEC constructs a single reusable nested sequence using a position-dependent marginal surrogate, concentrating evidence in early ranks while progressively expanding temporal context in later ranks. Section A further discusses video large multimodal models, Matryoshka representations, and visual-token compression.

3 Method

Figure 2 summarizes MEC's three stages: building a reusable sparse index, discovering query-conditioned evidence, and constructing an evidence-to-context Matryoshka ranking.

3.1 Problem Formulation

Let $\mathcal{V}=\{x_1,\ldots,x_N\}$ be a video, where x_n is its n-th frame and N is the total number of frames, and let q be a question about the video. Given a frame budget K, a predefined-budget selector returns

$$\mathcal{F}_K = \Phi(\mathcal{V}, q, K), \qquad |\mathcal{F}_K| = K, \tag{1}$$

where $\mathcal{F}_K$ is optimized for K, so changing the budget requires reevaluating Φ. MEC instead invokes its selector once with a budget-independent sequence length $1 \le M \le N$:

$$\begin{aligned} \mathcal{R} &= \Psi(\mathcal{V}, q, M) = (r_1, r_2, \ldots, r_M), \\ \mathcal{F}_K^{\mathrm{MEC}} &= \{r_1, \ldots, r_K\}, \qquad 1 \le K \le M. \end{aligned} \tag{2}$$

Here r_k is the original-frame index at rank k, and all r_k are distinct. The selected sets are nested across budgets, but nesting alone does not ensure useful prefixes. Let $\mathcal{B} \subseteq \{1, \ldots, M\}$ be the supported budgets and $\pi_K \geq 0$ their deployment importance, with $\sum_{K \in \mathcal{B}} \pi_K = 1$. We use $\mathcal{Q}_K(\mathcal{R}_{:K}; \mathcal{V}, q)$ to denote the ideal downstream utility of prefix $\mathcal{R}_{:K} = (r_1, \ldots, r_K)$. This oracle utility is unavailable: correctness is unknown at inference time, and repeatedly evaluating candidate prefixes with the LMM is prohibitively expensive. The ideal Matryoshka objective is

$$\mathcal{R}^\star \in \arg\max_{\mathcal{R} \in \Pi_M} \sum_{K \in \mathcal{B}} \pi_K \, \mathcal{Q}_K(\mathcal{R}_{:K}; \mathcal{V}, q), \tag{3}$$

where Π_M contains all sequences of M distinct frame indices. Equation (3) specifies the desired all-prefix ranking, not a directly computable selector loss. A frame at rank k appears in every supported prefix with $K \geq k$, giving cross-budget exposure $\alpha_k = \sum_{K \in \mathcal{B},\, K \geq k} \pi_K$ with $\alpha_1 \geq \cdots \geq \alpha_M$. Early positions therefore affect a broader range of budgets. This positional asymmetry motivates prioritizing evidence useful under both tight and large budgets before expanding temporal and visual context. Section 3.4 implements this structure through a tractable marginal surrogate.

3.2 Reusable Sparse Video Index

Analyzing all N frames for each question repeats video-level computation. MEC caches appearance, change, and observability cues in a sparse index independent of q.

Duration-Adaptive Sparse Probing. For duration T in seconds, base probe count P_0, maximum increment P_1, and saturation horizon τ_p, we use the uniform probe count

$$P = \min\left\{ N, \left\lfloor P_0 + \min\left(\frac{T}{\tau_p}, 1\right) P_1 \right\rfloor \right\}. \tag{4}$$

We use P_0=P_1=256 and τ_p=1800 seconds. This schedule increases temporal resolution for longer videos while bounding the index size. Let $\mathcal{P}$=$\{p_1, \ldots, p_P\}$ be the temporally ordered original-frame indices of the probes. Consecutive groups of L probe intervals form segments $\{\mathcal{G}_1, \ldots, \mathcal{G}_J\}$; we set L=4. These segments provide coarse temporal units for later query-conditioned refinement.

Low-Resolution Appearance Encoding. We resize each probe to a short side of $d = 32$ pixels while preserving its aspect ratio, convert it to grayscale, flatten it, and apply ℓ_2 normalization to obtain $\mathbf{f}_{p_i}$. These cached descriptors support inexpensive appearance comparisons (Section C.1).

Local Visual-Change Estimation. Let $\mathrm{Adj}_{\mathcal{P}}(n)$ contain the available probes immediately before and after frame index n on the sparse grid. We define local visual-change as

$$\Delta_n = \frac{1}{2} \max_{m \in \mathrm{Adj}_{\mathcal{P}}(n)} \|\mathbf{f}_n - \mathbf{f}_m\|_2. \tag{5}$$

This lightweight, query-independent score highlights probes that differ from their temporal neighbors, directing evidence discovery toward visually distinctive moments.

Visual Observability Assessment. We associate each frame with sharpness $s_n^{\mathrm{sharp}} \in [0, 1]$ and well-exposedness $\ell_n^{\mathrm{light}} \in [0, 1]$. Following Crete et al. (2007), sharpness measures how strongly directional local variation is attenuated by controlled reblurring; following Mertens et al. (2007), well-exposedness downweights pixels whose intensities are far from the mid-range, including severely dark or saturated regions. Section C.2 gives their detailed definitions. We define the visual observability prior as O_n=$s_n^{\mathrm{sharp}} \ell_n^{\mathrm{light}}$. The product requires both adequate sharpness and exposure, preventing one component from compensating for failure of the other.

3.3 Query-Conditioned Evidence Discovery

The index summarizes the video, while the question guides the search for answer-supporting observations. MEC scores the probes and refines only the most promising regions.

Multi-Cue Evidence Scoring. Let $\rho_n(q) \in [0, 1]$ be the frame–question matching probability produced by the BLIP-2 ITM head (Li et al., 2023). We define evidence as the suitability of a frame to serve as an effective observation for q within a compact visual context:

$$E_n(q) = \lambda_r \rho_n(q) + \lambda_\Delta \Delta_n + \lambda_o O_n, \tag{6}$$

where λ_r=0.7, λ_Δ=0.2, and λ_o=0.1. Relevance aligns frames with the question, visual change favors distinctive local appearances, and observability downweights visually unclear frames.

Evidence-Guided Segment Activation. For segment $\mathcal{G}_j$, let $\bar{E}_j$ be its mean probe evidence score. Given an activation ratio $\eta \in (0, 1]$, we retain the $\lceil \eta J \rceil$ segments with the largest $\bar{E}_j$, forming the active set $\mathcal{A}(q)$. We use η=0.25. Aggregating scores by segment reduces sensitivity to isolated high scores and limits fine-grained search to a small part of the timeline.

Anchor-Centered Local Zooming. In each active segment $\mathcal{G}_j \in \mathcal{A}(q)$, the highest-evidence probe serves as the anchor a_j. Its available preceding and following probes, a_j^- and a_j^+, delimit the two local intervals to be refined. The operator $\mathrm{Uni}(a, b, L_z)$ samples L_z distinct, uniformly spaced interior frame indices between a and b, producing the anchor-centered zoom set

$$\mathcal{Z}_j(q) = \mathrm{Uni}(a_j^-, a_j, L_z) \cup \mathrm{Uni}(a_j, a_j^+, L_z). \tag{7}$$

We set L_z=2. Refining both sides of the anchor introduces fine-grained zoom probes that may capture informative observations missed by the global sparse grid.

