DES: DISCRIMINATIVE EVIDENCE SELECTION FOR LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Answering questions about long videos requires identifying relevant visual evidence distributed over time. With a limited frame budget, multiple-choice video question answering also requires evidence that distinguishes among candidate answers. However, question-relevant frames may provide little discriminative information, and several high-scoring frames may depict the same content. We introduce Discriminative Evidence Selection (DES), a training-free framework that estimates frame utility by combining semantic relevance, differences in frame–option matching, object and action cues, and local temporal evidence. DES constructs a candidate pool from local score maxima, temporal segment representatives, and globally high-scoring frames. It then selects a diverse subset using maximal marginal relevance, progressively relaxing the minimum temporal gap between selected frames as needed to fill the frame budget. The selected frames are provided in chronological order to a frozen video-language model. Experiments on four benchmarks with frame budgets of 8, 16, and 32 show that DES outperforms uniform sampling in most settings. With LLaVA-OneVision-7B and eight input frames, DES improves accuracy over BOLT by 2.0 and 3.3 percentage points on Video-MME and MLVU, respectively. Under the same eight-frame budget, evaluations on Video-MME without subtitles using five additional video-language models show accuracy gains of 3.2–6.1 percentage points over uniform sampling.

1 INTRODUCTION

Long-video question answering requires models to answer natural-language questions based on video content under a limited visual input budget. As an extension of image-based visual question answering (VQA) (Antol et al., 2015), video question answering (VideoQA) involves understanding objects, actions, and temporal relationships (Xiao et al., 2021). Recent video-language models combine visual encoders with large language models (LLMs) to support video understanding and question answering (Maaz et al., 2024; Li et al., 2025a). A common pipeline encodes sampled frames, projects their features into the LLM's embedding space, and processes them alongside the question and, for multiple-choice tasks, candidate answers. Context capacity and computational cost limit the number of frames that can be processed, making frame selection consequential for answering questions about long videos.

Existing research addresses this budget constraint through task-aware frame selection, using questions, candidate answers, or relationships among frames. Uniform sampling provides temporal coverage but does not adapt the input to the question (Xu et al., 2024). BOLT (Liu et al., 2025b) and Q-Frame (Zhang et al., 2025c) use question–frame matching to guide selection toward relevant content. Video–query–options similarity incorporates candidate answers into content matching (Wu et al., 2026), while MDP^3 considers query relevance, frame-set diversity, and sequential relationships (Sun et al., 2025). These approaches offer different ways to assess which visual content to include within a limited input budget.

For multiple-choice VideoQA, effective frame selection depends on both evidence that distinguishes among answer options and complementary information across the selected frames. Question relevance alone may not capture the former. In Figure 1, the Query-only method (Bolt(Liu et al., 2025b)) selects frames showing a mobile phone, yet the model predicts "A. Phone call." DES includes

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.

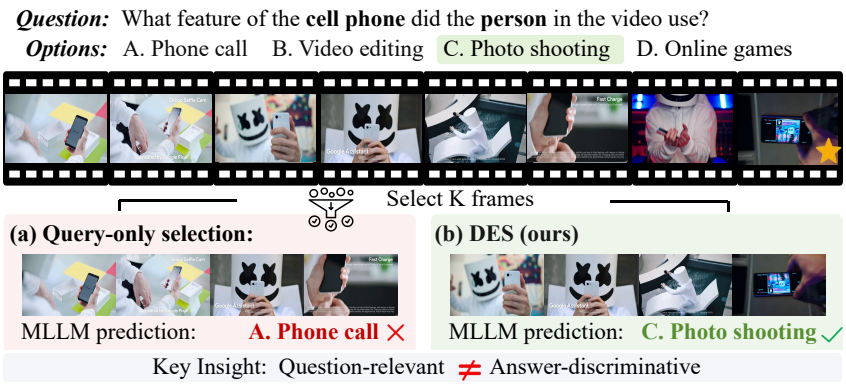

Figure 1: Question-relevant frames are not necessarily answer-discriminative. Frames selected by Query-only and DES, together with the corresponding downstream model predictions for the same video and question. The example is from the Video-MME test set (video ID: `3E4Zca1p64s`; question ID: `260-3`).

frames with camera-interface details, and the model instead predicts "C. Photo shooting." High-scoring frames may also contain redundant information, so their usefulness depends on how the final set is selected. In a controlled comparison on Video-MME (Fu et al., 2025) without subtitles, using LLaVA-OneVision-7B and eight frames, multi-cue scoring provides no accuracy gain over question-similarity scoring under direct top-K selection, but yields a gain when both use diversity-aware selection (Table 4). This observation motivates jointly considering frame utility estimation and set selection.

