SCORES HIGH ALONE, RANKS LOW IN CONTEXT: FULL-CONTEXT KNOCKOUT FRAME SELECTION WITH AN MLLM FOR LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video question answering asks a multimodal large language model (MLLM) to answer a question about a video that runs for tens of minutes. To make this long visual input manageable, keyframe selection reduces the video to a few dozen frames. A frame can be scored alone, by what it brings when other frames are absent, or in context, by what is lost when it is removed from the complete candidate pool. At deployment the selected frames enter the model together, motivating the second condition. We find that the highest-scoring frame alone can rank near the bottom in context because other frames already supply its evidence, and the two rankings are almost unrelated. We propose KEYSTONE, which removes one temporal unit, a short run of consecutive frames, and scores the displacement of a frozen answering model's answer distribution. The reference is determined jointly by all leave-one-out distributions. The whole chain runs on the frozen answering model, without training or an external scorer. On the Video-MME, MLVU, LongVideoBench, and LVBench benchmarks, with Qwen2.5-VL, LLaVA-OneVision, LLaVA-Video, and InternVL3.5, our KEYSTONE method outperforms existing training-free frame selectors. Code will be made publicly available upon acceptance.

1 INTRODUCTION

In long-video question answering, a multimodal large language model (MLLM) (Lin et al., 2024a; Cheng et al., 2024; Li et al., 2025b; Ye et al., 2025; Chen et al., 2024c; Wang et al., 2024b; Lin et al., 2024b; Zhang et al., 2025b; Chen et al., 2025) answers a question about a video that runs for tens of minutes. Among the ways to fit such a video into the model, this paper studies keyframe selection, which keeps a few dozen frames and fixes the visual evidence available to all subsequent reasoning (Tang et al., 2025; Liu et al., 2025; Zhang et al., 2025c). Training-free selectors, which leave the answering model untouched and require no additional annotation, are a common way to carry out this step.

Since keyframe selection determines the input the model receives, an important question is how a selector scores a single frame. Selectors score frames individually (Liu et al., 2025; Zhang et al., 2025c; Huang et al., 2025a; Kim et al., 2026), retrieve them in rounds (Wang et al., 2024a; 2025c; Zou et al., 2026), or add deduplication or coverage terms (Tang et al., 2025; Sun et al., 2025; Yang et al., 2026; Ma et al., 2026). Scores come from an external encoder, a trained selector, or the answering model's response to a frame. Regardless of their source or selection rule, they can be measured with that frame alone or with the other candidate frames present. This paper asks which condition a frame should be scored under.

A frame that scores highest alone can rank near the bottom when the other frames are present (Figure 1). Scored alone, its value is what it brings on its own; scored in context, its value is what it adds beyond the rest of the pool. The conditions keep different frames, change answers on a sizable share of questions, and yield higher accuracy when scored in context (Figure 1c). A frame can make the answer sharp alone yet leave it unchanged in context because other frames already supply its evidence. At deployment a frame sits among the other frames, so its score should be measured in context as well. Existing work incorporates other frames through a selection rule, a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

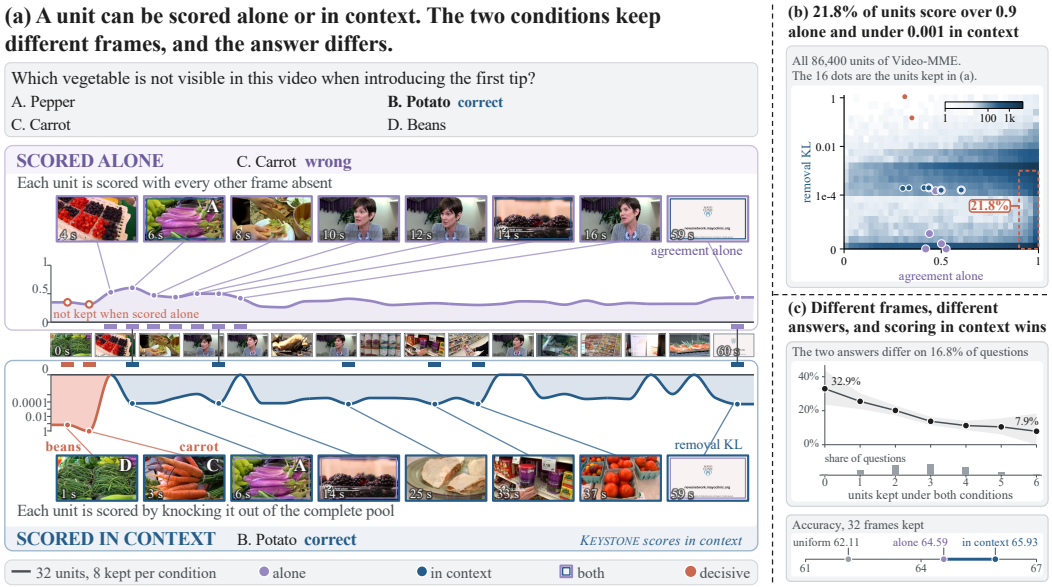

Figure 1: Scoring a unit alone versus removing it from the complete pool. (a) A Video-MME question with Qwen2.5-VL-7B: 128 frames form 32 four-frame units; each condition selects eight, shown with one frame per unit. (b) Both scores for 86,400 units across 2,700 questions; color is log-scaled count. (c) Answer-disagreement rate versus selection overlap between scoring alone and the deployed KEYSTONE, with a 95% bootstrap band, and accuracy below.

set-level training objective, or encoder-space distances (Sun et al., 2025; Tang et al., 2025; Yang et al., 2026; Ma et al., 2026). We propose KEYSTONE, which places this dependence where the score is measured, on the answering model's answer distribution.

Scoring in context raises a second question: which distribution should the displacement be measured from? The most immediate reference is the model's single prediction on the complete candidate pool, which is wrong on a substantial fraction of questions. On those questions, ranking frames by their support for that prediction returns evidence for the same wrong answer, repeating the error. The reference therefore cannot be any single prediction.

KEYSTONE answers both questions. It scores in context: it removes a temporal unit, a short run of consecutive frames, from the complete pool and measures the frozen answering model's posterior displacement. The score thus depends on what the remaining video already supplies. Its reference is the centroid of all leave-one-out distributions, so no single prediction anchors it. Attention masks let all removals share one visual-prefix prefill, and the selected frames return to the same frozen model. Scoring by removal has served as training supervision with ground-truth answers (Li et al., 2026a) and as a leave-one-out margin on an external encoder (Lin et al., 2025). KEYSTONE runs it at inference, without labels, on the answering model itself.

Our contributions are as follows.

- We show that scoring a frame alone and scoring it in context give two rankings that barely agree, and that the ranking in context is the one the answer depends on. We introduce KEYSTONE, which scores in context, by how much the answer distribution of the frozen answering model changes when a unit is removed.
- We show that the reference distribution decides whether the score repeats the error of the model itself, and we take the centroid of all leave-one-out distributions.
- Across four benchmarks and four answering models, KEYSTONE improves over uniform sampling by up to 7.1 points and over the strongest competing selector by up to 2.3 points.

2 RELATED WORK

Frame selection signals for long-video question answering. Existing work either compresses the tokens emitted by sampled frames (Chen et al., 2024b; Shang et al., 2025; Bolya et al., 2023;

Huang et al., 2025b; Zhang et al., 2025d; Jin et al., 2024; Qi et al., 2026; Zhang et al., 2025g; Li et al., 2025c) or selects a frame subset within a budget, as studied here. Frame-selection signals fall into four groups. The first uses question-frame matching scores from an external encoder (Radford et al., 2021; Li et al., 2022; Zhai et al., 2023; Tang et al., 2025; Liu et al., 2025; Zhang et al., 2025c; Zhu et al., 2026; Sun et al., 2025; Park et al., 2026; Chen et al., 2026). The second trains a dedicated importance predictor (Buch et al., 2022; Yu et al., 2025; Yao et al., 2025; Yu et al., 2023; Tang et al., 2026; Yang et al., 2026; Gao et al., 2023; Xu et al., 2026; Hu et al., 2025). The third lets a language or vision-language model decide which segment to inspect over multiple rounds (Wang et al., 2024a; 2025c; Zou et al., 2026; Ma et al., 2025; Liao et al., 2024; Yang et al., 2025; Zhang et al., 2025e). The fourth reads outputs or cross-modal attention from the answering multimodal model or another paired with it (Li et al., 2026b; Huang et al., 2025a; Eltahir et al., 2026; Kim et al., 2026; Wang et al., 2026). The proposed KEYSTONE measures each removal in the frozen answering model's posterior while retaining all other candidate units.

Set-level objectives and selection rules. How much a frame contributes changes with which other frames are present. Existing methods model this dependence through rules over per-frame scores, such as temporal coverage or semantic-boundary segmentation (Tang et al., 2025; Chen et al., 2026); direct subset coverage, using determinantal point processes, maximum spanned volume, or cluster representatives (Sun et al., 2025; Li et al., 2026c; Wang et al., 2025c); and set-level training objectives (Yang et al., 2026) or reinforcement-learning rewards (Lee et al., 2026). Another approach measures SigLIP-space distance from candidates with higher question relevance (Ma et al., 2026; Zhai et al., 2023). These methods encode dependence in selection rules, training objectives, or encoder geometry. Our KEYSTONE measures it in the answer posterior; a unit tends to rank low when the rest of the pool already supplies its evidence.

Attribution by removal. Scoring a unit by its absence from an otherwise complete input is a removal construction. In frame selection, removing or duplicating a frame and reading the change in a model's score for the ground-truth answer has served as training supervision (Li et al., 2026a; Sun et al., 2026), and a divergence between the answer distributions on a selected and a counterfactual frame subset has formed one term of a reinforcement-learning reward (Zhou et al., 2026). A leave-one-out margin on the video-text similarity of an external encoder has been computed at inference (Lin et al., 2025). Removal also appears in post-hoc attribution, masking an image region (Zeiler & Fergus, 2014), fitting a local surrogate on perturbed inputs (Ribeiro et al., 2016; Lundberg & Lee, 2017), or ablating context spans (Cohen-Wang et al., 2024), and in evaluating explanations by erasing the marked rationale (DeYoung et al., 2020). KEYSTONE needs no ground-truth answer, and its removal runs at inference on the frozen answering model, deciding which frames that model then answers from.

3 METHOD

KEYSTONE formulates frame selection as attribution by removal over temporal units, where a unit's value is the change in the frozen answering model's answer posterior when removed from the complete candidate pool (Figure 2). Section 3.1 defines units, posteriors, and the frame budget. Section 3.2 defines full-context contribution. Section 3.3 gives the centroid reference and KL score. Section 3.4 covers attention masking, cache reuse, and budgeted selection.

3.1 PRELIMINARIES

Evidence for long-video questions can span multiple moments. We define value over independently removable temporal units. The candidate pool $\mathcal{P} = \{p_i\}_{i=1}^{n}$ is a dense uniform sample of the video, and "full context" refers to this pool. Partition it into m temporal units $\mathcal{U}_1, \ldots, \mathcal{U}_m$ along independently maskable boundaries. They are disjoint and jointly cover the pool, $\mathcal{P} = \dot{\bigcup}_{j=1}^{m} \mathcal{U}_j$. The count m fixes the number of removals, trading attribution resolution against cost. Write $[m] = \{1, \ldots, m\}$ and, for $\mathcal{I} \subseteq [m]$, $\mathcal{P}_{\mathcal{I}} = \bigcup_{j \in \mathcal{I}} \mathcal{U}_j$. Attribution operates on unit indices while the model consumes the corresponding frames.

Let q be the question, $\mathcal{A}$ the finite answer set, and $\mathcal{M}$ the frozen answering model. Its answer-candidate logits induce the answer distribution, or posterior,

$$\boldsymbol{\pi}_{\mathcal{I}} = \mathrm{softmax}(\mathcal{M}(\mathcal{P}_{\mathcal{I}}, q)) \in \Delta^{|\mathcal{A}|-1}. \tag{1}$$

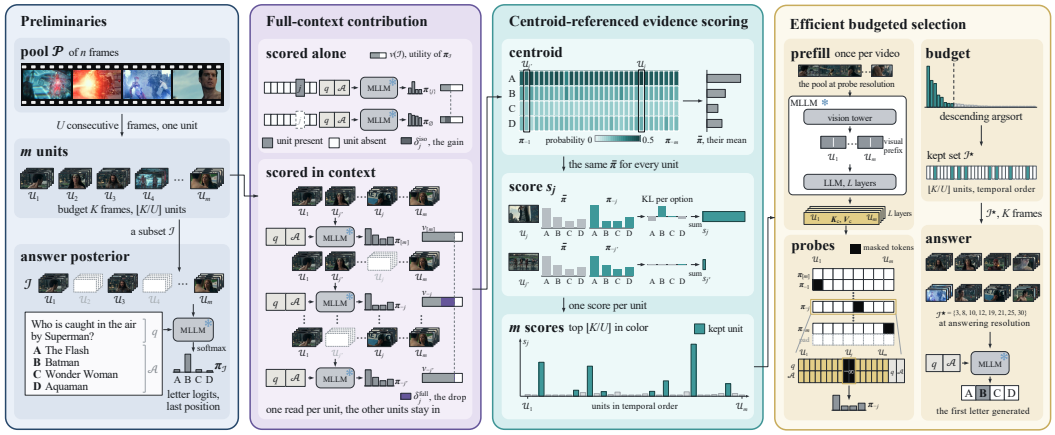

Figure 2: Overview of KEYSTONE. From left to right: partition the candidate pool into temporal units and define subset posteriors; contrast the score of a unit alone with its score in context; score leave-one-out posteriors against their shared centroid; and reuse one prefill for m masked suffix probes. The highest-scoring units within the budget return in temporal order to the same frozen answering model.

Given budget K, the selector returns $\mathcal{I}^\star \subseteq [m]$ with $|\mathcal{P}_{\mathcal{I}^\star}| \leq K$ and passes those frames to $\mathcal{M}$ for the answer. Unit scores must depend on context and use this model's posteriors alone, without ground-truth labels or an external scorer.

3.2 FULL-CONTEXT CONTRIBUTION

Isolated scoring measures the value of each frame or unit alone, with the other units absent, so the score depends only on that unit, the question, and the model, whatever selection rule is applied to the scores afterwards. Copies of evidence already present then score high together, and units that support the answer only together with material elsewhere score low and are dropped. KEYSTONE keeps the score-and-rank form and changes the context in which value is measured. Let $v(\mathcal{I})$ be a scalar set utility that summarizes the answer posterior induced by $\mathcal{P}_{\mathcal{I}}$. Its form is left open in this section, and Section 3.3 instantiates it as the negative divergence from that posterior to a shared reference.