Candidate Pool Construction. The candidate pool $\mathcal{C}(q)$ combines sparse probes $\mathcal{P}$ and query-conditioned zoom probes $\mathcal{Z}(q)$: $\mathcal{C}(q) = \mathcal{P} \cup \mathcal{Z}(q)$. For each zoom probe, $\mathbf{f}_n$, Δ_n, O_n, $\rho_n(q)$, and $E_n(q)$ are computed as above. This budget-independent pool supports ranking lengths $M \leq |\mathcal{C}(q)|$.

3.4 Evidence-to-Context Matryoshka Ranking

We approximate the marginal decision implied by the all-prefix objective, rather than downstream utility itself. The surrogate combines query-conditioned evidence with temporal and visual information absent from the current prefix. Section 3.1 motivates their relative weights.

Temporal Coverage and Visual Diversity Assessment. For rank $k \geq 2$, the selected prefix is $\mathcal{S}_{k-1}$=$\{r_1, \ldots, r_{k-1}\}$. An unselected candidate n is assessed by

$$H_n(k) = \min\left(1, \frac{\min_{m \in \mathcal{S}_{k-1}} |n - m|}{N/k}\right), \tag{8}$$

$$D_n(k) = 1 - \max_{m \in \mathcal{S}_{k-1}} \mathbf{f}_n^\top \mathbf{f}_m. \tag{9}$$

The interval N/k normalizes distance to the nearest selected frame. The marginal context cues $H_n(k)$ and $D_n(k)$ favor temporal regions and appearances absent from $\mathcal{S}_{k-1}$, respectively.

Prefix-Exposure Weight Scheduling. The discrete exposure α_k is a non-increasing step profile determined by $\mathcal{B}$ and $\{\pi_K\}$. We use $1 - (k/M)^\gamma$ as a smooth profile shared across budget sets. In a continuous relaxation, this is the survival function of the normalized-budget CDF $F(x) = x^\gamma$ on $[0, 1]$; $\gamma = 1$ gives a uniform budget prior and linear decay. The evidence weight follows this prior:

$$w_e(k) = w_e^- + (w_e^+ - w_e^-)\left[1 - \left(\frac{k}{M}\right)^\gamma\right], \tag{10}$$

Here, w_e^+ and w_e^- specify the upper and lower evidence-weight references, and $\gamma > 0$ controls the transition rate. The positive floor w_e^- retains evidence sensitivity at large budgets. Temporal coverage receives the remaining adaptive weight, $w_c(k) = 1 - w_e(k) - w_d$, while w_d stays fixed to control visual redundancy throughout the ranking. We use w_e^+=0.6, w_e^-=0.2, γ=1, and w_d=0.2. Early ranks emphasize evidence; later ranks favor complementary temporal context.

Greedy Matryoshka Ranking. At rank $k \geq 2$, we score each remaining candidate n as

$$u_k(n) = w_e(k)E_n(q) + w_c(k)H_n(k) + w_d D_n(k). \tag{11}$$

We interpret $u_k(n)$ as a tractable proxy for the unobservable marginal contribution of assigning n to rank k under Equation (3), conditioned on $\mathcal{S}_{k-1}$. Evidence $E_n(q)$ captures answer-supporting potential, while $H_n(k)$ and $D_n(k)$ capture complementarity to the current prefix. The score is position-aware through its weights and set-aware through these context terms. Starting from the highest-evidence candidate, the ranking greedily appends the candidate with the largest $u_k(n)$. Selected frames are never replaced, ensuring exact nesting without combinatorial lookahead. Algorithm 1 gives the complete procedure; Section 4.3 evaluates the surrogate structure empirically. For any budget $K \leq M$, MEC chronologically sorts $\mathcal{R}_{:K} = (r_1, \ldots, r_K)$ before LMM inference.

Table 1: Comparison of frame selectors using Qwen3.5-9B. All values are percentages. Video-MME-v2 reports the nonlinear group score (Non-Lin.) and micro accuracy (Acc.); LVB abbreviates LongVideoBench, and MLVU reports macro accuracy. Best results within each budget are bold.

K	Selector	Video-MME				Video-MME-v2		LVB	MLVU
		Short	Medium	Long	Overall	Non-Lin.	Acc.		
8	Uniform	69.00	57.44	50.44	58.96	8.59	23.84	54.15	57.40
8	AKS	70.89	62.00	53.33	62.07	10.59	**26.19**	61.18	66.40
8	WFS-SB	73.22	64.11	53.22	63.52	**10.93**	26.06	62.00	68.33
8	**MEC** *(ours)*	**73.78**	**64.33**	**53.44**	**63.85**	10.81	26.13	**62.45**	**69.46**
16	Uniform	73.67	60.67	53.56	62.63	10.06	25.44	57.52	60.29
16	AKS	75.67	65.00	54.56	65.07	11.30	27.03	60.81	68.65
16	WFS-SB	76.56	65.56	55.22	65.78	**12.23**	**28.22**	61.93	**68.94**
16	**MEC** *(ours)*	**77.22**	**66.89**	**55.44**	**66.52**	12.07	27.78	**62.30**	68.39
32	Uniform	77.78	65.11	55.56	66.15	11.41	27.47	58.86	65.00
32	AKS	79.00	68.67	57.00	68.22	12.95	29.50	61.33	70.43
32	WFS-SB	**79.56**	68.56	57.56	68.56	**13.42**	**29.69**	63.35	**70.90**
32	**MEC** *(ours)*	78.78	**70.33**	**58.22**	**69.11**	13.05	29.47	**63.58**	70.36
64	Uniform	80.44	68.78	59.22	69.48	12.79	28.78	60.73	69.69
64	AKS	81.33	72.33	59.78	71.15	14.23	31.00	64.62	72.46
64	WFS-SB	80.67	71.44	**61.67**	71.26	14.64	31.31	**66.57**	73.83
64	**MEC** *(ours)*	**81.89**	**73.00**	60.56	**71.81**	**14.77**	**31.69**	65.89	**74.30**
128	Uniform	**83.00**	73.11	61.33	72.48	14.33	31.00	65.00	73.77
128	AKS	82.56	74.67	**62.78**	73.33	15.50	32.47	67.61	74.88
128	WFS-SB	82.56	75.11	62.44	73.37	**15.71**	**33.16**	68.06	**75.85**
128	**MEC** *(ours)*	82.56	**75.89**	**62.78**	**73.74**	15.42	33.00	**68.36**	**75.85**
256	Uniform	82.33	74.33	64.44	73.70	15.78	32.75	66.87	76.01
256	AKS	82.67	75.89	64.44	74.33	16.31	33.75	68.21	76.19
256	WFS-SB	82.67	**76.11**	64.00	74.26	16.81	34.56	**69.78**	**77.01**
256	**MEC** *(ours)*	**83.11**	75.56	**65.22**	**74.63**	**17.08**	**34.63**	68.51	76.65

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Benchmarks. We evaluate on four complementary long-video benchmarks. Video-MME contains 900 videos and 2,700 multiple-choice questions across short-, medium-, and long-duration groups (Fu et al., 2025a). Video-MME-v2 contains 800 videos and 3,200 questions organized around information aggregation, temporal dynamics, and complex reasoning; in addition to micro accuracy, its nonlinear group score measures consistency across related questions (Fu et al., 2026). LongVideoBench contains 3,763 videos and 6,678 questions that emphasize long-context referring and temporal reasoning (Wu et al., 2024). MLVU contains 1,730 videos, 3,102 question–answer pairs, and nine tasks spanning global and local understanding (Zhou et al., 2025). Section B.1 provides additional benchmark scope and task details.