To address these two aspects, we propose DES (Discriminative Evidence Selection), a training-free framework that combines option-aware frame scoring with diversity-aware selection under a fixed frame budget. Its Multi-Evidence Utility Estimator (MEUE) uses relative differences in matching scores across options as a proxy for discriminative value, combining this signal with semantic relevance, object and action cues, and local temporal information. Diversity-Aware Selection (DAS) then constructs a candidate pool and applies Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) to balance frame utility against visual redundancy, together with adaptive temporal spacing. The selected frames are passed in chronological order to a frozen downstream model. We evaluate DES with LLaVA-OneVision-7B at budgets of 8, 16, and 32 frames on Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025), with additional evaluation on NExT-QA (Xiao et al., 2021). DES outperforms uniform sampling in most tested settings. On Video-MME without subtitles at eight frames, it also improves accuracy across five additional downstream models, with gains of 3.2–6.1 percentage points over uniform sampling.

Our contributions are threefold:

1. We introduce a training-free frame utility estimator for multiple-choice VideoQA that uses relative matching differences across answer options as a proxy for discriminative value, alongside semantic, visual, and local temporal cues.

2. We develop a fixed-budget selection pipeline that integrates multi-source candidate construction, MMR-based redundancy control, and adaptive temporal spacing to select frames using these utility estimates.

3. We evaluate DES on four VideoQA benchmarks, with additional cross-model evaluation on Video-MME. The results show gains over uniform sampling in most tested settings, while controlled comparisons examine the interaction between frame scoring and set selection.

2 Related Work

2.1 Video Large Language Models and Long Video Understanding

Long videos generate large numbers of visual tokens, straining context capacity and computational resources (Li et al., 2025b). Existing methods address this challenge through context extension or representation compression. LongVILA supports extended visual sequences (Chen et al., 2025), while LLaMA-VID and MovieChat reduce representation overhead through compact encoding and memory mechanisms, respectively (Li et al., 2024; Song et al., 2024). LongVU further reduces spatiotemporal redundancy using query information and inter-frame dependencies (Shen et al., 2025).

Beyond input capacity, long-video understanding requires identifying task-relevant content. Text–video retrieval uses semantic matching (Luo et al., 2022), question-driven localization identifies relevant segments (Yu et al., 2023), and interactive reasoning acquires information during inference (Wang et al., 2024). Query-aware frame selection directly allocates a limited frame budget according to question relevance (Liu et al., 2025b; Zhang et al., 2025c). Our work focuses on this input-selection stage, selecting a fixed number of frames for a frozen video QA model without additional selector training.

2.2 Frame Selection for Video Question Answering

Frame selection methods can be categorized by whether the selector requires training. Learning-based methods derive selection strategies from QA feedback, frame composition ranking, or pseudo-label supervision (Yu et al., 2023; 2025; Hu et al., 2025), even when the downstream QA model remains frozen. Training-free methods instead use pretrained cross-modal representations to rank or sample frames based on frame–text matching scores (Liu et al., 2025b; Zhang et al., 2025c).

In multiple-choice video question answering, candidate answers can indicate which objects, actions, or states must be distinguished. High QA accuracy, however, does not necessarily imply reliable evidence localization (Xiao et al., 2024). Video–question–option matching combines the question with each option and uses the highest clip-matching score to estimate relevance (Wu et al., 2026). Hypothesis verification instead treats candidate answers as propositions to test (Liu et al., 2025a). VideoHV-Agent (Wang et al., 2026) and PACE (Xu et al., 2026) use discriminative or contrastive cues to guide local content retrieval and verification.

The quality of a selected frame set also depends on its redundancy. BOLT constructs a sampling distribution based on question relevance (Liu et al., 2025b). MDP^3 jointly models question relevance and frame-set diversity through a determinantal point process (Sun et al., 2025), while GIFT assesses frame irreplaceability through inter-frame differences and preserves relevance (Ma et al., 2026).

Our method uses relative differences in matching scores across options within each frame as a proxy for discriminative value. It combines this signal with question relevance, background, action, and visual-change cues to score frames and construct a candidate pool. Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) then selects frames iteratively, balancing their utility scores against visual similarity to previously selected frames. This training-free framework couples multi-cue utility estimation with explicit redundancy control under a fixed frame budget.