The marginal contribution of a unit $j \notin \mathcal{I}$ relative to a retained context $\mathcal{I}$ is

$$\delta_j(\mathcal{I}) = v(\mathcal{I} \cup \{j\}) - v(\mathcal{I}). \tag{2}$$

Isolated scoring fixes the retained context at the empty set. The model reads the unit on its own, and the value of the unit is what it brings when nothing else is present,

$$\delta_j^{\text{iso}} = v(\{j\}) - v(\varnothing). \tag{3}$$

KEYSTONE fixes the retained context at every other unit. The model reads the complete pool once, then reads the same pool with $\mathcal{U}_j$ removed and every other unit retained, which we call the leave-one-out input of unit j (Figure 2), and the value of the unit is what the rest of the video cannot replace,

$$\delta_j^{\text{full}} = v([m]) - v([m] \setminus \{j\}). \tag{4}$$

The two conditions agree only when the utility is additive, $v(\mathcal{I}) = v(\varnothing) + \sum_{i \in \mathcal{I}} w_i$, so that removing unit j changes the utility by the same w_j whatever else is retained. Once the answer rests on evidence spread over several moments, the value of a unit depends on which other units are present, and only scoring in context keeps that dependence.

3.3 CENTROID-REFERENCED EVIDENCE SCORING

Scoring in context still needs a common reference to compare the posterior changes of different units. The reference decides which answer coordinates the divergence weights, so unit-specific references give incomparable distances. The single prediction on the complete pool puts that weight on its argmax answer, which is wrong on a substantial fraction of questions (Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Chen et al., 2024a), so the supporters of that answer would rank first. KEYSTONE instead derives a shared reference from the posteriors obtained when each

unit in turn is absent. For each $j \in [m]$, removing $\mathcal{U}_j$ while retaining all other units yields the leave-one-out posterior

$$\boldsymbol{\pi}_{-j} \coloneqq \boldsymbol{\pi}_{[m]\setminus\{j\}}. \tag{5}$$

These posteriors correspond to the same question and candidate pool and differ only in the removed unit (Figure 2), and their arithmetic centroid

$$\bar{\boldsymbol{\pi}} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\pi}_{-j} \tag{6}$$

is a permutation-invariant posterior reference determined jointly by all leave-one-out posteriors, which, like consensus over sampled outputs (Wang et al., 2023; Manakul et al., 2023), privileges neither a particular unit nor a single prediction as the coordinate center. It is computed once per question and held fixed across the scoring of every unit, so the utility defined below is a well-defined set function in which only $\boldsymbol{\pi}_{\mathcal{I}}$ varies with $\mathcal{I}$.

Using $\bar{\boldsymbol{\pi}}$ as the reference, the set utility and the score of unit j are

$$v_{\bar{\boldsymbol{\pi}}}(\mathcal{I}) = -D_{\mathrm{KL}}(\bar{\boldsymbol{\pi}} \,\|\, \boldsymbol{\pi}_{\mathcal{I}}), \tag{7}$$

$$s_j = D_{\mathrm{KL}}\big(\bar{\boldsymbol{\pi}} \,\big\|\, \boldsymbol{\pi}_{-j}\big), \tag{8}$$

where D_{KL} denotes the KL divergence. The utility is larger when the posterior induced by a subset lies closer to the shared centroid, and a larger s_j indicates that removing unit j moves the answer posterior farther from the consensus of all leave-one-out posteriors. Under this utility the full-context marginal is

$$\delta_j^{\mathrm{full}} = s_j - D_{\mathrm{KL}}\big(\bar{\boldsymbol{\pi}} \,\big\|\, \boldsymbol{\pi}_{[m]}\big), \tag{9}$$

whose second term is independent of j, so ranking by s_j yields the same order as ranking by the full-context marginal.

Because the KL divergence is nonnegative, s_j is a magnitude. Resolving the direction of the displacement would require knowing which answer is correct, which label-free inference does not provide, so KEYSTONE uses the magnitude alone. It selects the units the model's answer state actually depends on.

3.4 EFFICIENT BUDGETED SELECTION

Direct full-context scoring would rerun the video forward pass for every unit, making latency grow with both the unit count and visual input length. Split the input into a visual prefix containing all candidate frames and a text suffix containing the question and options. The prefix is shared across leave-one-out inputs, so KEYSTONE encodes it once and reruns only the suffix under each mask, reusing cached computation (Zheng et al., 2024). Let $\mathcal{R}_j$ denote the prefix token positions of $\mathcal{U}_j$, and $\boldsymbol{K}_{\mathrm{c}}, \boldsymbol{V}_{\mathrm{c}}$ the cached prefix keys and values. The mask for unit j excludes those positions from suffix attention:

$$\boldsymbol{H}^{(j)} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{sfx}}\boldsymbol{K}_{\mathrm{c}}^{\top}}{\sqrt{d_{\mathrm{h}}}} + \boldsymbol{B}^{(j)}\right)\boldsymbol{V}_{\mathrm{c}}, \qquad [\boldsymbol{B}^{(j)}]_{ab} = \begin{cases} -\infty, & b \in \mathcal{R}_j, \\ 0, & b \notin \mathcal{R}_j, \end{cases} \tag{10}$$

where $\boldsymbol{Q}_{\mathrm{sfx}}$ is the suffix query matrix, d_{h} is the attention-head dimension, and a and b index a suffix query and a cached key position. The mask hides only the visual tokens of $\mathcal{U}_j$ from the text suffix, so the question is answered as if that unit were absent, while the sequence length, the positional indices, and every other visual token stay as they are. Applying one such mask and recomputing the text suffix yields one leave-one-out posterior, and we call that single forward pass a probe (Figure 2). The cached keys and values of the other units were computed with $\mathcal{U}_j$ present, so a probe is an approximation of the posterior on a pool re-encoded without $\mathcal{U}_j$.

Under frame budget K, the final index set maximizes the total score,

$$\mathcal{I}^{\star} = \operatorname*{argmax}_{\mathcal{I}\subseteq[m],\, |\mathcal{P}_{\mathcal{I}}|\le K} \sum_{j\in\mathcal{I}} s_j, \tag{11}$$

where $\mathcal{P}_{\mathcal{I}^\star}$ is the corresponding selected frame set. For equal-width units of U frames, the exact solution takes the $\lfloor K/U \rfloor$ highest-scoring units. The selected frames return in temporal order to the same frozen model for the same question at answering resolution.

Let c_{shared} denote the cost of shared visual encoding and prefix prefill, and c_{probe} the cost of one masked suffix forward pass. Computing all unit scores requires

$$c_{\text{score}}(m) = c_{\text{shared}} + m\, c_{\text{probe}} + O(m|\mathcal{A}|)\,, \tag{12}$$

where the last term accounts for the centroid of Eq. (6) and the m KL scores of Eq. (8).

4 Experiments

4.1 Setup

We evaluate the proposed method and baselines on four benchmarks, including Video-MME (Fu et al., 2025), MLVU-Test (Zhou et al., 2025), LongVideoBench-Val (Wu et al., 2024), and LVBench (Wang et al., 2025a). Video-MME is reported with and without subtitles, split by official duration buckets; later tables omit subtitles unless noted. MLVU-Test uses the official macro average over nine question types. The four frozen answering models are Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2025a), LLaVA-Video-7B (Zhang et al., 2025f), and InternVL3.5-8B (Wang et al., 2025b), evaluated with lmms-eval (Zhang et al., 2025a). We compare with uniform sampling and eight training-free selectors, grouped by the model that scores the frames. Q-Frame (Zhang et al., 2025c), BOLT (Liu et al., 2025), AKS (Tang et al., 2025), mDP^3 (Sun et al., 2025), and WFS-SB (Chen et al., 2026) score with an image-text encoder, A.I.R. (Zou et al., 2026) and VideoTree (Wang et al., 2025c) with an MLLM rater, and the UG Framework (Kim et al., 2026) with the answering model. Every selector uses the same 128-frame candidate pool sampled uniformly on a fixed grid and a 32-frame answering budget. We call the model's answer on the complete candidate pool its dense answer. Keystone uses four-frame units. A.I.R. and VideoTree return variable frame counts, marked as at most 32 in Table 1.

4.2 Main Results

Table 1 reports the competitor matrix over four frozen answering models and four benchmarks. Keystone leads in 18 of 20 cells, ties mDP^3 on Qwen2.5-VL LVBench, and trails WFS-SB by 0.15 points on LLaVA-OneVision LongVideoBench. Averaged over the 20 cells, mDP^3 leads the competitors, and Keystone is 1.34 points above it. Against the best competitor of each cell, the margin averages 0.85 points and reaches 2.29 points on MLVU with InternVL3.5. Averaged over the five cells of each answering model, the most accurate competitor is a different method on each of the four models. Keystone ranks first on all four with one scoring rule, which reads how much the answer distribution of whichever model answers changes when a unit leaves the complete pool.

The gain over uniform sampling is smallest on short videos. On Video-MME without subtitles, averaged over the four models, Keystone adds 1.50 points on short videos, 3.39 on medium videos, and 2.33 on long videos, and subtitles keep this order. A short video lasts under two minutes, so 32 uniformly spaced frames already sample it every few seconds and leave selection little to change. Against the best competitor of each duration bucket without subtitles, Keystone leads on all four models for short and long videos and on three of the four for medium videos. Table 2 scales Qwen2.5-VL from 7B to 72B on MLVU. The margin over uniform narrows from 6.89 to 6.22 points, so the gain persists at larger model scale.

4.3 Ablations

Table 3 removes one component of Keystone per row. *w/o in-context scoring* scores each unit alone, where the probe retains a single unit, hides the remaining 31 with the attention mask, and scores that unit by $-D_{\text{KL}}(\boldsymbol{\pi}_{[m]} \,\|\, \boldsymbol{\pi}_{\{j\}})$. Every cell falls below the deployed arm. *w/o centroid reference* replaces the centroid of all leave-one-out posteriors by the answer distribution on the complete candidate pool. *w/o KL readout* scores by the change in the maximum answer probability and lowers all four cells. *w/o attention masking* removes a unit by black-frame replacement and stays within one point of the deployed arm in all four cells. Attention masking is kept because it needs one prefill per video instead of two. Table 4 answers how the unit width U is set. Width 4 is the widest setting that stays within one point of width 2 on every benchmark, and we use it throughout.

4.4 Why Scoring in Context Helps

Table 1: Main results on long-video question answering, accuracy (%). Within each block the best result is **bold** and the runner-up underlined. Later tables abbreviate Video-MME to VMME and LongVideoBench to LVB. † marks numbers we reproduced under this shared protocol.

Model	Size	#Frames	Video-MME (w/o / w/ sub.)				MLVU	LongVideoBench	LVBench
			Short	Medium	Long	Overall			
Avg. Duration			*1 min*	*9 min*	*41 min*	*17 min*	*18 min*	*8 min*	*67 min*
Qwen2.5-VL†	7B	32	73.89/76.56	60.11/64.78	52.33/54.33	62.11/65.22	45.98	56.84	39.06
+Q-Frame†	7B	32	72.33/75.56	61.44/63.11	52.22/54.11	62.00/64.26	50.92	60.81	41.12
+BOLT†	7B	32	75.00/76.67	61.22/63.00	52.11/54.44	62.78/64.70	50.97	59.09	42.48
+AKS†	7B	32	74.33/77.11	63.11/64.78	54.67/54.78	64.04/65.56	51.19	60.73	42.35
+MDP^3†	7B	32	76.44/78.22	60.89/63.89	53.67/55.22	63.67/65.78	51.57	60.66	**42.80**
+A.I.R.†	7B	≤32	71.67/75.22	60.56/62.78	52.89/54.33	61.70/64.11	50.16	57.52	37.12
+VideoTree†	7B	≤32	73.56/73.78	61.22/60.11	53.11/55.33	62.63/63.07	49.13	56.92	39.51
+WFS-SB†	7B	32	75.67/78.78	62.67/64.56	52.78/57.11	63.70/66.81	49.99	61.03	41.64
+UG Framework†	7B	32	74.56/76.89	61.78/65.22	52.56/55.11	62.96/65.74	47.65	61.48	41.58
+KEYSTONE	7B	32	77.33/78.78	64.56/67.22	55.89/57.33	**65.93/67.78**	**52.87**	**61.93**	**42.80**
LLaVA-OneVision†	7B	32	71.56/75.22	56.56/58.89	47.67/51.22	58.59/61.78	47.13	57.29	39.12
+Q-Frame†	7B	32	67.44/73.67	58.11/62.22	48.56/52.44	58.04/62.78	47.52	58.19	43.00
+BOLT†	7B	32	70.44/75.22	57.89/60.00	49.89/52.11	59.41/62.44	48.97	58.71	42.93
+AKS†	7B	32	70.67/74.78	57.89/61.22	49.78/53.44	59.44/63.15	48.63	57.97	43.19
+MDP^3†	7B	32	70.44/75.44	58.33/60.89	50.22/51.89	59.67/62.74	49.48	58.49	43.19
+A.I.R.†	7B	≤32	69.22/74.33	56.44/57.44	48.11/52.44	57.93/61.41	46.01	56.69	41.58
+VideoTree†	7B	≤32	70.11/73.56	56.89/57.67	49.00/51.67	58.67/60.96	45.14	55.80	40.67
+WFS-SB†	7B	32	71.78/75.11	59.89/61.44	49.44/51.44	60.37/62.67	50.16	**59.54**	42.09
+KEYSTONE	7B	32	71.89/75.22	60.11/62.00	50.89/52.33	**60.96/63.19**	**50.69**	59.39	**43.96**
LLaVA-Video†	7B	32	75.67/77.89	60.67/62.11	55.22/55.78	63.85/65.26	45.64	58.12	40.41
+Q-Frame†	7B	32	72.22/76.67	59.33/63.78	53.56/55.67	61.70/65.37	46.63	58.34	44.09
+BOLT†	7B	32	74.11/78.33	61.67/64.56	54.56/56.89	63.44/66.59	47.99	59.24	45.19
+AKS†	7B	32	75.89/78.44	62.00/65.67	53.78/56.78	63.89/66.96	45.93	58.94	43.83
+MDP^3†	7B	32	75.89/79.00	61.22/63.22	54.56/56.67	63.89/66.30	48.43	58.86	44.74
+A.I.R.†	7B	≤32	72.89/77.11	58.33/61.89	53.22/56.33	61.48/65.11	43.74	58.12	38.80
+VideoTree†	7B	≤32	74.67/75.44	59.67/59.56	51.89/55.22	62.07/63.41	45.44	57.44	41.77
+WFS-SB†	7B	32	77.00/79.67	62.11/64.89	53.33/55.78	64.15/66.78	45.03	59.39	44.42
+KEYSTONE	7B	32	77.78/79.89	64.78/67.89	55.11/56.56	**65.89/68.11**	**50.59**	**59.99**	**45.32**
InternVL3.5†	8B	32	77.56/80.22	63.89/66.33	53.78/54.11	65.07/66.89	48.51	60.36	42.48
+Q-Frame†	8B	32	72.56/75.56	61.89/65.22	54.56/56.44	63.00/65.74	51.72	60.73	46.93
+BOLT†	8B	32	76.89/79.89	64.56/66.00	53.89/55.56	65.11/67.15	53.30	60.06	47.06
+AKS†	8B	32	77.11/79.11	64.89/67.44	54.44/55.56	65.48/67.37	51.45	62.75	46.80
+MDP^3†	8B	32	77.11/80.11	63.33/65.89	55.44/57.00	65.30/67.67	51.96	62.15	47.84
+A.I.R.†	8B	≤32	74.00/76.89	60.22/63.56	54.22/54.89	62.81/65.11	49.21	60.06	44.54
+VideoTree†	8B	≤32	75.56/76.22	61.89/63.44	54.67/53.78	64.04/64.48	47.89	59.61	44.74
+WFS-SB†	8B	32	76.56/79.67	63.89/66.00	54.67/55.00	65.04/66.89	51.86	62.00	47.90
+UG Framework†	8B	32	75.33/78.44	65.78/68.22	55.67/56.33	65.59/67.67	53.24	63.28	45.84
+KEYSTONE	8B	32	77.67/79.78	65.33/67.78	56.44/57.56	**66.48/68.37**	**55.59**	**64.25**	**47.97**

Table 2: Scaling Qwen2.5-VL from 7B to 72B on MLVU, accuracy (%).