Implementation Details. We evaluate with LMMs-Eval (Zhang et al., 2025a) on 8 NVIDIA H20 GPUs. The downstream LMMs are Qwen3.5-9B, LLaVA-OneVision-2-8B-Instruct, and InternVL3.5-8B-Instruct. We compare with uniform sampling, AKS (Tang et al., 2025), and WFS-SB (Chen et al., 2026b). All methods use the same question text, downstream prompt and dataset-specific scoring protocol. Every MEC experiment uses one shared configuration across all benchmarks, downstream LMMs, and frame budgets, without model-, dataset-, or budget-specific tuning. We set M=256, $(\lambda_r, \lambda_\Delta, \lambda_o)$=$(0.7, 0.2, 0.1)$, and $(w_e^+, w_e^-, w_d, \gamma)$=$(0.6, 0.2, 0.2, 1)$. Baselines follow their released procedures and default settings. Table 5 gives the full configuration.

4.2 MAIN RESULTS

Comparison with State-of-the-Art. We first compare MEC with training-free frame selectors using Qwen3.5-9B over six budgets. Table 1 shows that the benefit of MEC extends across datasets and budgets. Averaging Video-MME Overall, Video-MME-v2 micro accuracy, LongVideoBench

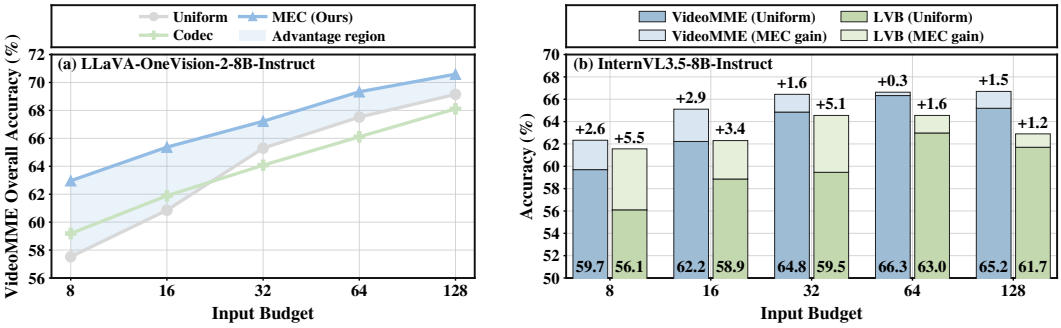


Figure 3: Generalization across downstream LMMs and input budgets $K \in \{8, 16, 32, 64, 128\}$. (a) Video-MME Overall with LLaVA-OneVision-2-8B-Instruct, comparing uniform frames, native codec-stream input, and MEC. (b) Uniform accuracy and the corresponding MEC gain on Video-MME and LongVideoBench (LVB) with InternVL3.5-8B-Instruct.

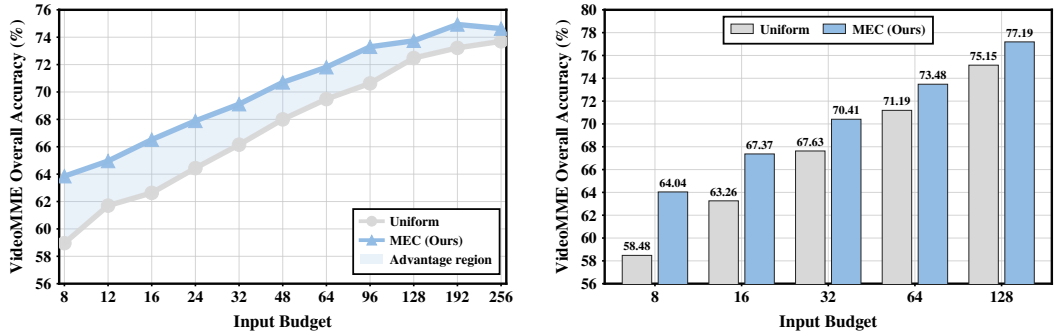


Figure 4: Budget behavior and thinking-mode evaluation on Video-MME Overall with Qwen3.5-9B. Left: Uniform and MEC at 11 budgets; shading marks the MEC gain. Right: the two selectors at five budgets with thinking enabled.

accuracy, and MLVU accuracy over all six budgets, MEC improves uniform sampling by 3.77 pp. The advantage is especially pronounced for tight prefixes: at K=8, the gains over uniform sampling are 8.30 pp on LongVideoBench and 12.06 pp on MLVU. Against stronger selectors, MEC obtains the best Video-MME Overall result at every budget, including 66.52 versus 65.78 at K=16. WFS-SB remains strongest in several Video-MME-v2, LongVideoBench, and MLVU settings, but MEC stays competitive while uniquely serving all budgets with one shared ranking.

Generalization Across LMMs. We next test whether the same selector configuration transfers across architectures and input representations. With LLaVA-OneVision-2, MEC outperforms uniform frames and codec-stream input at every budget. Codec leads uniform sampling at K=8 and 16, but falls behind from K=32. Codec-stream tokenization follows bit-cost dynamics and motion-residual cues to concentrate capacity on event-bearing intervals (An et al., 2026). We interpret the crossover as a possible consequence of this allocation: it may favor sparse motion-relevant evidence at small budgets, while additional decoded frames may preserve more local texture, fine-grained static evidence, and continuous trajectories. MEC remains strongest across the crossover, showing that query-conditioned evidence discovery and evidence-to-context organization improve over model-specific input allocation while retaining original-frame evidence. The InternVL3.5 results in Figure 3(b) provide a complementary test. MEC improves over uniform sampling at all five budgets on both Video-MME and LongVideoBench, despite using exactly the same configuration as the other experiments. The gains transfer across downstream architecture, benchmark, and context size rather than being tied to Qwen3.5-9B or a single evaluation set.

4.3 Further Quantitative Analyses

Fine-Grained Budget Behavior. We evaluate eleven budgets to determine whether the gains persist between the six operating points used in Table 1. Figure 4 shows a positive gain at every budget from K=8 to K=256, including five intermediate values absent from Table 1. The margin is largest

Table 2: Component and ranking-schedule ablations on Video-MME Overall with Qwen3.5-9B. Low, Med., and High average $K = \{8, 16\}$, $\{32, 64\}$, and $\{128, 256\}$; Avg. covers all six budgets. In (b), fixed-weight tuples give (w_e, w_c, w_d); decay schedules use $(w_e^+, w_e^-) = (.6, .2)$, $w_d = .2$ and $w_c(k) = 1 - w_e(k) - w_d$. Best values in each panel are bold.

(a) Component contributions

Variant	Low	Med.	High	Avg.
Uniform	60.80	67.81	73.09	67.23
MEC	**65.19**	**70.46**	**74.19**	**69.94**
w/o coverage	61.69	68.13	73.63	67.81
w/o diversity	64.63	70.06	73.72	69.47
w/o local zoom	64.26	69.94	73.91	69.37
w/o visual change	64.11	69.59	73.85	69.19
w/o observability	64.44	70.26	74.04	69.58

(b) Ranking-weight schedules

Schedule	Low	Med.	High	Avg.
Uniform	60.80	67.81	73.09	67.23
MEC ($\gamma = 1$)	**65.19**	70.46	74.19	**69.94**
Fixed weights $(.6, .2, .2)$	64.89	70.09	73.56	69.51
Fixed weights $(.2, .6, .2)$	63.44	69.46	74.06	68.99
Fixed weights $(.4, .4, .2)$	64.66	70.15	73.82	69.54
Decay $\gamma = 1.5$	64.70	**70.61**	**74.22**	69.85
Decay $\gamma = 2/3$	64.49	70.22	73.89	69.60

Table 3: Ranking horizon M vs. budget K on Video-MME Overall. Bold: column best; –: $K > M$.