Method	7B	72B
Uniform	45.98	51.81
MDP^3	51.57	57.30
KEYSTONE	**52.87**	**58.03**

Figure 3 examines the rank the same unit receives when scored alone and when scored in context. In Figure 3a the color does not concentrate on the diagonal. For each question the 32 units are ranked once under each score, and the median rank correlation between the two rankings over all questions is only 0.03 to 0.07. In Figure 3b the unit ranked first alone remains first in context in roughly 20% of the questions. When each score selects eight units, only 2.5 to 2.7 of them coincide on average. Figure 3c answers with the frames each score selects, and scoring in context is more than two points higher on both benchmarks.

A selector could also gain by keeping frames that are informative for any question about the video. We therefore score each question's frames on Qwen2.5-VL with the next question of the same video in cyclic order, and then answer the original question (Appendix D.3). Accuracy falls by 5.8 points on Video-MME, 2.2 points on LongVideoBench, and 2.8 points on MLVU, to the level of null controls built from evenly spaced units, so the gains depend on the question.

With the answer distribution on the complete pool as the reference in place of the centroid, accuracy changes by about one point at most (Table 3 and Appendix D.3). Figure 4 tests what separates the two references, namely whether the score indirectly reads the gold answer and whether gains persist when the dense answer is wrong. On questions where the dense answer is wrong, the score's rank correlation with the drop in gold-answer probability when a unit is removed stays within the

Table 3: Component ablation, accuracy (%). Each row removes one component of KEYSTONE. The VMME columns are with subtitles.

Method	LLaVA-Video		InternVL3.5	
	VMME	LVB	VMME	LVB
Uniform	65.26	58.12	66.89	60.36
KEYSTONE (full)	**68.11**	**59.99**	68.37	**64.25**
w/o in-context scoring	65.33	59.54	65.41	61.18
w/o centroid reference	67.30	59.54	67.37	63.72
w/o KL readout	67.15	59.76	66.96	62.45
w/o attention masking	67.70	59.46	**68.44**	63.43

Table 4: Unit width sweep, Qwen2.5-VL-7B accuracy (%). The width is in frames, the number of probes is the candidate pool size divided by the width, and selection time is the whole-process time per question. Distinct sets counts the 32-frame selections a width can express.

Width	Probes	Sel. s	Distinct sets	VMME	LVB	MLVU
2	64	2.35	4.9×10^{14}	65.78	61.03	**53.41**
4	32	2.07	1.1×10^{7}	**65.93**	61.93	52.87
8	16	2.02	1,820	64.70	61.63	53.28
16	8	1.82	28	62.96	**62.08**	49.12
32	4	1.79	4	60.70	59.09	45.64

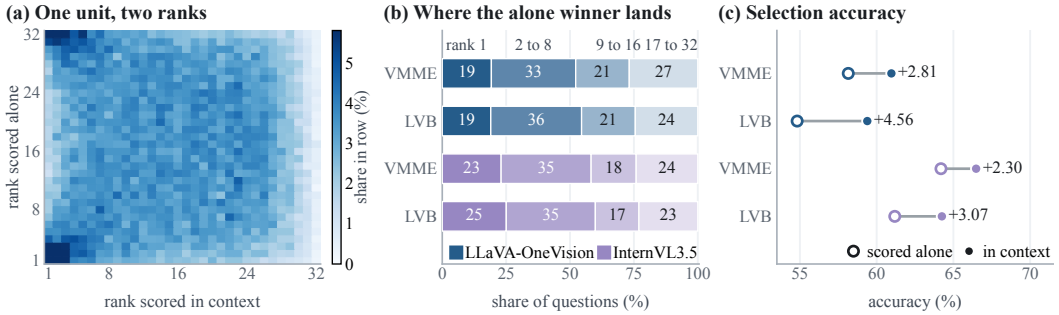

Figure 3: Scoring the same units alone versus in context. (a) Unit ranks under both scores; color shows row-normalized shares, and rank 1 is highest. (b) In-context rank of the unit ranked first alone. (c) Accuracy of scoring alone and of KEYSTONE, with gains at right. Panel a uses LLaVA-OneVision on VMME; b and c use LLaVA-OneVision and InternVL3.5 on VMME and LVB.

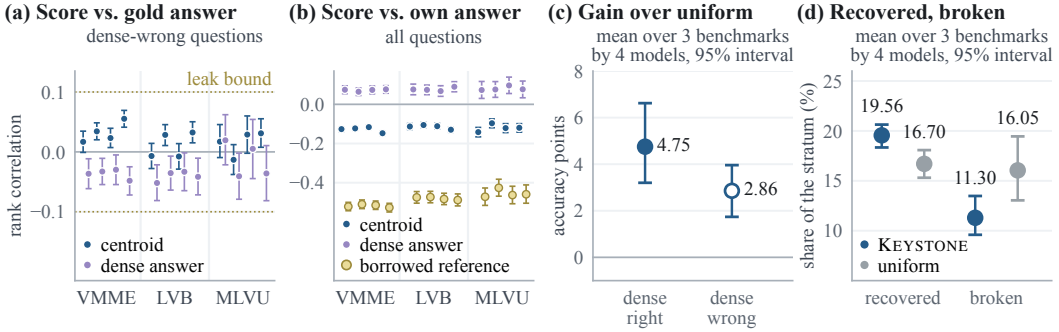

Figure 4: Reference diagnostics and answering outcomes. (a) Within-question rank correlation of the score with the drop in gold-answer probability. (b) Correlation with the drop in the model's own answer probability under three references, including a reference borrowed from another question as a null control. (c) Gain over uniform, split by whether the dense answer is correct. (d) Fractions of questions recovered or broken by selection. In a and b, models run left to right within each benchmark: Qwen2.5-VL, LLaVA-OneVision, LLaVA-Video, InternVL3.5. Panels c and d average twelve evaluation cells; whiskers show 95% intervals.

leakage bound of 0.10 set before the runs, in all twelve cells under both references (panel a). With the dense answer as reference, higher scores track larger drops in that answer, and with the centroid the correlation turns negative (panel b). The centroid privileges no single prediction and its score does not follow the model's own answer, motivating its deployment. Even when the dense answer is wrong, KEYSTONE gains 2.86 points over uniform averaged over the twelve cells (panel c), and it recovers more dense-wrong questions while breaking fewer dense-right ones (panel d).

4.5 MORE FRAMES VERSUS SELECTION

KEYSTONE reads the complete pool before it selects, so one could simply keep the dense answer. Table 5 compares the two, and the 32 selected frames answer better in 11 of 12 cells. On LLaVA-OneVision the dense answer falls below 32 uniformly sampled frames on Video-MME and LongVideoBench (Table 1), so reading more frames does not substitute for choosing them. Table 6

Table 5: The dense answer versus the answer from 32 selected frames, accuracy (%). The dense rows take the dense answer from the pass over the complete pool that KEYSTONE runs before it selects. On Qwen2.5-VL that pass uses a quarter of the per-frame pixel budget.

Model	Method	#F	VMME	LVB	LVBench
Qwen2.5-VL	Dense	128	64.37	59.91	40.22
	KEYSTONE	32	**65.93**	**61.93**	**42.80**
LLaVA-OneVision	Dense	128	58.48	56.47	40.87
	KEYSTONE	32	**60.96**	**59.39**	**43.96**
LLaVA-Video	Dense	128	64.07	**60.81**	43.58
	KEYSTONE	32	**65.89**	59.99	**45.32**
InternVL3.5	Dense	128	65.26	60.58	46.55
	KEYSTONE	32	**66.48**	**64.25**	**47.97**

Table 6: Frame budget sweep, Qwen2.5-VL-7B accuracy (%). Each budget is compared with uniform sampling of the same budget drawn evenly from the same 128-frame candidate pool. The 32-frame column is the budget used in every other table.

Benchmark	Method	#Frames			
		8	16	32	64
VMME	Uniform	55.52	59.33	62.11	63.74
	KEYSTONE	**61.89**	**64.07**	**65.93**	**66.26**
LVB	Uniform	55.20	57.14	56.84	60.43
	KEYSTONE	**59.39**	**60.81**	**61.93**	**61.41**
MLVU	Uniform	42.62	41.65	45.98	49.87
	KEYSTONE	**52.27**	**52.09**	**52.87**	**53.68**

Table 7: Seconds per question. Video-MME with Qwen2.5-VL-7B, 30 questions from 10 videos, mean of 3 repetitions. Selection time is the wall clock of the selection process divided by the number of questions.

Scorer type	Method	Sel. s	Total s
None	Uniform	0.00	0.80
	Full-resolution dense	0.00	1.63
Image-text encoder	MDP^3	1.57	2.37
	AKS	2.46	3.26
MLLM rating	A.I.R.	8.16	8.96
	VideoTree	9.64	10.44
Answering model	KEYSTONE	2.07	2.87
	UG Framework	15.16	15.96

Table 8: Dense answering and total cost. Video-MME accuracy (%) with Qwen2.5-VL-7B. The full-resolution dense row feeds all 128 candidate frames to the answering model. Total seconds are the whole-process selection time plus the answering time. Accuracy is measured on the full benchmark and the seconds on the 30-question timing sample of Table 7.

Method	#F	Acc.	Sel. s	Total s
Full-resolution dense	128	65.96	0.00	1.63
Uniform	32	62.11	0.00	0.80
KEYSTONE	32	65.93	2.07	2.87
Full-res-probe KEYSTONE	32	66.11	3.20	4.00
Coalition-reg. KEYSTONE	32	65.59	2.53	3.33

examines how the gain varies with the frame budget K. On Video-MME the margin of KEYSTONE over uniform sampling narrows steadily as the budget grows. On LongVideoBench it is largest at a budget of 32 frames and narrows at 64.

4.6 SELECTION COST

Tables 7 and 8 report selection and total time, amortizing each selector's shared computation over questions of the same video. In this same-environment timing run, selection plus answering costs more than dense 128-frame answering at matching accuracy. Full-resolution probes raise the per-frame pixel cap to the budget assigned to the complete candidate pool and cost more. The coalition-regression variant of Table 8 scores units in sparse contexts that match the final frame budget. Twelve random partitions of the 32 units into four groups of eight give 48 subsets, each of which keeps 32 frames. A ridge regression of each subset's negative KL divergence from the complete-pool distribution on unit-presence indicators gives one coefficient per unit, and the eight units with the largest coefficients are selected (Appendix A.4). Its 48 subset probes come on top of the leave-one-out probes, so it costs more than KEYSTONE without raising accuracy.

5 CONCLUSION

Keyframe selection is one way to answer a question about a video lasting tens of minutes. Scoring a frame alone can overvalue evidence already supplied by the rest of the video. We propose a novel keyframe selection method, KEYSTONE, which instead scores in context through leave-one-out answer distributions, using their centroid as a shared reference on the same frozen answering model. KEYSTONE needs neither training nor an external scorer. Across four benchmarks and four answering models, it improves over uniform sampling by up to 7.1 points and over the strongest competing selector by up to 2.3 points. The gain depends on both the context and the question, since scoring the same units alone costs up to 4.6 points and swapping the question costs up to 5.8 points. The gains persist at 72B and across all four budgets from 8 to 64 frames. Extending KEYSTONE to open-ended questions on long videos and adapting its frame budget to each question are left for future work.

AI USE STATEMENT

In this work, we used generative AI tools for aiding or polishing the writing, for retrieval and discovery, and for research ideation or execution. We made the research decisions and reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Upon acceptance we will release the code. Appendix A gives the evaluation protocol and the implementation details, with the implementation-level algorithm and its constants in Appendix A.3, the verbatim prompts in Appendix A.5, the evaluation harness in Appendix A.6, and the timing protocol in Appendix A.7.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in video-language understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2907–2917, 2022. doi: 10.1109/CVPR52688.2022.00293. URL https://doi.org/10.1109/CVPR52688.2022.00293.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=Zj12nzlQbz.

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 *European Conference on Computer Vision*, pp. 19–35, 2024b. doi: 10.1007/978-3-031-73004-7_2. URL https://doi.org/10.1007/978-3-031-73004-7_2.

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

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. LongVILA: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wCXAlfvCy6.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024c. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in Video-LLMs, 2024. URL https://arxiv.org/abs/2406.07476.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, pp. 95764–95807, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/adbea136219b64db96a9941e4249a857-Abstract-Conference.html.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Mohamed Eltahir, Lama Ayash, Ali Habibullah, Tanveer Hussain, and Naeemullah Khan. GridProbe: Posterior-probing for adaptive test-time compute in long-video VLMs, 2026. URL https://arxiv.org/abs/2605.10762.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. MIST: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14773–14783, 2023. doi: 10.1109/CVPR52729.2023.01419. URL https://doi.org/10.1109/CVPR52729.2023.01419.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13702–13712, 2025. doi: 10.1109/CVPR52734.2025.01279. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_M-LLM_Based_Video_Frame_Selection_for_Efficient_Video_Understanding_CVPR_2025_paper.html.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. FRAG: Frame selection augmented generation for long video and long document understanding, 2025a. URL https://arxiv.org/abs/2504.17447.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025b. doi: 10.18653/v1/2025.findings-acl.1024. URL https://aclanthology.org/2025.findings-acl.1024/.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024. doi: 10.1109/CVPR52733.2024.01300. URL https://doi.org/10.1109/CVPR52733.2024.01300.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom

Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Sanghwan Kim, Rui Xiao, Stephan Alaniz, Yongqin Xian, and Zeynep Akata. Training-free uncertainty guidance for complex visual tasks with MLLMs. In *European Conference on Computer Vision*, pp. 606–624, 2026. doi: 10.1007/978-3-032-37020-4_33. URL https://doi.org/10.1007/978-3-032-37020-4_33.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Hosu Lee, Junho Kim, Hyunjun Kim, and Yong Man Ro. ReFoCUS: Reinforcement-guided frame optimization for contextual understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Findings)*, pp. 8291–8302, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/papers/Lee_ReFoCUS_Reinforcement-guided_Frame_Optimization_for_Contextual_Understanding_CVPRF_2026_paper.pdf.

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*, 2025a. URL https://openreview.net/forum?id=zKv8qULV6n.

Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. FrameOracle: Learning what to see and how much to see in videos. In *International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=Vd5bjrFQBh.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11369–11380, 2026b. 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, Caiming Xiong, and Steven C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025b. doi: 10.1007/s11432-024-4321-9. URL https://doi.org/10.1007/s11432-024-4321-9.

Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan V. Oseledets. MaxInfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7198–7207, 2026c. doi: 10.1109/WACV61042.2026.00695. URL https://doi.org/10.1109/WACV61042.2026.00695.

Yulin Li, Haokun Gui, Ziyang Fan, Junjie Wang, Bin Kang, Bin Chen, and Zhuotao Tian. Less is more, but where? Dynamic token compression via LLM-guided keyframe prior. In *Advances in Neural Information Processing Systems*, 2025c. URL https://openreview.net/forum?id=uhFx1RGD1g.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. VideoINSTA: Zero-shot long video understanding via informative spatial-temporal reasoning with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6577–6602, 2024. doi: 10.18653/v1/2024.findings-emnlp.384. URL https://aclanthology.org/2024.findings-emnlp.384/.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024a. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26679–26689, 2024b. doi: 10.1109/CVPR52733.2024.02520. URL https://doi.org/10.1109/CVPR52733.2024.02520.

Shih-Yao Lin, Sibendu Paul, and Caren Chen. From captions to keyframes: KeyScore for multimodal frame scoring and video-language understanding, 2025. URL https://arxiv.org/abs/2510.06509.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: Boost large vision-language model without training for long-form video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3318–3327, 2025. doi: 10.1109/CVPR52734.2025.00315. 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.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Junpeng Ma, Sashuai Zhou, Guanghao Li, Xin Gao, Yue Cao, Hengyu Zeng, Yuxiang Yan, Zhibin Wang, Jun Song, Bo Zheng, Shanghang Zhang, and Jian Pu. GIFT: Global irreplaceability frame targeting for efficient video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25610–25620, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ma_GIFT_Global_Irreplaceability_Frame_Targeting_for_Efficient_Video_Understanding_CVPR_2026_paper.html.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. doi: 10.1109/CVPR52734.2025.01764. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://doi.org/10.18653/v1/2023.emnlp-main.557.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video QA. In *Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3569–3588, 2026. doi: 10.18653/v1/2026.eacl-long.164. URL https://aclanthology.org/2026.eacl-long.164/.

Haozhe Qi, Kevin Qu, Mahdi Rad, Rui Wang, Alexander Mathis, and Marc Pollefeys. AdaptToken: Entropy-based adaptive token selection for MLLM long video understanding, 2026. URL https://arxiv.org/abs/2603.28696.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya

Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025. doi: 10.1109/ICCV51701.2025.02122. URL https://doi.org/10.1109/ICCV51701.2025.02122.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP^3: A training-free approach for list-wise frame selection in Video-LLMs. In *IEEE/CVF International Conference on Computer Vision*, pp. 24090–24101, 2025. doi: 10.1109/ICCV51701.2025.02233. URL https://doi.org/10.1109/ICCV51701.2025.02233.

Xiaokun Sun, Yubo Wang, Haoyu Cao, and Linli Xu. When thinking hurts: Mitigating visual forgetting in video reasoning via frame repetition. In *European Conference on Computer Vision*, pp. 168–184, 2026. doi: 10.1007/978-3-032-37026-6_10. URL https://doi.org/10.1007/978-3-032-37026-6_10.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. TSPO: Temporal sampling policy optimization for long-form video language understanding. In *AAAI Conference on Artificial Intelligence*, pp. 9368–9376, 2026. doi: 10.1609/aaai.v40i11.37896. URL https://doi.org/10.1609/aaai.v40i11.37896.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025. doi: 10.1109/CVPR52734.2025.02711. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_LVBench_An_Extreme_Long_Video_Understanding_Benchmark_ICCV_2025_paper.html.

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, 2025b. URL https://arxiv.org/abs/2508.18265.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76, 2024a. doi: 10.1007/978-3-031-72989-8_4. URL https://doi.org/10.1007/978-3-031-72989-8_4.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pp. 396–416, 2024b. doi: 10.1007/978-3-031-73013-9_23. URL https://doi.org/10.1007/978-3-031-73013-9_23.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient frame selection for long videos at test time with attention-based MLLM selectors, 2026. URL https://arxiv.org/abs/2607.15689.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025c. doi: 10.1109/CVPR52734.2025.00311. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VideoTree_Adaptive_Tree-based_Video_Representation_for_LLM_Reasoning_on_Long_CVPR_2025_paper.html.

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*, pp. 28828–28857, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long video understanding with learnable retrieval in video-language models. *IEEE Transactions on Multimedia*, 28:5254–5261, 2026. doi: 10.1109/TMM.2026.3664996. URL https://doi.org/10.1109/TMM.2026.3664996.

Yiqing Yang, Yun Li, Daiqing Qi, Lehan Yang, Tianlong Wang, Wenhao Zhang, Sheng Li, and Kinman Lam. HFS: Holistic query-aware frame selection for efficient video understanding. In *ACM International Conference on Multimedia*, 2026. URL https://arxiv.org/abs/2512.11534. To appear.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *IEEE/CVF International Conference on Computer Vision*, pp. 20168–20179, 2025. doi: 10.1109/ICCV51701.2025.01876. URL https://doi.org/10.1109/ICCV51701.2025.01876.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17900–17917, 2025. doi: 10.18653/v1/2025.findings-acl.921. URL https://aclanthology.org/2025.findings-acl.921/.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pr37sbuhVa.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, pp. 76749–76771, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f22a9af8dbb348952b08bd58d4734b50-Abstract-Conference.html.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-Voyager: Learning to query frames for video large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LNL7zKvm7e.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53. URL https://doi.org/10.1007/978-3-319-10590-1_53.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision*, pp. 11941–11952, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

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.

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. URL https://openreview.net/forum?id=30RAWQVGlx.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025c. doi: 10.1109/ICCV51701.2025.02048. URL https://doi.org/10.1109/ICCV51701.2025.02048.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, 2025d. URL https://openreview.net/forum?id=UQJ7CDW8nb.

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*, pp. 89863–89895, 2025e. URL http://papers.nips.cc/paper_files/paper/2025/hash/8190b210e9808e54ee16263b673a847d-Abstract-Conference.html.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025f. URL https://openreview.net/forum?id=EElFGvt39K.

Yunzhu Zhang, Yu Lu, Tianyi Wang, Fengyun Rao, Yi Yang, and Linchao Zhu. FlexSelect: Flexible token selection for efficient long video understanding. In *Advances in Neural Information Processing Systems*, pp. 102751–102777, 2025g. URL https://openreview.net/forum?id=0D3ja9s17M.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems*, pp. 62557–62583, 2024. URL https://openreview.net/forum?id=VqkAKQibpq.

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 *IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278. 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 *AAAI Conference on Artificial Intelligence*, pp. 13791–13799, 2026. doi: 10.1609/aaai.v40i16.38387. URL https://doi.org/10.1609/aaai.v40i16.38387.

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://openreview.net/forum?id=1OQKqLFcbB.

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://openreview.net/forum?id=SZVpOKw0YD.

Section A gives the evaluation protocol and implementation details. Section B reports extended results across frame budgets, pool sizes, and data groups. Section C compares scoring-chain variants, Section D adds mechanism diagnostics, Section E applies the score to open-ended questions, Section F presents examples and failures, and Section G states the scope and limitations. Notation follows Table 1; bold marks the best result within the stated comparison group, and shading marks the deployed KEYSTONE configuration.

A PROTOCOL AND IMPLEMENTATION DETAILS

A.1 BENCHMARKS

All four benchmarks are multiple-choice with a single gold answer per question, and accuracy is the share of questions answered correctly, except on MLVU-Test, whose official score averages the accuracy of its nine question types. Video-MME holds 2,700 questions from 900 videos, with 300 videos and 900 questions per duration bucket, and the settings with and without subtitles use the same questions. MLVU-Test holds 502 questions from 349 videos. LongVideoBench-Val holds 1,337 questions, and its four official duration groups hold 189, 172, 412, and 564 questions from the shortest to the longest. LVBench holds 1,549 questions from 103 videos. The model receives no audio on any benchmark.

A.2 CANDIDATE POOL

Each video is decoded once and cached. The evenly spaced positions of the candidate pool start at the first cached frame and end at the last, each rounded down to a frame index. The probes and the answering pass each resize the cached frames to their resolution in Table 9. When uniform sampling takes evenly spaced positions over the candidate pool, each position maps to the nearest pool frame at every frame budget. The 32-frame uniform cells of LLaVA-OneVision on LongVideoBench-Val and of LLaVA-Video on MLVU-Test in Table 1 instead round each position down to a pool frame.

Table 9: Frame resolution of the probes and of the answering pass, per answering model. Qwen2.5-VL-72B in Table 2 uses the settings of the Qwen2.5-VL row.

Model	**Probes**	**Answering pass**	**Notes**
Qwen2.5-VL	50,176 pixels per frame	602,112 pixels per frame	the probe cap is one quarter of the per-frame budget that our evaluation setting gives a 128-frame input, bicubic resizing
LLaVA-OneVision	pooling stride 2	pooling stride 2	checkpoint image processor, bilinear pooling
LLaVA-Video	pooling stride 2	pooling stride 2	checkpoint image processor, bilinear pooling in the probes and average pooling in the answering pass
InternVL3.5	one 448 by 448 tile	one 448 by 448 tile	the default video setting of the model

A.3 IMPLEMENTATION OF THE SCORING CHAIN

In Algorithm 1 the row that masks no unit gives the posterior $\boldsymbol{\pi}_{[m]}$ on the complete candidate pool, which both the variants of Section A.4 that take $\boldsymbol{\pi}_{[m]}$ as their reference and the reference split of Figure 4 read. Table 10 lists the values of the constants in Algorithm 1. Every batch is padded to at least four rows, because a smaller batch calls a different attention kernel and drifts from the single forward pass. The probes run through the scaled-dot-product attention kernel, because the FlashAttention path drops the masked tokens and repacks the sequence instead of masking the attention scores. On Qwen2.5-VL the two-dimensional key-padding mask gives unit scores that correlate with those of the additive mask of Section 3.4 at 1.0000 and identical selections, and the attention implementations of all four answering models accept this form.

Algorithm 1 KEYSTONE scoring and selection, implementation level.

Require: the questions of a benchmark grouped by video, and for each video its cached candidate pool $\mathcal{P}$ of n evenly spaced frames; frozen answering model $\mathcal{M}$ with its tokenizer; unit width U; frame budget K; chunk size κ; floor ε
Ensure: for every question, K frames in temporal order, written to the index file of the question

for each video and its cached pool $\mathcal{P}$ **do**
  $F \leftarrow$ the frames of $\mathcal{P}$ at the probe resolution ▷ per-model budget in Table 9
  $E \leftarrow$ vision tower and projector on F
  $\mathcal{R}_j \leftarrow$ the positions of the tokens of unit j in the visual prefix, for $j = 1, \dots, m$ ▷ unit j holds frames $Uj{-}U{+}1$ to Uj
  $(\boldsymbol{K}_c, \boldsymbol{V}_c) \leftarrow$ prefill of the visual prefix with a key-value cache ▷ the prefix precedes the question text
  for each question q of the video **do**
    $\Pi \leftarrow$ the multiple-choice prompt of the benchmark for q (Table 11), without subtitles
    $\mathcal{A} \leftarrow$ the first token id of each option letter of q
    $D_0 \leftarrow \varnothing$; $D_j \leftarrow \{j\}$ for $j = 1, \dots, m$ ▷ row 0 is unmasked
    for each chunk G of at most κ consecutive rows of D **do**
      pad G to at least 4 rows by repeating its last row
      $\boldsymbol{\Omega} \leftarrow$ key-padding mask of ones over the cached and suffix positions, one row per row of G; set $\boldsymbol{\Omega}[r, \mathcal{R}_j] \leftarrow 0$ for every row r of G and every $j \in D_r$
      $z \leftarrow \mathcal{M}(\text{text suffix of } \Pi \mid \boldsymbol{K}_c, \boldsymbol{V}_c, \boldsymbol{\Omega})$, logits at the last position ▷ cache expanded to the rows of G
      $\boldsymbol{p}_r \leftarrow \mathrm{softmax}(z_r[\mathcal{A}])$ for every row r of G that is not padding
    end for
    $\boldsymbol{\pi}_{-j} \leftarrow \boldsymbol{p}_j$ clipped at ε and renormalized, for $j = 1, \dots, m$
    $\bar{\boldsymbol{\pi}} \leftarrow \frac{1}{m} \sum_j \boldsymbol{\pi}_{-j}$, clipped at ε and renormalized
    $s_j \leftarrow \sum_a \bar{\pi}_a \left(\log \bar{\pi}_a - \log \pi_{-j,a}\right)$ for $j = 1, \dots, m$
    $\mathcal{I}^\star \leftarrow$ the $\lfloor K/U \rfloor$ units with the largest s_j, by a descending argsort
    write the frames of $\mathcal{I}^\star$, sorted by timestamp, to the index file of q
  end for
end for

Table 10: Constants of the scoring chain and the answering pass.

Symbol	Value	Meaning
n	128	candidate frames per video
U, m	4, 32	frames per unit and units per question
K	32	frame budget, 8 units
κ	13, or 9 on InternVL3.5	rows per batched suffix pass
ε	10^{-12}	floor before every logarithm

A.4 VARIANTS USED IN THE ABLATIONS

The probe rows of scoring alone are computed in one batched pass with the unmasked row. The maximum-probability readout takes the option $a^\star$ with the highest probability under the centroid $\bar{\boldsymbol{\pi}}$ and scores unit j by the drop $\bar{\pi}_{a^\star} - \pi_{-j,a^\star}$ in the probability of that option when the unit is removed.

Black-frame replacement runs a second prefill per video on the same number of all-black frames. The probe of unit j replaces the cached keys and values of that unit by those of the black frames at the same positions and leaves the other units unchanged.

Stepwise greedy starts from the empty set and runs $\lfloor K/U \rfloor$ steps. At each step, every unit u outside the held set gets one probe that keeps the held set and u and masks the other units, and the unit whose probe posterior has the smallest KL divergence from $\boldsymbol{\pi}_{[m]}$ joins the held set, which adds 228 probes per question to the 32 leave-one-out probes.

Coalition regression uses the question identifier as the random seed. The ridge fit uses the 48 subset responses, an intercept, and a penalty of 1.0 on the unit coefficients, with no penalty on the intercept. The reported implementation also computes the 32 leave-one-out probes and the unmasked row, giving 81 probe rows per question; the leave-one-out responses are not used in the regression.

Full-resolution probes raise the pixel cap of the Qwen2.5-VL probes to the per-frame budget that our evaluation setting gives a 128-frame input and keep every other step. The width sweep of Table 4 keeps the batch size. The budget sweeps of Table 6 and Section B.1 reuse one set of scores per question and change only the number of units kept.

A.5 PROMPT TEMPLATES

The probes and the answering pass without subtitles receive the same prompt from Table 11, and each answering model wraps it in its own chat template with its default system prompt. In the setting with subtitles, the harness inserts the subtitle lines that cover 32 evenly spaced frames of the whole video. These lines do not change with the frames a selector picks, so every arm receives the same subtitles for a given question.

Table 11: The question-answering prompt of each benchmark in our runs. Each prompt is its lmms-eval v0.6 task template verbatim, except that on MLVU-Test we add the final instruction line, which the template leaves out. Typewriter text is literal and bracketed italic text is a placeholder filled per question.