$M \backslash K$	8	16	32	64	128	256
8	62.44	–	–	–	–	–
16	63.48	**66.70**	–	–	–	–
32	**64.04**	65.78	68.26	–	–	–
64	63.89	66.11	**69.22**	71.04	–	–
128	63.52	66.00	68.52	71.70	73.26	–
256	63.85	66.52	69.11	**71.81**	**73.74**	**74.63**

Table 4: Mean time costs over six budgets on Video-MME Long.

	MEC	AKS	WFS-SB
Candidates	**635.40**	2,466.82	2,466.82
Prep. (s)	**5.48**	12.09	12.09
Sel. (s)	**7.89**	22.06	25.50
E2E (s)	**23.09**	43.87	47.31

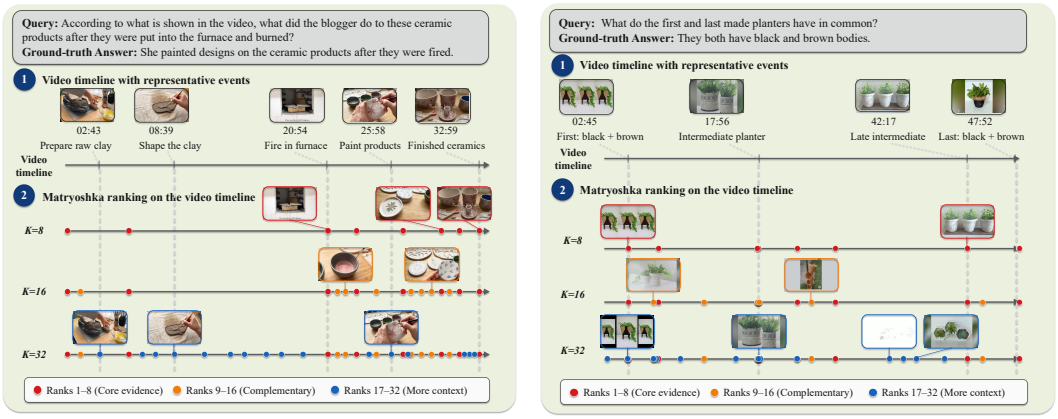
Prep.: preprocessing; Sel.: selection; E2E: end-to-end time per query.

at K=8 (+4.89 pp) and remains positive at K=256 (+0.93 pp). Although the gap is not strictly monotonic, it generally narrows as the input grows: tight prefixes benefit most from concentrating diagnostic evidence, whereas larger inputs allow uniform sampling to recover more context. The persistent residual gain indicates that the later ranks of MEC continue to organize context more effectively rather than merely exploiting a few favorable budget choices.

Thinking-Mode Evaluation. For Figure 4 (right), we enable Qwen3.5 thinking mode through its official interface (Qwen Team, 2026); standard evaluations disable it. MEC improves over Uniform at all five tested budgets, with gains decreasing from 5.56 pp at $K = 8$ to 2.04 pp at $K = 128$ (Section E.3). Frame selection remains useful with downstream reasoning enabled.

Component and Ranking Ablations. Table 2 shows the complementary contributions of candidate discovery and prefix construction. Removing coverage, diversity, local zoom, visual change Δ_n, or observability O_n reduces average accuracy by 2.13, 0.47, 0.57, 0.75, and 0.36 pp, respectively. The full configuration improves Low, Medium, and High accuracy over Uniform by 4.39, 2.65, and 1.10 pp. With candidate discovery and downstream evaluation fixed, the position-adaptive ranking exceeds all three fixed-weight configurations by 0.40–0.95 pp on average, supporting different surrogate weights across positions under Equation (3). Table 6 further shows that reversing the evidence-to-context transition reduces average accuracy from 69.94 to 69.00, whereas the five tested evidence-decay schedules span only 0.46 pp across decay rates and endpoints. These results support the evidence-first, context-later structure more strongly than any particular decay curve. Together with the cross-model and cross-benchmark results in Figure 3 and Table 1, they support using one default configuration across models, benchmarks, and budgets.

Effect of Ranking Horizon. Table 3 compares a shared long ranking with rankings terminated at each target budget. Setting $M = K$ completes the evidence-to-context transition within that budget; $M = 256$ keeps short prefixes focused on evidence and reserves later positions for context expansion. Across $K \in \{8, 16, 32, 64, 128\}$, the shared ranking improves mean accuracy by 0.67 pp over the $M = K$ diagonal, including gains of 1.41, 0.85, 0.77, and 0.48 pp at $K = 8, 32, 64$, and 128. It achieves the best results at $K = 64$ and 128 and stays within 0.19 pp of the best at each smaller budget. A single Matryoshka ranking therefore provides an accuracy benefit across budgets while eliminating the need to construct a new sequence for each one.



(a) Evidence-to-context expansion

(b) Missed endpoint evidence

Figure 5: Success and failure of MEC. (a) Larger prefixes retain ceramic-making evidence and add earlier preparation stages. (b) Sparse discovery misses a clear view of the last completed planter, leaving the first–last comparison unresolved. Red, orange, and blue mark frames introduced at ranks 1–8, 9–16, and 17–32. All prefixes share one ranking and are shown chronologically.

Selector Efficiency. Table 4 reports time costs on Video-MME Long with cached RGB frames that can be reused across questions about the same video. For a fair comparison at each budget, the reported Sel. cost includes an independent run of MEC selection for every evaluated budget. In deployment, changing the budget for the same query incurs no additional selector cost: the existing ranking is truncated to the required prefix and the selected frames are restored to chronological order. MEC evaluates 635.40 candidates on average, 74.24% fewer than AKS and WFS-SB, and reduces preprocessing from 12.09 to 5.48 seconds per query. Selection takes 7.89 seconds, compared with 22.06 for AKS and 25.50 for WFS-SB, reductions of 64.23% and 69.06%. Including 9.72 seconds of LMM inference, end-to-end latency falls from 43.87 and 47.31 seconds to 23.09 seconds, respectively. These savings accompany the accuracy gains in Figure 1b.

4.4 Qualitative Analysis

In Figure 5a, the $K = 8$ prefix concentrates on the furnace, painted surfaces, and finished ceramics, providing the evidence that painting follows firing. The $K = 16$ prefix retains these frames and adds pigment and complementary product views; $K = 32$ further covers raw-clay preparation and shaping. Larger budgets reveal more of the manufacturing process while preserving answer-relevant observations. Figure 5b shows a first–last comparison in a 48.3-minute video. The first completed planter at 164.6 s and the last at 2871.7 s both have black-and-brown bodies. Sparse probes capture the first planter, intermediate products, and observations near the end, but miss a clear view of the last completed planter. This brief state receives no anchor for local zooming and never enters the candidate pool. The LMM answers incorrectly with MEC at all six budgets, while Uniform yields correct answers from $K = 32$ onward. All prefixes draw from the same pool, so later ranks add coverage and diversity only among discovered frames. Candidate recall, particularly for brief endpoint events, is the bottleneck: larger prefixes cannot recover evidence missed during discovery.

5 Conclusion

We formulate long-video frame selection as a Matryoshka ranking problem: one query-conditioned sequence serves multiple budgets, concentrating evidence in short prefixes and expanding context in longer ones. We introduce MEC, a training-free framework that combines reusable sparse indexing, query-conditioned evidence discovery, and position-adaptive ranking to construct this sequence efficiently. Experiments across four benchmarks, three downstream LMMs, and six frame budgets validate its effectiveness, generalization, and favorable accuracy–latency trade-off. By decoupling frame selection from the target budget, MEC offers a practical foundation for long-video systems that must adapt their visual context to changing deployment constraints.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript for clarity and readability and to assist in searching for and identifying relevant literature. We take responsibility for the final content of this work, including all AI-assisted text and citations.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025. doi: 10.1109/CVPR52734.2025.00877. URL https://arxiv.org/abs/2503.02175.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. arXiv preprint arXiv:2605.25979, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/722fcbc1a6667f2075d75ea79a1b3552-Abstract-Conference.html.

Wang Chen, Yongdong Luo, Yuhui Zeng, Luojun Lin, Tianyu Xie, Fei Chao, Rongrong Ji, and Xiawu Zheng. Event-anchored frame selection for effective long-video understanding. arXiv preprint arXiv:2603.00983, 2026a. URL https://arxiv.org/abs/2603.00983.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24052–24061, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Wavelet-based_Frame_Selection_by_Detecting_Semantic_Boundary_for_Long_Video_CVPR_2026_paper.html.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611, 2026. URL https://arxiv.org/abs/2601.10611.

Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. The blur effect: Perception and estimation with a new no-reference perceptual blur metric. In *Human Vision and Electronic Imaging XII*, volume 6492, pp. 64920I. SPIE, 2007. doi: 10.1117/12.702790.

Devvrit, Sneha Kudugunta, Aditya Kusupati, Tim Dettmers, Kaifeng Chen, Inderjit Dhillon, Yulia Tsvetkov, Hannaneh Hajishirzi, Sham Kakade, Ali Farhadi, and Prateek Jain. MatFormer: Nested transformer for elastic inference. In *Advances in Neural Information Processing Systems*, volume 37, pp. 140535–140564, 2024. doi: 10.52202/079017-4461. URL https:

//proceedings.neurips.cc/paper_files/paper/2024/hash/fe066022bab2a6c6a3c57032a1623c70-Abstract-Conference.html.

Sunqi Fan, Meng-Hao Guo, and Shuojin Yang. Agentic keyframe search for video question answering. arXiv preprint arXiv:2503.16032, 2025. URL https://arxiv.org/abs/2503.16032.

Bo Fang, Yuxin Song, Haoyuan Sun, Qiangqiang Wu, Wenhao Wu, and Antoni B. Chan. Threading keyframe with narratives: MLLMs as strong long video comprehenders. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/675a31be2d051be53d57bced52c6631f-Abstract-Conference.html.

Gabriel Fiastre, Antoine Yang, and Cordelia Schmid. CaptionFormer: Unified segmentation, tracking, and captioning for spatio-temporal objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39518–39528, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fiastre_CaptionFormer_Unified_Segmentation_Tracking_and_Captioning_for_Spatio-Temporal_Objects_CVPR_2026_paper.html.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025a. URL https://arxiv.org/abs/2405.21075.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the next stage in benchmarks for comprehensive video understanding. arXiv preprint arXiv:2604.05015, 2026. URL https://arxiv.org/abs/2604.05015.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. FrameFusion: Combining similarity and importance for video token reduction on large vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025b. doi: 10.1109/ICCV51701.2025.02103. URL https://arxiv.org/abs/2501.01986.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79ba1b827d3fc58e129d1cbfc8ff69f2-Abstract-Conference.html.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13702–13712, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_M-LLM_Based_Video_Frame_Selection_for_Efficient_Video_Understanding_CVPR_2025_paper.html.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50168–50188, 2024. doi: 10.52202/079017-1588. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/59c147c7d4fdb732daea3064eab949bf-Abstract-Conference.html.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11369–11380, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Divide_then_Ground_Adapting_Frame_Selection_to_Query_Types_for_CVPR_2026_paper.html.

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 *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b14d7175755b180dc2163e15e3110cb6-Abstract-Conference.html.

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3318–3327, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_BOLT_Boost_Large_Vision-Language_Model_Without_Training_for_Long-form_Video_CVPR_2025_paper.html.

Yudong Liu, Jingwei Sun, Yueqian Lin, Jianyi Zhang, Jingyang Zhang, Ming Yin, Qinsi Wang, Hai Li, and Yiran Chen. Keyframe-oriented vision token pruning: Enhancing efficiency of large vision language models on long-form video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20802–20811, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liu_Keyframe-oriented_Vision_Token_Pruning_Enhancing_Efficiency_of_Large_Vision_Language_ICCV_2025_paper.html.

Yongdong Luo, Wang Chen, Xiawu Zheng, Weizhong Huang, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Jiebo Luo, and Rongrong Ji. QuoTA: Query-oriented token assignment via CoT query decouple for long video comprehension. arXiv preprint arXiv:2503.08689, 2025. URL https://arxiv.org/abs/2503.08689.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications*, pp. 382–390. IEEE, 2007. doi: 10.1109/PG.2007.17.

Yaxuan Qin, Hefei Li, Wenqi Mu, and Yancheng He. Efficient frame selection for long video understanding via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16944–16953, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Qin_Efficient_Frame_Selection_for_Long_Video_Understanding_via_Reinforcement_Learning_CVPR_2026_paper.html.

Qwen Team. Qwen3.5-9B model card. Official model card, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed September 26, 2026.

Shaden Shaar, Bradon Thymes, Sirawut Chaixanien, Claire Cardie, and Bharath Hariharan. MovieRecapsQA: A multimodal open-ended video question-answering benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Shaar_MovieRecapsQA_A_Multimodal_Open-Ended_Video_Question-Answering_Benchmark_CVPR_2026_paper.html.

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 the IEEE/CVF International Conference on Computer Vision*, pp. 24090–24101, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Sun_MDP3_A_Training-free_Approach_for_List-wise_Frame_Selection_in_Video-LLMs_ICCV_2025_paper.html.

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. doi: 10.1109/CVPR52734.2025.02711. URL https://arxiv.org/abs/2502.21271.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025. URL https://arxiv.org/abs/2508.18265.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Yuhui Zeng, Wang Chen, Jinfa Huang, Tianyu Xie, Yongdong Luo, Jiayi Ji, Xiawu Zheng, and Jiebo Luo. WaveZip: Wavelet-driven space-time decoupling for video token condensation. arXiv preprint arXiv:2607.23265, 2026. URL https://arxiv.org/abs/2607.23265.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025a. doi: 10.18653/v1/2025.findings-naacl.51. URL https://aclanthology.org/2025.findings-naacl.51/.

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*, pp. 22056–22065, 2025b. doi: 10.1109/ICCV51701.2025.02048. 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. arXiv preprint arXiv:2510.02778, 2025c. URL https://arxiv.org/abs/2510.02778.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 89863–89895, 2025d. doi: 10.52202/085713-3005. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8190b210e9808e54ee16263b673a847d-Abstract-Conference.html.