Benchmark	**Question-answering prompt**

Video-MME	`Select the best answer to the following multiple-choice question based on the video and the subtitles. Respond with only the letter (A, B, C, or D) of the correct option.`
	[question]
	[options, one per line, A. to D.]
	[blank line]
	`Answer with the option's letter from the given choices directly.`

Video-MME with subtitles	`This video's subtitles are listed below:`
	[subtitle lines]
	`Select the best answer to the following multiple-choice question based on the video and the subtitles. Respond with only the letter (A, B, C, or D) of the correct option.`
	[question]
	[options, one per line, A. to D.]
	`The best answer is:`

MLVU-Test	*[question with the lettered options embedded as (A), (B), and so on]* `Answer with the option letter from the given choices directly.`
LongVideoBench-Val	*[question]* *[options, one per line, A. to E.]* `Answer with the option's letter from the given choices directly.`
LVBench	*[question with the options embedded as (A) to (D)]* `Answer the question with the option letter`

A.6 EVALUATION HARNESS

Each lmms-eval adapter is patched with a frame-index argument that replaces uniform sampling with the selected frames and preserves all other processing. Decoding follows the generation settings of each task definition, with a new-token cap of 16 on Video-MME, MLVU-Test, and LVBench and of 32 on LongVideoBench-Val. The predicted letter is read by the stock parser of each task. Video-MME and LVBench strip standard answer prefixes and take the first option letter, MLVU-Test reads the character before the first closing parenthesis and requires an exact match, and LongVideoBench-Val takes the first option letter and falls back to a random option when none is found. All experiments run on NVIDIA H200 GPUs, and every accuracy cell comes from one evaluation run.

A.7 TIMING PROTOCOL

The videos of the timing sample are drawn with a fixed seed, four long, three medium, and three short. The wall clock of a selection process runs from interpreter start to process exit, so it covers the imports, the model load, the reading of the frame pool, the scoring, and the writing of the index file, and uniform sampling and the dense arm have no selection process. The answering time is measured in a separate process that loads the answering model once and times each question under

device synchronization, from reading the frames through the processor to the generation of at most 8 new tokens, at batch size 1. In the evaluation runs of the timed arms every answer is a single letter, so this cap generates the same tokens as the evaluation cap of 16. The per-frame pixel budget of the answering pass follows the evaluation runs, and on Qwen2.5-VL it falls to 200,704 pixels when all 128 frames enter at once.

B Extended Results

B.1 Frame Budget on LVBench

KEYSTONE exceeds uniform sampling at every budget for every answering model, with the smallest margin at 64 frames (Table 12).

Table 12: Frame budget sweep on LVBench for the four answering models, accuracy (%). Columns give the frame budget.

Answering model	Method	8	16	32	64
Qwen2.5-VL	Uniform	33.25	37.38	39.06	40.41
	KEYSTONE	**40.41**	**42.48**	**42.80**	**43.38**
LLaVA-OneVision	Uniform	35.18	37.64	39.12	41.58
	KEYSTONE	**41.83**	**42.61**	**43.96**	**42.35**
LLaVA-Video	Uniform	36.41	36.48	40.41	43.00
	KEYSTONE	**45.06**	**45.71**	**45.32**	**45.13**
InternVL3.5	Uniform	38.09	41.77	42.48	44.03
	KEYSTONE	**47.19**	**47.51**	**47.97**	**48.61**

B.2 Candidate Pool Size

In Table 13, pools of 64 and 96 frames fall below the 128-frame pool on all three benchmarks. A smaller pool also has fewer probes, yet the last row is on a par with the other 64-frame row on all three benchmarks, so more probes do not recover the loss, and the loss comes from the sparser candidate frames.

Table 13: Candidate pool size, Qwen2.5-VL-7B accuracy (%). Width and probes follow Table 4, and the last row halves the width so that the number of probes matches the deployed row.

Pool	Width	Probes	VMME	LVB	MLVU
64	4	16	64.07	60.28	50.39
96	4	24	64.74	61.18	51.80
128	4	32	**65.93**	**61.93**	**52.87**
64	2	32	64.22	60.06	51.03

B.3 Unit Width on the Other Answering Models

U is a hyperparameter introduced by the method. A width of 4 halves the probes of a width of 2 and moves accuracy by under one point on each benchmark. At a width of 8 Video-MME begins to drop, at 16 MLVU drops as well, and at 32 the method can only select one of four contiguous blocks and all three benchmarks drop markedly. On the other three answering models, changing the unit width from 2 to 4 changes accuracy by at most 1.1 points, with no consistent direction (Table 14).

B.4 Where the Gain Lands

Gains in Tables 15 to 18 extend across most groups. LVBench summarization falls below uniform sampling for all four answering models, while the remaining declines depend on the model, with the shortest LongVideoBench group declining only for LLaVA-OneVision.

Table 14: Unit width 2 against 4, accuracy (%). Both columns are the deployed centroid-reference arm on the candidate grid of the same cell in Table 1.

Benchmark	Answering model	Width 2	Width 4
VMME	LLaVA-OneVision	**61.30**	60.96
	LLaVA-Video	65.37	**65.89**
	InternVL3.5	**66.74**	66.48
LVB	LLaVA-OneVision	58.94	**59.39**
	LLaVA-Video	**59.99**	**59.99**
	InternVL3.5	63.13	**64.25**
MLVU	LLaVA-OneVision	**51.37**	50.69
	LLaVA-Video	**50.77**	50.59
	InternVL3.5	54.78	**55.59**

Table 15: MLVU accuracy (%) by official question type. The benchmark metric is the macro average over the nine types.

Answering model	Method	Anomaly	Count	Ego	Needle	Order	Plot	Sports	Topic	Tutorial
Qwen2.5-VL	Uniform	**46.15**	21.67	43.40	48.33	34.29	46.00	**50.00**	86.81	37.21
	KEYSTONE	**46.15**	**31.67**	**49.06**	**55.00**	**51.43**	**54.00**	47.22	**90.11**	**51.16**
LLaVA-OneVision	Uniform	53.85	23.33	**60.38**	53.33	24.29	50.00	**44.44**	**89.01**	25.58
	KEYSTONE	**56.41**	**26.67**	52.83	**68.33**	**41.43**	**56.00**	36.11	83.52	**34.88**
LLaVA-Video	Uniform	46.15	23.33	62.26	40.00	35.71	54.00	41.67	68.13	**39.53**
	KEYSTONE	**51.28**	**33.33**	**67.92**	**43.33**	**50.00**	**64.00**	**47.22**	**70.33**	27.91
InternVL3.5	Uniform	46.15	25.00	54.72	46.67	35.71	50.00	**47.22**	**84.62**	46.51
	KEYSTONE	**51.28**	**26.67**	**66.04**	**60.00**	**51.43**	**64.00**	**47.22**	80.22	**53.49**

Table 16: Video-MME accuracy (%) by (a) task category and (b) video category, with uniform sampling and KEYSTONE on four answering models.

(a) Task category

Task category	#Q	Qwen2.5-VL		LLaVA-OneVision		LLaVA-Video		InternVL3.5	
		Unif.	KEYSTONE	Unif.	KEYSTONE	Unif.	KEYSTONE	Unif.	KEYSTONE
Object reasoning	454	57.49	**61.45**	56.17	**57.71**	60.79	**63.00**	58.81	**64.32**
Object recognition	354	66.38	**71.75**	65.54	**69.49**	70.34	**72.03**	**72.88**	70.62
Information synopsis	323	75.85	**78.02**	73.68	**75.54**	78.33	**79.57**	**79.88**	**79.88**
Action recognition	313	63.90	**70.29**	**57.83**	57.51	63.58	**67.41**	64.54	**67.41**
Action reasoning	285	56.84	**60.35**	51.93	**57.19**	**57.19**	56.84	58.60	**58.95**
Counting	268	**42.16**	**42.16**	34.70	**35.82**	42.91	**45.15**	**45.52**	43.66
Attribute perception	222	75.68	**76.58**	73.42	**76.13**	77.03	**77.93**	**79.28**	78.38
Temporal reasoning	177	44.63	**55.93**	41.81	**49.72**	48.59	**51.41**	46.33	**49.72**
OCR	139	67.63	**71.94**	**64.03**	61.87	63.31	**70.50**	68.35	**76.26**
Spatial reasoning	56	78.57	**82.14**	76.79	**83.93**	**87.50**	85.71	82.14	**85.71**
Temporal perception	55	65.45	**70.91**	**63.64**	56.36	**76.36**	74.55	76.36	**78.18**
Spatial perception	54	**74.07**	66.67	57.41	**62.96**	61.11	**66.67**	**77.78**	74.07

(b) Video category

Video category	#Q	Qwen2.5-VL		LLaVA-OneVision		LLaVA-Video		InternVL3.5	
		Unif.	KEYSTONE	Unif.	KEYSTONE	Unif.	KEYSTONE	Unif.	KEYSTONE
Knowledge	810	63.58	**67.28**	60.25	**61.60**	64.81	**66.67**	67.53	**68.64**
Life record	630	56.83	**66.03**	56.67	**60.95**	63.81	**65.40**	62.54	**65.87**
Sports competition	450	64.00	**65.56**	55.33	**57.33**	62.22	**62.44**	61.33	**61.78**
Film and television	360	65.28	**66.67**	61.39	**63.61**	65.83	**70.00**	68.06	**69.17**
Artistic performance	360	**63.89**	63.06	61.11	**62.78**	63.06	**66.11**	**66.94**	66.67
Multilingual	90	56.67	**63.33**	52.22	**55.56**	58.89	**62.22**	60.00	**63.33**

C SCORING-CHAIN REPLACEMENTS

C.1 SIX POSITIONS ON QWEN2.5-VL AND LLAVA-VIDEO

Reverse KL and total variation retain over 90% of the deployed selection and stay within 0.6 accuracy points, whereas the maximum-probability readout lowers accuracy in all four cells (Table 19). Greedy and coalition estimators have the lowest selection overlaps.

Table 17: LongVideoBench accuracy (%) by official duration group. Column labels give the upper bounds of four disjoint groups, rather than cumulative subsets.

Answering model	Method	≤15 s	≤60 s	≤600 s	≤3600 s
Qwen2.5-VL	Uniform	71.96	73.26	56.07	47.34
	KEYSTONE	**74.07**	**78.49**	**60.44**	**53.90**
LLaVA-OneVision	Uniform	**74.07**	71.51	55.83	48.40
	KEYSTONE	69.31	**76.16**	**58.01**	**51.95**
LLaVA-Video	Uniform	67.20	**72.09**	58.01	50.89
	KEYSTONE	**67.72**	**72.09**	**63.35**	**51.24**
InternVL3.5	Uniform	69.31	**75.00**	59.22	53.72
	KEYSTONE	**72.49**	73.84	**64.32**	**58.51**

Table 18: LVBench accuracy (%) by official question type on four answering models. Questions may carry multiple type labels and therefore contribute to more than one column; the columns do not form a partition of the benchmark.

Answering model	Method	Entity recog.	Event underst.	Key info. retrieval	Temporal grounding	Reasoning	Summarization
Qwen2.5-VL	Uniform	37.08	40.65	40.89	37.73	40.30	**29.31**
	KEYSTONE	**42.10**	**43.89**	**45.36**	**38.64**	**43.28**	27.59
LLaVA-OneVision	Uniform	40.18	34.47	41.58	35.45	39.80	**39.66**
	KEYSTONE	**46.53**	**39.57**	**44.67**	**41.36**	**47.26**	37.93
LLaVA-Video	Uniform	41.65	38.33	40.21	36.82	42.29	**37.93**
	KEYSTONE	**46.53**	**43.12**	**47.42**	**37.73**	**46.77**	36.21
InternVL3.5	Uniform	42.98	40.80	43.64	36.36	45.77	**41.38**
	KEYSTONE	**49.78**	**46.06**	**49.83**	**40.91**	**50.25**	39.66

The dense reference changes nearly 30% of selected frames while staying within 0.5 accuracy points; its different score alignment is shown in Figure 4b. The DPP term gives no consistent improvement, and black-frame replacement stays within one point.

The three probe resolutions stay within 0.8 accuracy points of the deployed arm on Qwen2.5-VL. Reduced LLaVA-Video probe resolution lowers all three Video-MME results. Repeating the deployed scoring gives selection overlap above 0.95 and accuracy within one point on both answering models.

C.2 Scoring-Chain Replacements on the Other Two Models

The observations of Section C.1 extend to the other two answering models (Table 20). The dense reference and the full-distribution readouts remain close to the deployed accuracy, while the maximum-probability readout, the DPP term, and coalition regression lower it in all four cells. Black-frame replacement stays within one point except for LLaVA-OneVision on LVB, where it is 2.40 points lower. Repeating the deployed scoring gives selection overlap above 0.95 and accuracy within one point on both answering models. On LLaVA-OneVision, stepwise greedy lowers accuracy in both cells and has the lowest selection overlap, and the three probe resolutions stay within 1.3 accuracy points of the deployed arm (Table 21).

D Mechanism Diagnostics

D.1 Gold-Answer Leakage Check

The own-answer correlations of the dense reference in Figure 4b are small in magnitude, so Table 22 reports the within-question top-1 hit rate. The top-1 hit rate is around one half in every cell, far above chance, so the check has the sensitivity to detect a dependence, which makes the near-zero gold-answer correlation of Figure 4a informative. The question-level correlation is negative in all twelve cells. A reference borrowed from another question gives an own-answer correlation of about -0.5 in Figure 4b, showing that the measure detects dependence where present.

Table 19: Replacing the implementation at each of the six positions of the scoring chain. Each cell is the accuracy (%) followed by the selection overlap, the per-question mean share of selected frames that coincide with the deployed arm. The deployed self-overlap is 1 by definition and is omitted. The dense reference is the *w/o centroid reference* arm of Section 4.3, and black-frame replacement, coalition regression, stepwise greedy, and the maximum-probability readout follow Section A.4. Reverse KL scores a unit by $D_{\mathrm{KL}}(\boldsymbol{\pi}_{-j} \,\|\, \bar{\boldsymbol{\pi}})$, and total variation by half the sum over options of $|\pi_{-j,a} - \bar{\pi}_a|$. The determinantal point process term adds eight units one at a time, each maximizing the log-determinant of the kernel $\mathrm{diag}(\boldsymbol{q})\, \boldsymbol{S}\, \mathrm{diag}(\boldsymbol{q})$ restricted to the selected units, where q_j is the exponential of the score rescaled to the unit interval times a temperature of 2 and $\boldsymbol{S}$ is a Gaussian kernel with a bandwidth of 0.10 on unit start times rescaled to the unit interval. On Qwen2.5-VL the High row is the full-resolution probe arm of Table 8.

Variant	Qwen2.5-VL		LLaVA-Video	
	VMME	LVB	VMME	LVB
Centroid-reference KEYSTONE	65.93	61.93	65.89	59.99
Reference				
Dense reference	66.11 / 0.711	61.78 / 0.715	66.15 / 0.710	59.54 / 0.722
Removal				
Black-frame replacement	65.44 / 0.608	61.26 / 0.604	66.59 / 0.576	59.46 / 0.586
Estimator				
Coalition regression	65.59 / 0.339	61.18 / 0.351	63.96 / 0.327	59.61 / 0.341
Stepwise greedy	64.11 / 0.268	60.06 / 0.280	63.37 / 0.238	59.31 / 0.264
Resolution				
High at 200,704 pixels or stride 3	66.11 / 0.451	62.30 / 0.451	64.93 / 0.434	60.43 / 0.447
Medium at 100,352 pixels or stride 4	66.00 / 0.465	62.60 / 0.465	64.81 / 0.420	61.03 / 0.424
Low at 25,088 pixels or stride 6	65.22 / 0.438	61.93 / 0.442	64.85 / 0.395	59.61 / 0.404
Readout				
Reverse KL	66.11 / 0.988	62.30 / 0.989	65.89 / 0.987	60.21 / 0.989
Total variation	65.52 / 0.919	62.53 / 0.907	66.19 / 0.912	59.61 / 0.907
Maximum-probability readout	64.96 / 0.415	59.54 / 0.402	64.70 / 0.414	59.76 / 0.411
Diversity				
Determinantal point process term	65.07 / 0.620	61.48 / 0.613	65.00 / 0.637	60.06 / 0.642
Replication				
Re-run of the deployed scoring	65.96 / 0.996	61.85 / 0.998	66.30 / 0.958	59.39 / 0.951

Table 20: Scoring-chain replacements on the other two answering models. Cells report accuracy (%) / selection overlap as in Table 19.

Variant	LLaVA-OneVision		InternVL3.5	
	VMME	LVB	VMME	LVB
Centroid-reference KEYSTONE	60.96	59.39	66.48	64.25
Reference				
Dense reference	61.15 / 0.721	58.49 / 0.748	66.78 / 0.704	63.72 / 0.736
Removal				
Black-frame replacement	61.19 / 0.582	56.99 / 0.615	66.44 / 0.597	63.43 / 0.625
Estimator				
Coalition regression	58.78 / 0.321	55.95 / 0.332	64.93 / 0.324	62.53 / 0.335
Readout				
Reverse KL	60.89 / 0.987	59.46 / 0.988	66.59 / 0.978	64.47 / 0.983
Total variation	60.85 / 0.916	58.86 / 0.911	66.85 / 0.912	64.40 / 0.912
Maximum-probability readout	58.96 / 0.418	57.29 / 0.422	64.96 / 0.416	62.45 / 0.414
Diversity				
Determinantal point process term	60.48 / 0.637	58.34 / 0.636	65.93 / 0.637	64.10 / 0.639
Replication				
Re-run of the deployed scoring	60.63 / 0.956	59.24 / 0.958	66.41 / 0.997	64.25 / 0.995

Table 21: Stepwise greedy and probe resolution on LLaVA-OneVision. Cells report accuracy (%) / selection overlap as in Table 19.

Variant	VMME	LVB
Centroid-reference KEYSTONE	60.96	59.39
Estimator		
Stepwise greedy	58.93 / 0.225	56.84 / 0.233
Resolution		
High at stride 3	61.44 / 0.430	58.12 / 0.442
Medium at stride 4	61.37 / 0.415	58.49 / 0.434
Low at stride 6	60.63 / 0.391	59.31 / 0.403

Table 22: Within-question top-1 hit and across-question correlation for the dense-reference variant. Top-1 hit is the percentage of questions where the highest-scoring unit also has the largest own-answer probability drop; chance is $1/32$. Across-question ρ is the rank correlation between the mean top-8 score and whether the dense answer is correct.

Model	Top-1 hit (%)	Across-question ρ
VMME		
Qwen2.5-VL	59.3	−0.268
LLaVA-OneVision	58.3	−0.275
LLaVA-Video	57.6	−0.266
InternVL3.5	59.8	−0.367
LVB		
Qwen2.5-VL	54.6	−0.179
LLaVA-OneVision	54.4	−0.193
LLaVA-Video	55.0	−0.182
InternVL3.5	61.2	−0.277
MLVU		
Qwen2.5-VL	57.8	−0.258
LLaVA-OneVision	49.0	−0.129
LLaVA-Video	60.6	−0.175
InternVL3.5	59.8	−0.279

D.2 OUTCOMES BY DENSE-ANSWER CORRECTNESS

Table 23 gives the per-cell results, and in every cell more than a third of the questions fall in the stratum where the dense answer is wrong. Table 24 compares the dense-right minus dense-wrong gain gap across methods, and for the deployed arm the difference is positive in nine of the twelve cells. The stratification could create such a difference on its own, since the questions with a wrong dense answer are harder for the model and leave a different margin to any selector. AKS and BOLT score frames without reading the answering model, and under the same stratification their mean difference over the cells is negative, so the stratification does not produce a positive difference on its own. The positive difference of KEYSTONE is consistent with a score read from the posterior of the answering model.

D.3 QUESTION-SWAP CONTROL

The swap shifts the question cyclically over the question identifiers of one video, selects frames with the shifted question, and still answers the original one. The three groups of rows in Table 25 score with the true question, score with a swapped question, and hold null and baseline selections.

The paired-uniform control takes eight evenly spaced units on the 32-unit grid of each question, so its frames arrive in eight blocks of four. A question alone in its video has no sibling to be swapped with and keeps its own frames, so the swapped arm carries a share of true-question selections. The dilution-matched null reproduces that share by taking the paired-uniform selection wherever a sibling exists and the true-question frames elsewhere. Video-MME holds three questions per video, so the two nulls coincide there. The degenerate baseline takes 32 contiguous frames at the center of the video.

Table 23: Gains of KEYSTONE over uniform sampling, split by whether the dense answer is correct. Δ_R and Δ_W are accuracy gains in points on dense-right and dense-wrong questions. Dense-answer accuracy, macro-averaged over question types on MLVU as in Table 1, ranges from 46.38% to 65.26% across these cells. On MLVU, Δ_R and Δ_W are also macro averages over question types, so the overall gain in Table 1 need not lie between them.

Model	Dense-right gain Δ_R	Dense-wrong gain Δ_W
VMME		
Qwen2.5-VL	+4.03	+3.43
LLaVA-OneVision	+2.09	+2.77
LLaVA-Video	+3.76	−1.03
InternVL3.5	+1.48	+1.28
LVB		
Qwen2.5-VL	+4.74	+5.60
LLaVA-OneVision	+2.78	+1.20
LLaVA-Video	+2.34	+1.15
InternVL3.5	+2.72	+5.69
MLVU		
Qwen2.5-VL	+11.85	+2.59
LLaVA-OneVision	+5.31	+4.47
LLaVA-Video	+6.02	+5.04
InternVL3.5	+9.91	+2.10

Table 24: Cross-method comparison of the gap between gains on dense-right and dense-wrong questions over the same twelve benchmark-model cells as Table 23. The mean gap is $\Delta_R - \Delta_W$ in points; positive values mean a larger gain on dense-right questions. Positive cells is the number of cells with $\Delta_R > \Delta_W$. The KEYSTONE row uses the deployed centroid reference.

Method	Mean gap $\Delta_R - \Delta_W$	Positive cells
KEYSTONE	+1.89	9 / 12
AKS	−0.23	4 / 12
BOLT	−3.16	4 / 12

Within the true-question group the dense reference is on a par with the centroid reference, and the stepwise greedy estimator is lower on all three benchmarks. After the swap both references fall and neither is separable from either null control, and both stay above the degenerate baseline.

Table 25: Question-swap control, Qwen2.5-VL-7B accuracy (%). Swapped scoring selects frames with another question of the same video and then answers the original question.

Method	VMME	LVB	MLVU
Scored with the true question			
Centroid-reference KEYSTONE	65.93	61.93	52.87
Dense-reference leave-one-out	66.11	61.78	53.90
Dense-reference stepwise greedy	64.11	60.06	48.79
Scored with a swapped question			
Centroid-reference KEYSTONE	60.15	59.69	50.10
Dense-reference leave-one-out	59.85	58.34	51.10
Null controls and baselines			
Paired-uniform control	60.07	58.41	48.33
Dilution-matched null	60.07	59.46	52.20
Contiguous center frames	55.85	54.15	40.63
Uniform	62.11	56.84	45.98

E OPEN-ENDED QUESTIONS

KEYSTONE reads the answer posterior over the lettered options, and an open-ended question supplies no such posterior. To test how much of the scoring carries over without options, we replace the option posterior with a readout over a generated answer and evaluate on the NExT-QA open-ended test split (Xiao et al., 2021), with 9,178 questions over 1,000 videos, drawn from a dataset whose videos average 44 seconds. The candidate pool, the units and the frame budget are those of the main experiments, and the answering model is Qwen2.5-VL-7B.

The readout starts from the model's own answer. The answering model reads the complete candidate pool and greedily generates an answer $\hat{y}$ of T tokens. With that answer fixed, the output of $\mathcal{M}$ at

answer position t is a distribution over the vocabulary $\mathcal{V}$,

$$\boldsymbol{\pi}_I^{(t)} = \mathrm{softmax}(\mathcal{M}(\mathcal{P}_I, q, \hat{y}_{<t})) \in \Delta^{|\mathcal{V}|-1}, \tag{13}$$

which takes the place of the answer posterior $\boldsymbol{\pi}_I$ of Section 3.1. The leave-one-out distributions $\boldsymbol{\pi}_{-j}^{(t)}$ and their centroid $\bar{\boldsymbol{\pi}}^{(t)}$ are defined at each position as in Section 3.3, and the score of unit j sums the divergence over at most the first four answer positions,

$$s_j = \sum_{t=1}^{\min(T,4)} D_{\mathrm{KL}}\left(\bar{\boldsymbol{\pi}}^{(t)} \,\middle\|\, \boldsymbol{\pi}_{-j}^{(t)}\right). \tag{14}$$

The ground-truth answer is read only by WUPS after answering, so selection never reads it, as in the main experiments.

In Table 26, KEYSTONE scores above uniform sampling at both thresholds, whereas AKS and BOLT exceed it at neither.

Table 26: WUPS (%) on the NExT-QA open-ended test split, 9,178 questions, with Qwen2.5-VL-7B answering from 32 frames selected out of a 128-frame candidate pool. WUPS@0.9 follows the lmms-eval scorer, which scales a WordNet word similarity below 0.9 by 0.1 against the first reference answer and applies no threshold against the second, and WUPS@0 applies no threshold.

Method	WUPS@0	WUPS@0.9
Uniform	32.27	25.85
AKS	32.17	25.72
BOLT	32.03	25.65
KEYSTONE	**32.68**	**26.31**

The benchmark leaves little room for frame selection, since 32 frames amount to one frame every 1.4 seconds. Whichever selector is used, 68% to 75% of the questions receive exactly the same answer as under uniform sampling, so the differences between selectors come from the remaining questions.

F QUALITATIVE ANALYSIS

F.1 WALKTHROUGH BOARDS

Figures 5 to 9 expand the contrast of Figure 1a into the complete scoring process on five questions. As in Figure 1, the upper curve plots the agreement $\exp(-D_{\mathrm{KL}}(\boldsymbol{\pi}_{[m]} \,\|\, \boldsymbol{\pi}_{\{j\}}))$ of each unit scored alone, and the lower curve plots the removal KL $D_{\mathrm{KL}}(\boldsymbol{\pi}_{[m]} \,\|\, \boldsymbol{\pi}_{-j})$ of each unit scored in context. Coral marks the unit ranked first in context. In Figure 1a it marks the units kept in context but not alone whose removal KL exceeds 0.01. On the first question, the eight units kept when scored alone all lie in the first 20 seconds of the video and include two pairs of nearly identical frames, whereas half of the units kept in context lie in the second half, including a shot of running along the water. On the second question, the units kept when scored alone spread from the start of the video to near its end, whereas the units kept in context concentrate in its middle and late parts, and the unit ranked first shows the food line at the event. On the third question, the unit ranked first in context is the title card that names the seventh handcraft. On the fourth, the unit that shows People's Square ranks 29th when scored alone and first in context. On the fifth, the unit that shows the final standings ranks last when scored alone and first in context.

F.2 REMOVING ONE UNIT AT A TIME

Figures 10 to 19 show what the in-context score responds to on ten further questions. The upper panel gives the probability that the frozen model assigns to the correct option after one unit is removed from the 128-frame pool, and the lower panel gives the removal KL of Figure 1a with the eight selected units marked. On every question the probability stays level across most units and falls at one of them, and that unit ranks first. Its frames show what the question asks for, such as a bedside clock at 6:00, a lipstick numbered 656, a street scene captioned 7:30pm, or a dish labeled $750. The correct option is used only to draw the upper panel and never enters the score.

F.3 FAILURE CASES

Figures 20 and 21 show two questions that KEYSTONE answers wrongly and uniform sampling answers correctly. The first asks which item does not appear. The traffic light spans two adjacent units, shaded gray, and each substitutes for the other under removal. They rank 32nd and 10th and are both discarded, and the model names the traffic light as the absent item, turning a correct dense answer into an error. Evidence shared across units can therefore be lost when the units are discarded together.

The second asks what happened between two events. The score retains activities supporting both the correct option and a distractor, because removing either can strongly displace the posterior. The return-to-dormitory and departure-to-study scenes that bound the interval, at the two ends of the gray band, both rank 13th and are omitted, leaving its boundaries unclear. The unit ranked first lies after the interval and shows the distractor, and removing it raises the probability of the correct option. A displacement-magnitude score can retain distractor evidence while missing temporal boundaries.

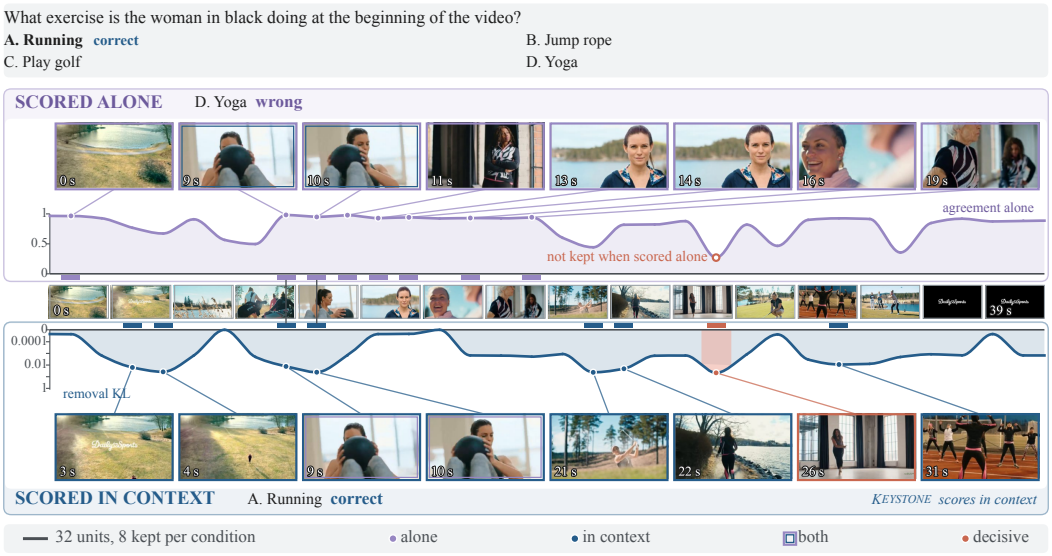

Figure 5: Walkthrough board for question 283-1 of Video-MME.