Mengyu Zhao, Di Fu, Yongyu Xie, Jiaxing Zhang, Zhigang Yuan, Shirin Jalali, and Yong Cao. Shot-aware frame sampling for video understanding. arXiv preprint arXiv:2603.17374, 2026. URL https://arxiv.org/abs/2603.17374.

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

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. ReaSon: Reinforced causal search with information bottleneck for video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13791–13799, 2026. doi: 10.1609/aaai.v40i16.38387. URL https://ojs.aaai.org/index.php/AAAI/article/view/38387.

Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. FOCUS: Efficient keyframe selection for long video understanding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/a010577f93a8301dada61f0b97cda923-Abstract-Conference.html.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.04428.

Appendix overview. Sections A and B provide additional related work and experimental details. Section C specifies the appearance descriptor, observability measures, and ranking algorithm. Section D discusses limitations, and Section E presents further ablations, model-scale and thinking-mode results, and a qualitative example.

A Additional Related Work

Video Large Multimodal Models. Video LMMs encode frames or clips as visual tokens for joint visual, temporal, and language reasoning. Recent systems improve this interface through dynamic visual resolution, spatiotemporal position modeling, grounded video supervision, tracking, and codec-aware tokenization (Bai et al., 2025; Wang et al., 2025; Clark et al., 2026; An et al., 2026). These advances do not eliminate the need for input selection: redundant observations can still occupy the input while brief or subtle events are missed. MEC selects original frames and treats the downstream LMM as a black box. The selector can therefore be used across architectures without access to internal token scores.

Matryoshka Representations. Matryoshka Representation Learning arranges an embedding so that nested dimensional prefixes retain progressively finer information; MatFormer extends this idea to nested Transformer subnetworks (Kusupati et al., 2022; Devvrit et al., 2024). For multimodal inputs, Matryoshka Multimodal Models learn nested visual-token sets, and the Matryoshka Query Transformer orders latent queries whose prefixes yield different token budgets (Cai et al., 2025; Hu et al., 2024). MEC applies the nesting principle to a query-conditioned ordering of original video-frame indices constructed at inference time. Early ranks concentrate answer-relevant evidence, while later ranks add temporal and visual context.

Visual-Token Compression. Visual-token compression reduces the input length through temporal aggregation, token merging or pruning, and adaptive budget allocation. Examples include hierarchical clip-to-video compression, joint similarity and importance scoring, diversity-based pruning, and query-oriented token assignment (Li et al., 2026b; Fu et al., 2025b; Alvar et al., 2025; Luo et al., 2025; Zeng et al., 2026). These methods operate within or around the visual-token pipeline and may use representations unavailable through a frame-only interface. MEC selects a nested sequence of original frames and leaves tokenization to the downstream model. It preserves the visual content of the selected frames until that step; token compression may discard fine details during input reduction.

B Additional Experimental Details

B.1 Benchmarks

Video-MME. Video-MME contains 900 videos and 2,700 manually annotated multiple-choice questions, with three questions per video. The videos cover six domains and 30 subfields and range from 11 seconds to one hour (Fu et al., 2025a). We report accuracy on the official short, medium, and long splits, together with Overall accuracy (Table 1). These splits allow us to compare selection performance across video durations.

Video-MME-v2. Video-MME-v2 contains 800 videos and 3,200 questions arranged in correlated groups. Its three-level hierarchy progresses from aggregating visual information across multiple points to modeling temporal dynamics and performing complex multimodal reasoning (Fu et al., 2026). We report micro accuracy over individual questions and the official nonlinear group score. The group score rewards consistent correctness across related questions and penalizes partially correct groups, complementing question-level accuracy.

LongVideoBench. LongVideoBench contains 3,763 videos and 6,678 human-authored multiple-choice questions in 17 categories. Videos cover movies, news, daily life, and knowledge-oriented content, with four duration ranges spanning roughly ten seconds to one hour (Wu et al., 2024). The benchmark pairs video with subtitle context and asks models to locate a referenced context before reasoning about its details. We report official question-level accuracy and use video inputs without subtitles, as specified in Table 5.

MLVU. MLVU contains 1,730 videos and 3,102 question–answer pairs across nine tasks. Videos range from three minutes to two hours, average approximately 15 minutes, and include movies, surveillance footage, egocentric recordings, cartoons, and games (Zhou et al., 2025). The full benchmark covers global understanding and localized reasoning in both multiple-choice and open-ended formats. Under our evaluation protocol, we report macro accuracy over the evaluated tasks, giving each task equal weight regardless of its number of questions.

B.2 Downstream Large Multimodal Models

Qwen3.5-9B. Qwen3.5-9B is a post-trained, 9-billion-parameter model with a unified image, video, and text interface. Its release describes early-fusion multimodal training, a vision encoder coupled to a causal language model, a hybrid Gated DeltaNet and attention backbone, and a native context length of 262,144 tokens (Qwen Team, 2026). Thinking is enabled by default and can be disabled through the official interface. We disable it for the four-benchmark comparison, component and schedule ablations, and efficiency analysis; Figure 4 (right) evaluates thinking separately. MEC supplies selected frames without accessing or modifying internal visual tokens.

LLaVA-OneVision-2-8B-Instruct. LLaVA-OneVision-2 uses a native OneVision encoder with windowed attention and shared 3D rotary position encoding for images, sampled frames, and codec-derived canvases. Its codec-stream pathway uses bit-cost dynamics to group temporal content and motion-residual cues to allocate spatial evidence within a visual-token budget (An et al., 2026). Uniform and MEC use decoded frames. Figure 3(a) also evaluates native codec-stream input, allowing us to compare query-conditioned frame selection with the model's compressed-video representation.

InternVL3.5-8B-Instruct. InternVL3.5 uses cascade reinforcement learning and a Visual Resolution Router to adjust visual-token resolution. Its release also describes decoupled vision and language deployment across devices (Wang et al., 2025). We use the 8B instruction-tuned checkpoint to test transfer across architectures. Figure 3(b) reports Video-MME and LongVideoBench results with the same MEC configuration used elsewhere, without model-specific selector tuning.

B.3 Comparison Methods

Adaptive Keyframe Sampling (AKS). AKS selects a fixed number of keyframes by balancing query relevance and temporal coverage (Tang et al., 2025). A lightweight vision–language scorer measures prompt–frame relevance, and recursively partitioned temporal bins measure coverage. The adaptive procedure ranges from selecting the highest-relevance frames to enforcing increasingly uniform coverage. Because selection depends on the target budget, AKS is rerun for each K. We use the released procedure and default settings at the same budgets as MEC.

Wavelet-based Frame Selection by Detecting Semantic Boundary (WFS-SB). WFS-SB applies multi-resolution wavelet analysis to a temporal query–frame relevance signal, suppressing high-frequency noise and identifying coarse semantic boundaries (Chen et al., 2026b). It allocates the target frame budget among the resulting clips according to their importance, then uses maximal marginal relevance to balance relevance and diversity within each clip. Both allocation and within-clip selection depend on the prescribed budget K. We use the released procedure and default settings without dataset- or budget-specific retuning.

B.4 Complete Experimental Configuration

We run all evaluations through LMMs-Eval (Zhang et al., 2025a). Except where an ablation specifies otherwise, MEC uses the configuration in Table 5 for every benchmark, downstream model, and frame budget. The candidate pool and length-M ranking are constructed independently of the target inference budget K.

For the reported multiple-choice evaluations, the original question and choices are followed by the same post-prompt for every selector and downstream model:

```
Answer with the option's letter from the given choices
directly.
```

Table 5: Complete shared configuration for the main MEC experiments.

Component	Shared setting
Hardware	8 NVIDIA H20 GPUs; Intel Xeon Platinum 8575C, 192 cores
Video decoding	Decord with `DECORD_THREADS=16`
Evaluation input	Decoded frames and question text; no subtitles; thinking disabled except for Figure 4 (right)
Ranking and budgets	$M = 256$; $K \in \{8, 16, 32, 64, 128, 256\}$
Sparse probes	$P_0 = P_1 = 256$; $\tau_p = 1800$ s
Temporal segments	$L = 4$ probe intervals
Appearance encoding	Short side $d = 32$, grayscale, ℓ_2 normalization
Evidence scorer	BLIP-2 ITM (Li et al., 2023)
Evidence mixture	$(\lambda_r, \lambda_\Delta, \lambda_o) = (0.7, 0.2, 0.1)$
Segment activation	$\eta = 0.25$
Local zoom	$L_z = 2$ interior frames on each available side
Ranking schedule	$(w_e^+, w_e^-, w_d, \gamma) = (0.6, 0.2, 0.2, 1)$
Sharpness	One-dimensional mean reblurring kernel, $L_b = 9$
Well-exposedness	$(\mu, \sigma) = (0.5, 0.2)$

This post-prompt specifies only the answer format. Dataset scoring, answer normalization, and other model-side evaluation procedures follow the corresponding LMMs-Eval task implementation. We use no subtitles and disable thinking except in Figure 4 (right). No additional model-, dataset-, or budget-specific decoding adjustments are made.

C Additional Method Details

C.1 Low-Resolution Appearance Encoding

We encode each probe as

$$\mathbf{f}_{p_i} = \mathrm{L2\text{-}Norm}(\mathrm{Vec}(\mathrm{Gray}(\mathrm{Resize}_d(x_{p_i})))) \,. \tag{12}$$

Here, Resize_d preserves the aspect ratio and sets the short side to $d = 32$ pixels; Gray and Vec denote grayscale conversion and vectorization. We cache the normalized descriptors so that appearance comparisons use compact vectors instead of original-resolution frames, reducing storage and computation.

C.2 Observability Measures

The sparse index estimates observability from query-independent sharpness and exposure measures. Let Ω_n be the pixel domain after resizing frame x_n to short side $d = 32$, and let $Y_n : \Omega_n \to [0, 1]$ be the grayscale intensity. For direction $\delta \in \{\mathrm{h}, \mathrm{v}\}$, let Ω_n^δ contain pixels with a valid forward neighbor and let ∇_δ be the corresponding finite difference. We reblur Y_n along direction δ with a one-dimensional mean kernel of length $L_b = 9$, producing B_n^δ. The original variation, reblurred variation, and nonnegative attenuation are

$$\Delta_{Y,n}^\delta(u) = |\nabla_\delta Y_n(u)|, \tag{13}$$

$$\Delta_{B,n}^\delta(u) = |\nabla_\delta B_n^\delta(u)|, \tag{14}$$

$$\Xi_n^\delta(u) = \max\big(0, \Delta_{Y,n}^\delta(u) - \Delta_{B,n}^\delta(u)\big) \,. \tag{15}$$

These pixelwise blur quantities are distinct from the frame-level visual-change cue Δ_n in Equation (5). Following the no-reference perceptual blur metric of Crete et al. (2007), the directional blur estimate and frame sharpness are

$$\beta_n^\delta = \frac{\sum_{u \in \Omega_n^\delta} \Delta_{Y,n}^\delta(u) - \sum_{u \in \Omega_n^\delta} \Xi_n^\delta(u)}{\sum_{u \in \Omega_n^\delta} \Delta_{Y,n}^\delta(u)}, \tag{16}$$

$$s_n^{\mathrm{sharp}} = 1 - \max_{\delta \in \{\mathrm{h}, \mathrm{v}\}} \beta_n^\delta. \tag{17}$$

If a direction has no measurable original variation, its denominator is zero and we set $\beta_n^\delta = 1$. Thus a flat image is not treated as sharp merely because reblurring leaves it unchanged.

For exposure, let $I_{n,c}(u) \in [0,1]$ denote color channel c at pixel u of the resized color thumbnail. Following the well-exposedness term of Mertens et al. (2007), we compute

$$W_n(u) = \prod_{c\in\{R,G,B\}} \exp\left(-\frac{(I_{n,c}(u)-\mu)^2}{2\sigma^2}\right), \tag{18}$$

$$\ell_n^{\text{light}} = \frac{1}{|\Omega_n|}\sum_{u\in\Omega_n} W_n(u), \tag{19}$$

with $\mu = 0.5$ and $\sigma = 0.2$. The channel product downweights pixels with any channel far from mid-range. We average over pixels and combine the resulting exposure score with sharpness:

$$O_n = s_n^{\text{sharp}}\ell_n^{\text{light}}. \tag{20}$$

Observability contributes only through $\lambda_o O_n$ in Equation (6); it does not impose a hard filtering threshold. This term downweights severely blurred or poorly exposed frames but does not measure query relevance or semantic informativeness.

C.3 Greedy Matryoshka Ranking Algorithm

Algorithm 1 implements the position-dependent surrogate from Section 3.4. Starting from the budget-independent candidate pool, it appends one frame at a time without replacing earlier selections. Each intermediate sequence is therefore an exact prefix of the final ranking.

Algorithm 1 Greedy construction of the Matryoshka ranking

Input: $\mathcal{C}(q)$ with $E_n(q)$ and $\mathbf{f}_n$; video length N
Parameters: $M, w_e^+, w_e^-, w_d, \gamma$
Output: Shared ranking $\mathcal{R}$

$r_1 \leftarrow \arg\max_{n\in\mathcal{C}(q)} E_n(q)$
$\mathcal{R} \leftarrow (r_1)$; $\mathcal{S} \leftarrow \{r_1\}$
for $k = 2, \ldots, M$ **do**
  Compute $w_e(k)$ using Equation (10)
  $w_c(k) \leftarrow 1 - w_e(k) - w_d$
  for each $n \in \mathcal{C}(q) \setminus \mathcal{S}$ **do**
    Compute $H_n(k)$ and $D_n(k)$ using Equations (8)–(9)
    $u_k(n) \leftarrow w_e(k)E_n(q) + w_c(k)H_n(k) + w_d D_n(k)$
  end for
  $r_k \leftarrow \arg\max_{n\in\mathcal{C}(q)\setminus\mathcal{S}} u_k(n)$
  Append r_k to $\mathcal{R}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{r_k\}$
end for
return $\mathcal{R}$

For a budget $K \le M$, we take the prefix $\mathcal{R}_{:K}$ and sort its frame indices by timestamp before passing the frames to the LMM. No additional selection run is needed.