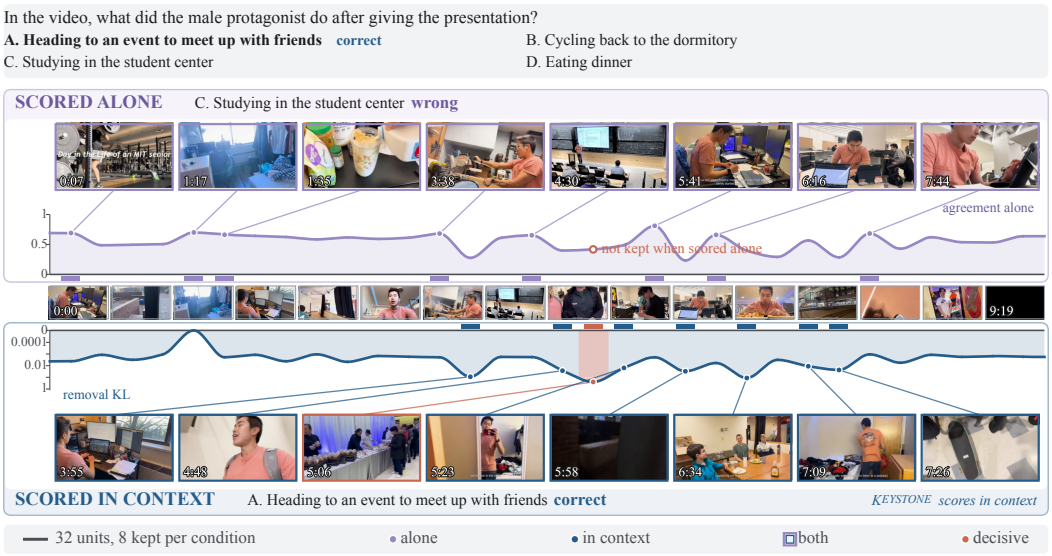

Figure 6: Walkthrough board for question 557-3 of Video-MME.

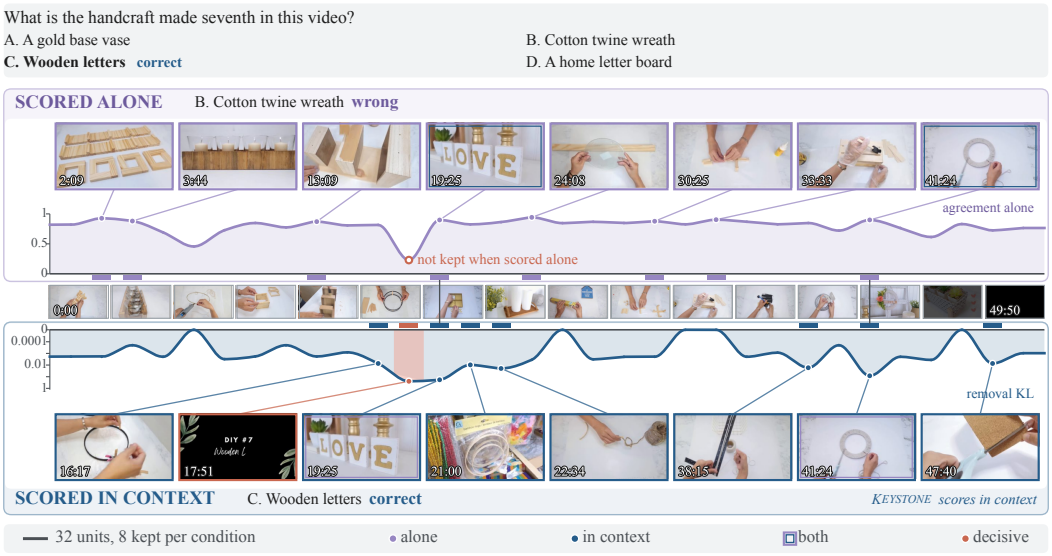

Figure 7: Walkthrough board for question 828-1 of Video-MME.

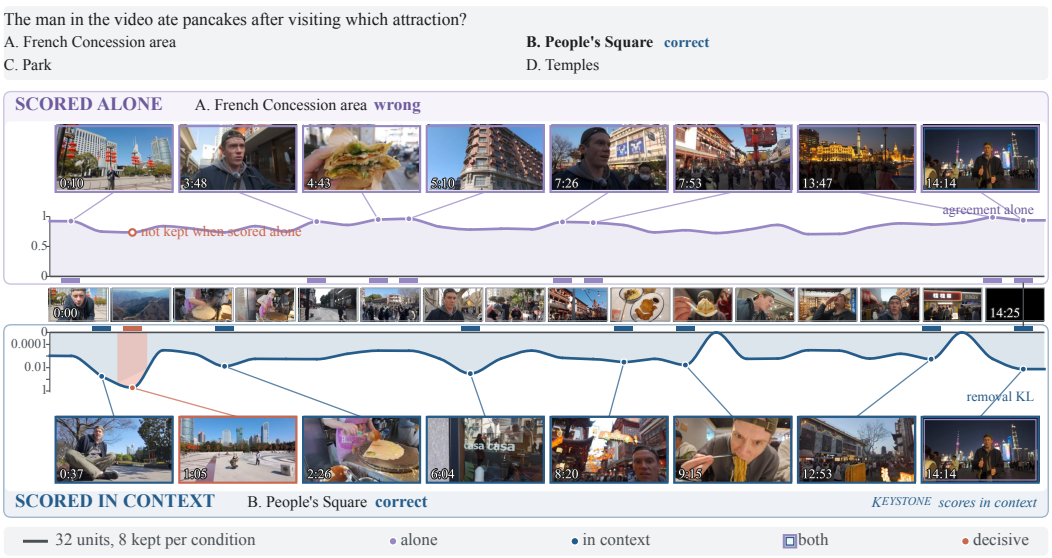

Figure 8: Walkthrough board for question 570-1 of Video-MME.

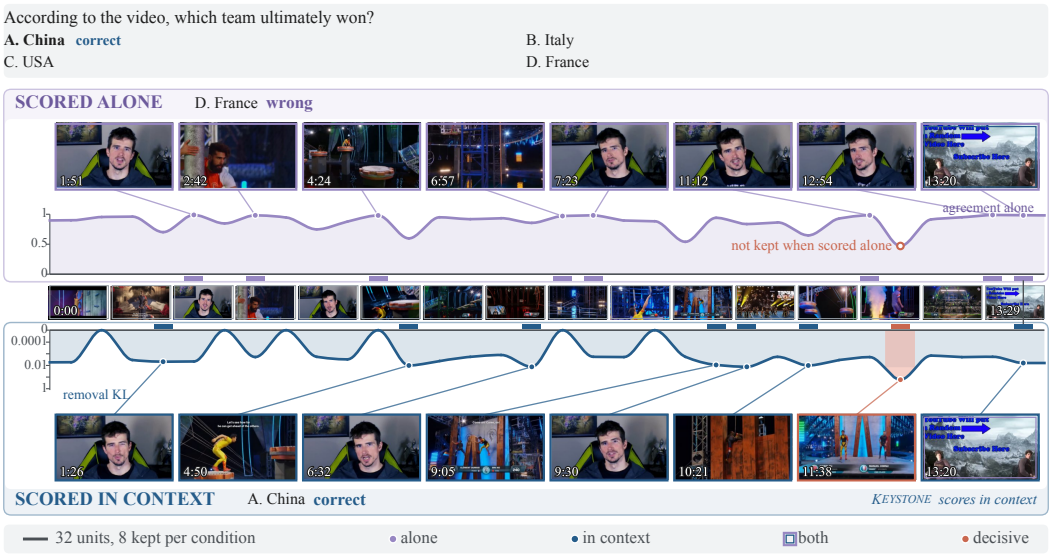

Figure 9: Walkthrough board for question 506-3 of Video-MME.

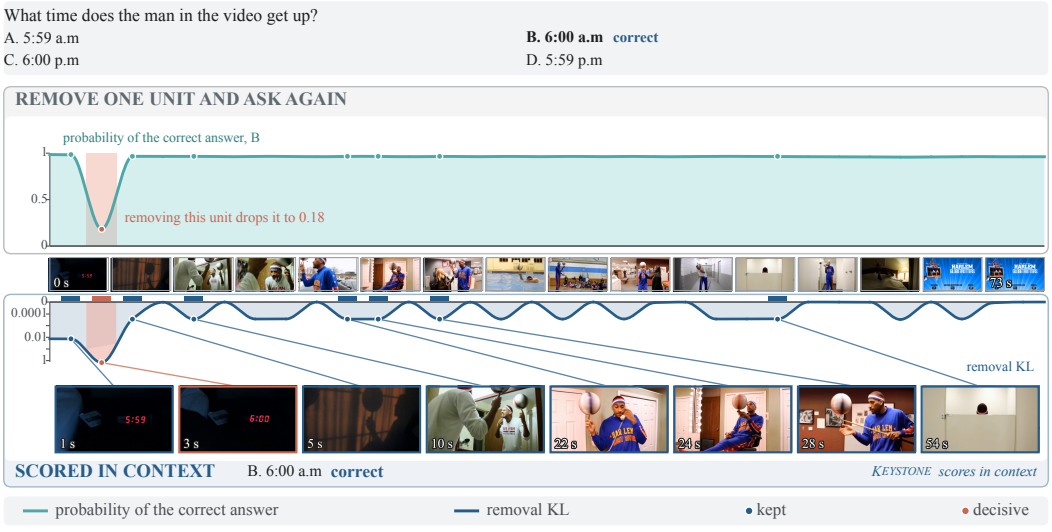

Figure 10: Removing one unit at a time, question 259-1 of Video-MME.

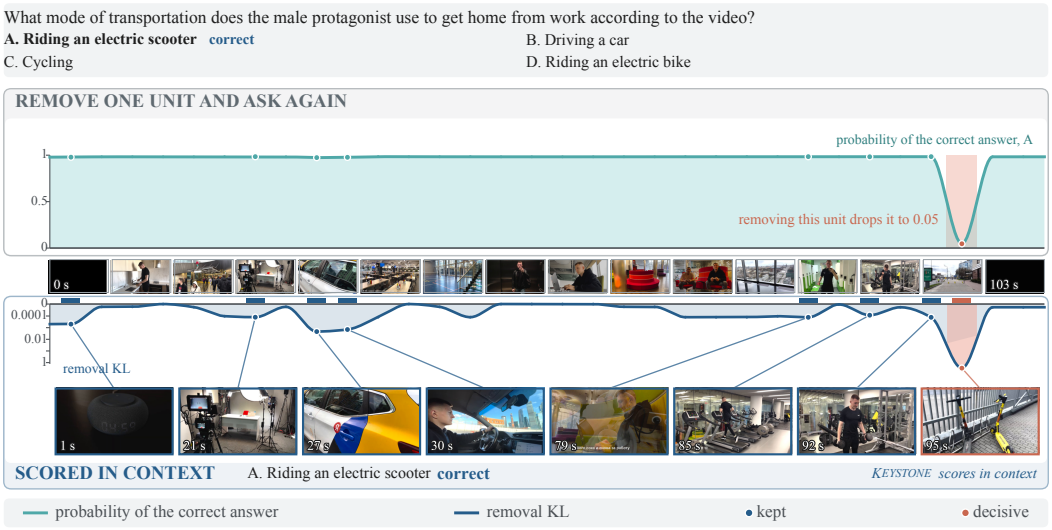

Figure 11: Removing one unit at a time, question 292-1 of Video-MME.

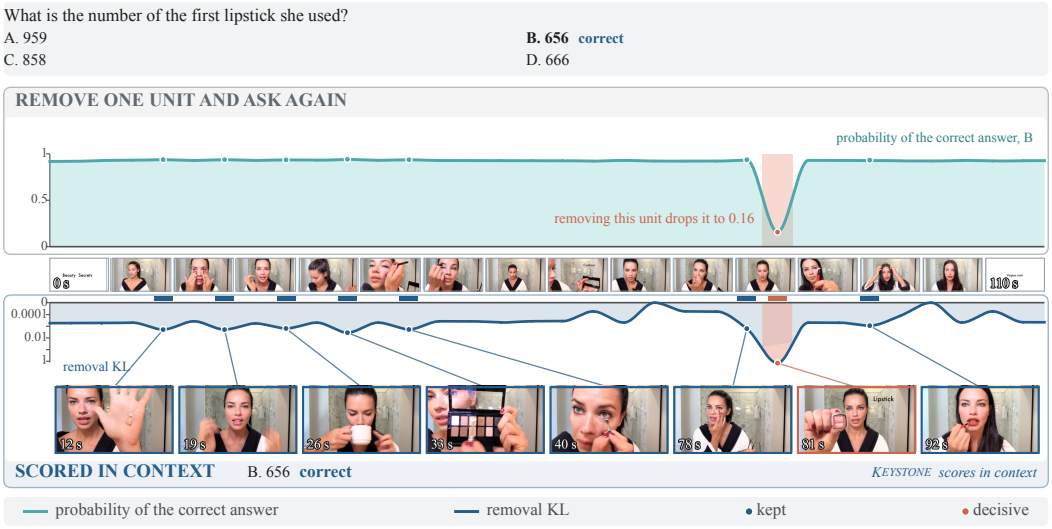

Figure 12: Removing one unit at a time, question 242-2 of Video-MME.

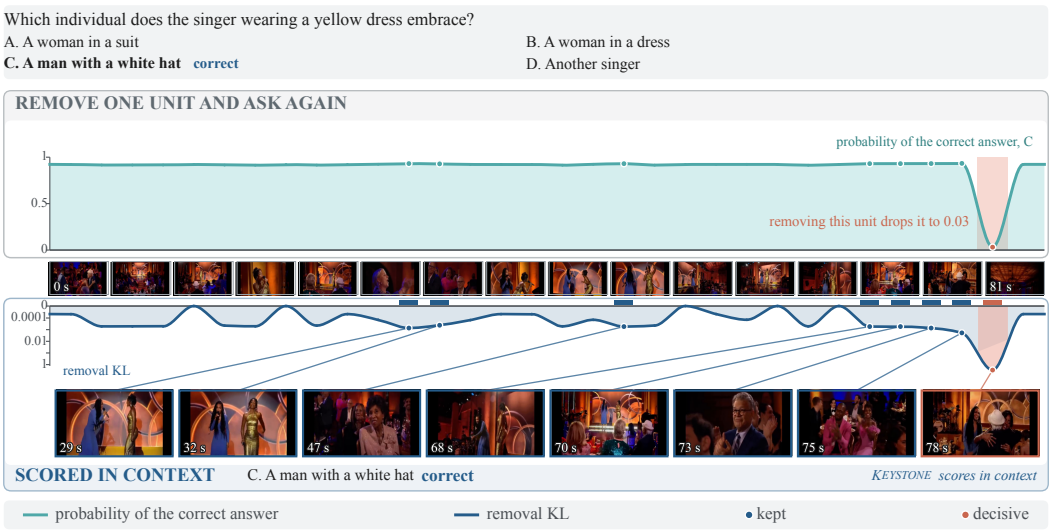

Figure 13: Removing one unit at a time, question 183-2 of Video-MME.