D Limitations

Sparse-discovery blind spots. MEC scores a duration-adaptive sparse grid, activates the highest-scoring temporal segments, and samples around one anchor per active segment. This limits computation but can miss brief events between probes. Segment averaging may further dilute weak evidence at neighboring probes and leave a relevant segment inactive. Because ranking operates only on the discovered candidates, subsequent coverage and diversity terms cannot recover omitted frames. This is particularly problematic for very long videos, single-frame state changes, or events with little visual continuity, as illustrated in Figure 5b. Denser probes, shorter segments, higher activation ratios, or multiple anchors could improve recall at the cost of more decoding and query scoring. Adaptive boundary proposals or refinement triggered by uncertainty are possible alternatives.

Approximate evidence cues. The evidence cues approximate downstream utility. BLIP-2 ITM scores frames individually and may miss evidence that depends on ordered observations or long temporal relationships. The descriptor with a 32-pixel short side discards color and fine detail; cuts, camera motion, and lighting changes can dominate its distances even when small semantic changes matter more. Sharpness and exposure can also downweight an indispensable frame that is blurred or poorly lit. These errors affect segment activation and early ranks, and the fixed evidence mixture cannot adapt its cue weights to each question. Temporal scorers, higher-resolution descriptors, or learned calibration could help, but may require additional training, memory, or per-query computation.

Greedy surrogate rather than global optimization. Equation (3) defines an oracle objective over all prefixes. Prefix exposure motivates position-dependent weights, but evidence, coverage, and diversity remain proxies for unknown downstream marginal utility. Greedy construction guarantees nesting, not global optimality at any budget or for their weighted aggregate. An individually relevant frame can enter early and remain selected even if another combination would be more useful. The smooth exposure profile is a design prior, not a fit to deployment weights π_K; deployments concentrated on a few budgets may benefit from a different transition. The ablations provide empirical support for the chosen schedule. Beam search, local swaps, or learned marginal-utility estimates could improve selection at additional computational cost.

E Additional Experiments

E.1 Additional Ranking Ablations

Table 6: Additional ablations on Video-MME Overall with Qwen3.5-9B. Budget groups follow Table 2; Uniform and the default MEC configuration are repeated for reference. Arrows give the start and end evidence weights of linear schedules, with $w_d = .2$ and $w_c(k) = 1 - w_e(k) - w_d$. Best values are bold.

Variant	Low	Med.	High	Avg.
Uniform	60.80	67.81	73.09	67.23
MEC (default)	**65.19**	70.46	74.19	**69.94**
Evidence only	61.37	67.54	73.19	67.36
Linear $.2 \rightarrow .6$	63.19	69.52	74.28	69.00
Linear $.7 \rightarrow .3$	63.80	70.17	**74.48**	69.48
Linear $.5 \rightarrow .1$	65.00	**70.59**	73.67	69.75

Table 6 extends the comparisons in Table 2. Combining evidence, coverage, and diversity improves average accuracy by 2.58 pp over Evidence only. With candidate discovery fixed, reversing the default evidence-to-context transition reduces the average by 0.94 pp. Changing the linear endpoints from $.6 \rightarrow .2$ to $.7 \rightarrow .3$ or $.5 \rightarrow .1$ gives averages of 69.48 and 69.75. Together with the three decay rates in Table 2b, these five evidence-decay schedules fall within 0.46 pp of one another and outperform Uniform by at least 2.25 pp. The gains hold across the tested decay rates and endpoints.

E.2 Generalization Across Qwen3.5 Model Scales

Table 7 compares frame selection across downstream model scales. The six-budget mean under Uniform rises from 63.39 for 4B to 67.23 for 9B and 70.14 for 35B-A3B. The corresponding MEC means are 65.69, 69.94, and 72.67. Larger models achieve higher accuracy with either selector, while MEC retains an advantage at every scale.

MEC improves all 18 model–budget pairs. The mean gains are 2.31, 2.71, and 2.53 pp for 4B, 9B, and 35B-A3B, respectively, rising to 4.33, 4.89, and 3.74 pp at $K = 8$. Thus, selecting relevant evidence remains useful even for the larger model under a small frame budget. The 35B-A3B model also gains 3.23 and 2.70 pp at $K = 32$ and 64, respectively.

Table 7: Video-MME Overall accuracy across Qwen3.5 model scales and frame budgets. All models use the same configuration.

Size	Selector	8	16	32	64	128	256
4B	Uniform	56.19	60.04	62.93	65.30	67.15	68.70
4B	**MEC**	**60.52**	**62.74**	**65.89**	**67.04**	**68.59**	**69.37**
9B	Uniform	58.96	62.63	66.15	69.48	72.48	73.70
9B	**MEC**	**63.85**	**66.52**	**69.11**	**71.81**	**73.74**	**74.63**
35B-A3B	Uniform	61.85	65.44	69.33	72.67	74.93	76.59
35B-A3B	**MEC**	**65.59**	**69.11**	**72.56**	**75.37**	**76.07**	**77.30**

At $K = 256$, the gains narrow to 0.67, 0.93, and 0.71 pp as both selectors cover more of the video. The positive margins are consistent with the intended ranking behavior: later frames expand context while shorter-prefix evidence remains selected.

E.3 Thinking-Mode Results

Our standard evaluations disable Qwen3.5 thinking through the official interface (Qwen Team, 2026), as described in Section B.2. Figure 4 (right) enables thinking for both Uniform and MEC. The gains over Uniform are 5.56, 4.11, 2.78, 2.29, and 2.04 pp at $K = 8$, 16, 32, 64, and 128, respectively. They decrease as the budget grows but remain positive at every tested budget.

E.4 Additional Qualitative Example

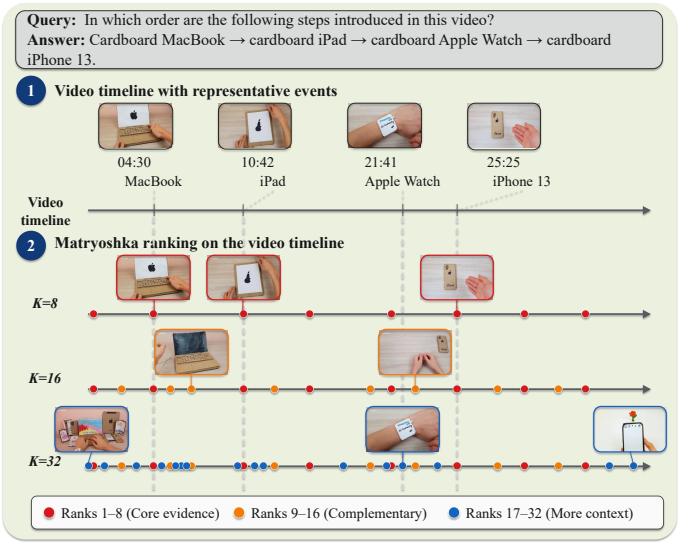


Figure 6: Ordering four cardboard-device introductions: MacBook (04:30), iPad (10:42), Apple Watch (21:41), and iPhone 13 (25:25). Red, orange, and blue mark frames entering at ranks 1–8, 9–16, and 17–32. Each prefix is shown in the chronological order used for LMM inference.

Figure 6 shows a query about four introductions spread across a video. The eight-frame prefix covers all four rather than concentrating on one construction sequence. In chronological order, these frames show the MacBook, iPad, Apple Watch, and iPhone introductions, and the LMM answers correctly. The $K = 16$ prefix retains these frames and adds MacBook and iPhone construction views. At $K = 32$, additions include an opening overview, another Apple Watch view, and a later iPhone demonstration. The nested prefixes extend coverage while preserving earlier selections, although this example does not establish that every added frame is needed for the answer.