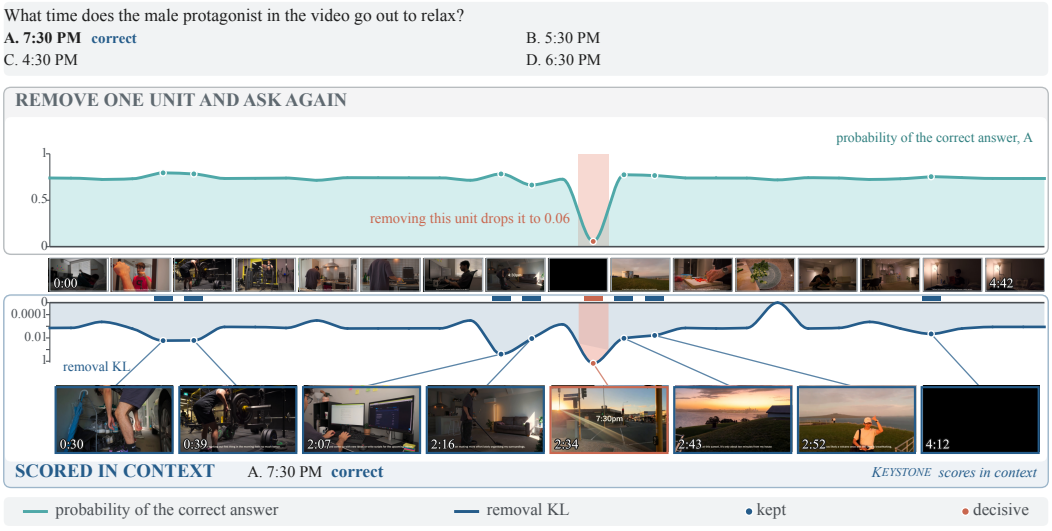

Figure 14: Removing one unit at a time, question 560-2 of Video-MME.

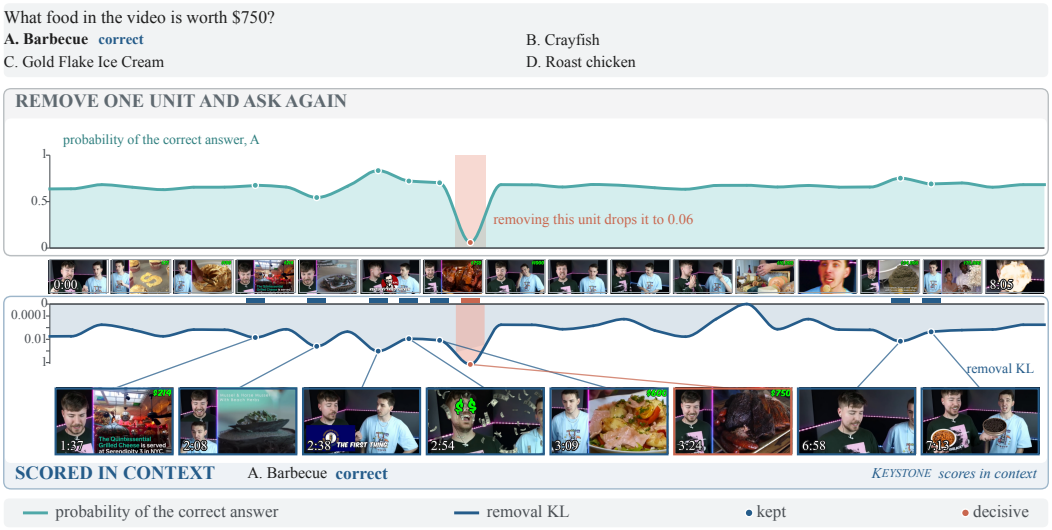

Figure 15: Removing one unit at a time, question 535-1 of Video-MME.

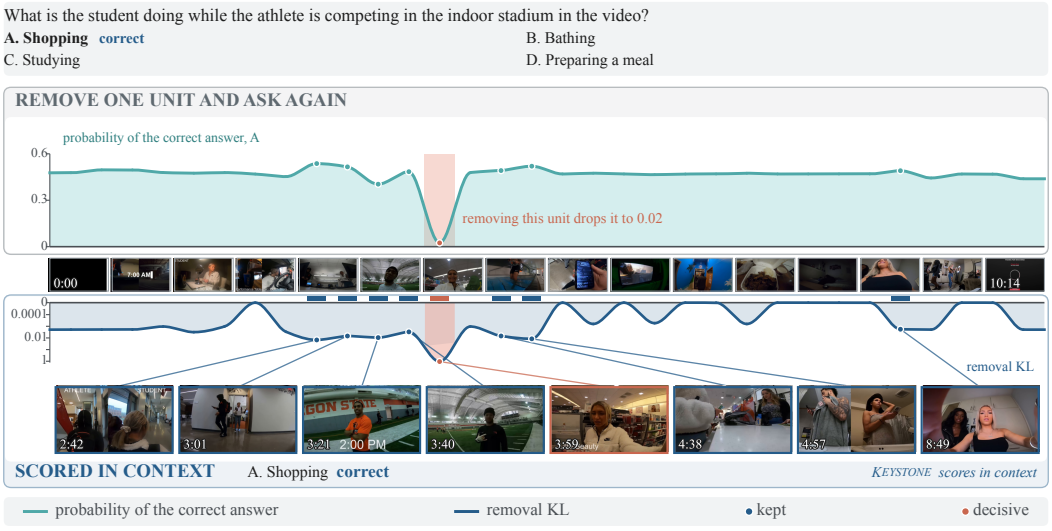

Figure 16: Removing one unit at a time, question 552-2 of Video-MME.

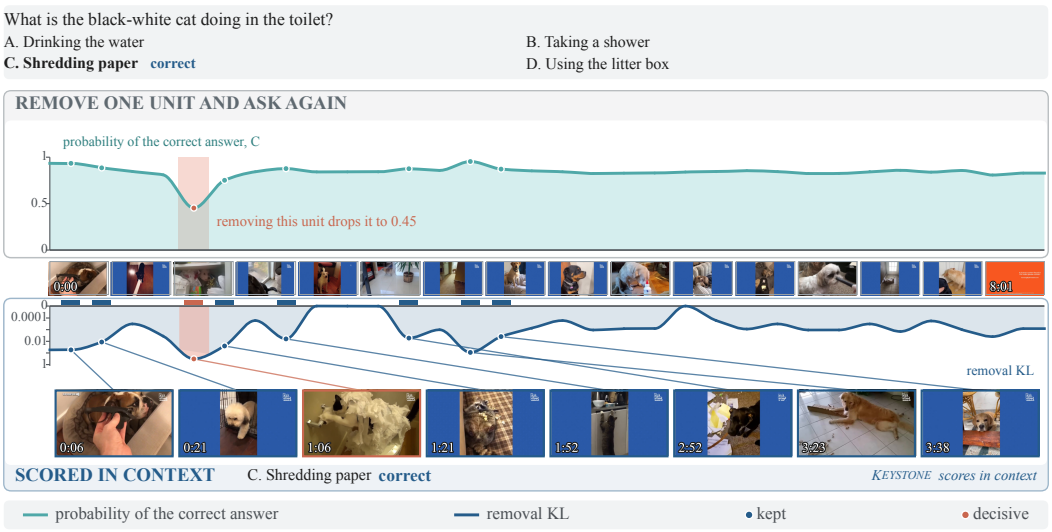

Figure 17: Removing one unit at a time, question 572-1 of Video-MME.

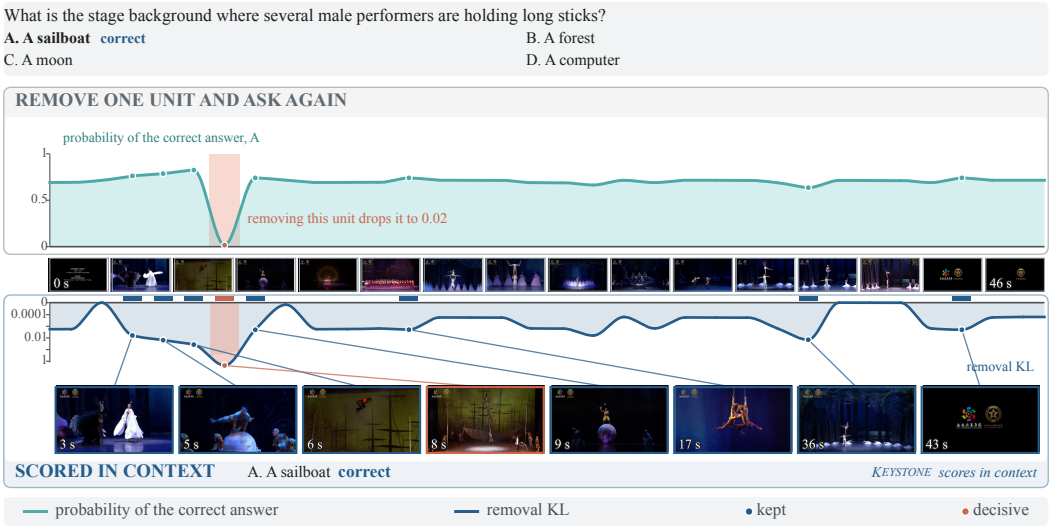

Figure 18: Removing one unit at a time, question 219-3 of Video-MME.

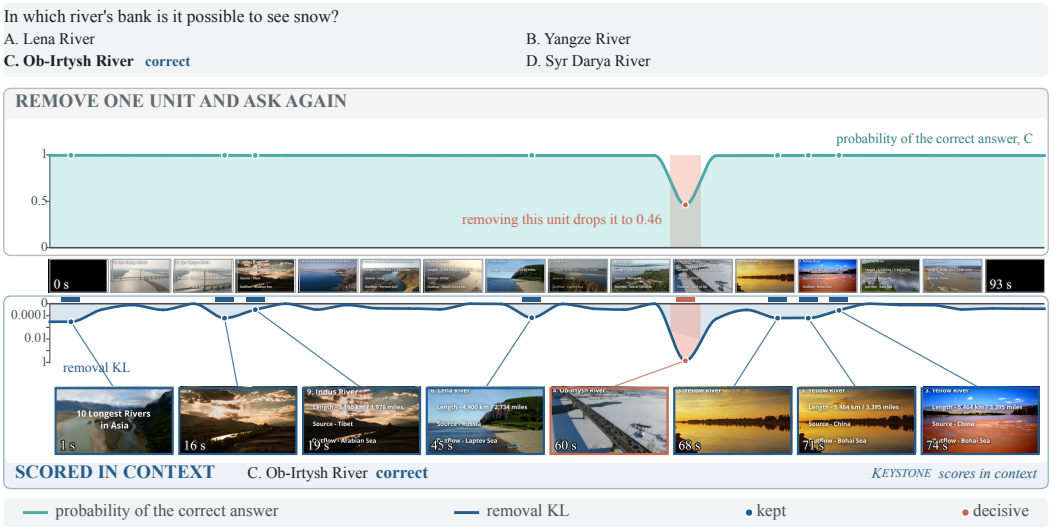

Figure 19: Removing one unit at a time, question 051-3 of Video-MME.

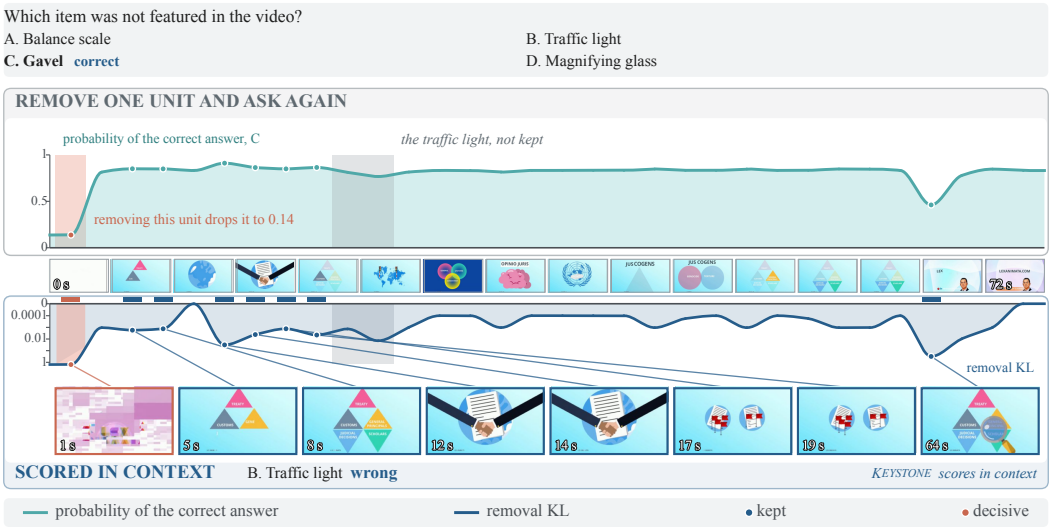

Figure 20: Failure case, question 061-2 of Video-MME. Gray marks the units that show the traffic light.

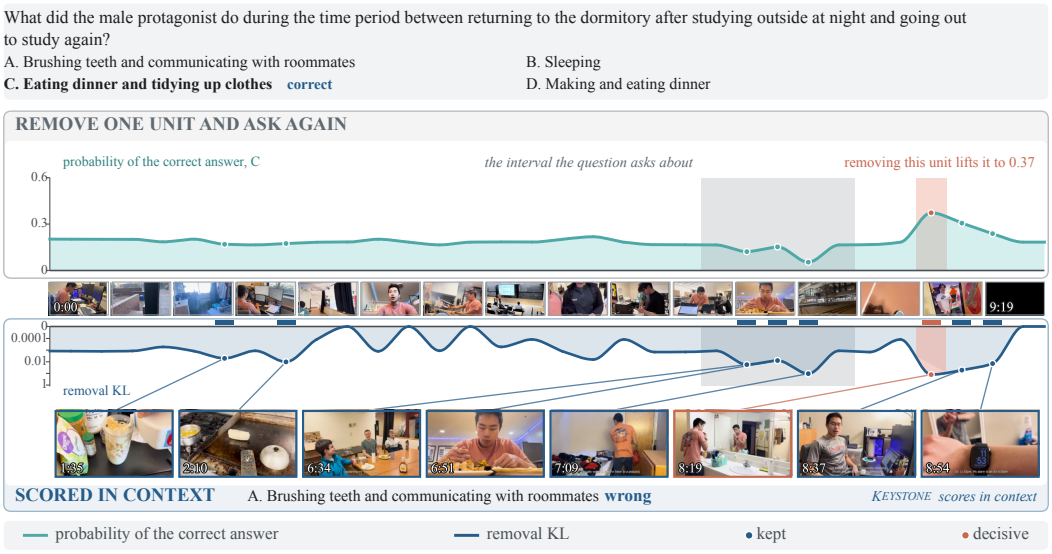

Figure 21: Failure case, question 557-2 of Video-MME. Gray marks the interval the question asks about.

G LIMITATIONS

KEYSTONE requires a finite-choice answer posterior. Its open-ended readout in Section E has been evaluated on only one benchmark of short videos and one answering model, and that benchmark leaves little room for frame selection. KEYSTONE uses a fixed frame budget, without per-question adaptation.