3 Method

We propose DES, a training-free framework for fixed-budget frame selection in multiple-choice video question answering. As shown in Figure 2, the Multi-Evidence Utility Estimator (MEUE) combines semantic, option-specific, visual, and temporal cues to estimate frame utility scores. Diversity-Aware Selection (DAS) then balances utility against redundancy to select complementary frames for a frozen downstream model.

3.1 Task Definition and Overview

Given a video $V = (f_1, \ldots, f_T)$, a question Q, and N candidate answers $\mathcal{O} = \{o_n\}_{n=1}^{N}$, we select $K \ll T$ frames, where K is the frame budget. Let $\mathbf{X} = (\mathbf{v}_1, \ldots, \mathbf{v}_T)$ denote precomputed visual

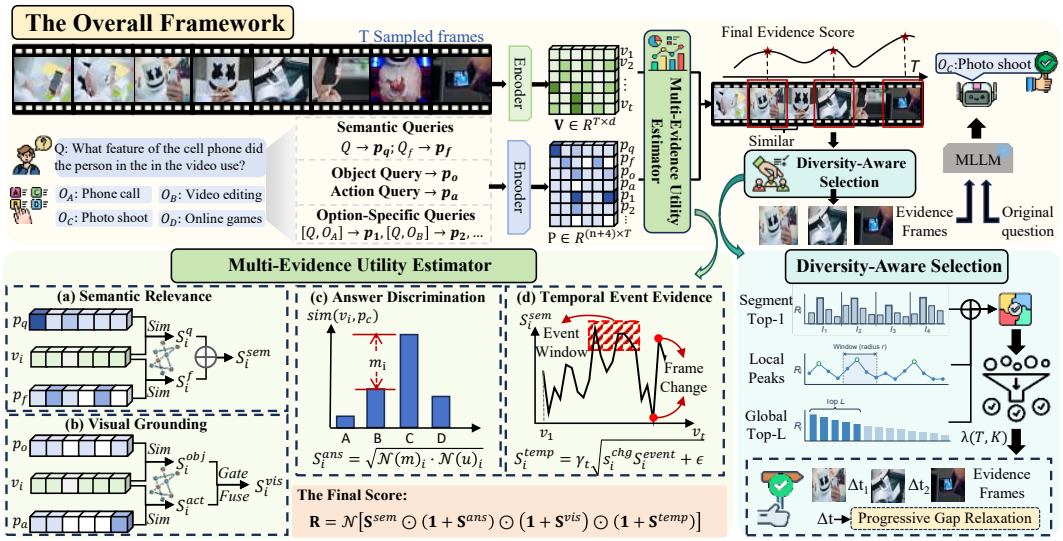

Figure 2: Overview of DES. MEUE integrates (a) semantic relevance, (b) object and action grounding, (c) answer discrimination, and (d) temporal event evidence. DAS combines local peaks, temporal segment representatives, and globally high-scoring frames, then applies MMR with progressively relaxed temporal spacing. The selected frames are passed in chronological order to a frozen video-language model.

features, with $\mathbf{v}_i$ corresponding to f_i. The selector returns

$$\mathcal{I} = \mathrm{Select}(\mathbf{X}, Q, \mathcal{O}; K), \qquad \mathcal{I} \subseteq \{1, \dots, T\}, \quad |\mathcal{I}| = K. \tag{1}$$

We order the selected indices as $s_1 < \cdots < s_K$ and predict the answer using

$$\hat{A} = \mathrm{VLM}_\theta\big((f_{s_1}, \dots, f_{s_K}), Q, \mathcal{O}\big), \tag{2}$$

where θ remains frozen. MEUE estimates per-frame utility (Section 3.2), and DAS constructs the final frame set (Section 3.3); neither requires additional training.

3.2 Multi-Evidence Utility Estimation

Question relevance alone may not identify frames that distinguish among candidate answers. MEUE therefore combines four complementary cues for all T frames, assuming $N \geq 2$. Let $\mathbf{S}^x = (S_1^x, \dots, S_T^x)$ denote the score vector for cue x. We use $\mathcal{N}(\cdot)$ to denote robust normalization across frames and ϵ for numerical stability; normalization details are provided in Appendix B.1.

Semantic relevance. We match each frame against the question Q and the full context $Q_f = \mathrm{Concat}(Q, \mathcal{O})$ (Figure 2(a)). Using pretrained frame–text matching scores $\mathrm{sim}(\mathbf{v}_i, t)$, we obtain

$$\begin{aligned} \mathbf{S}^q &= \mathcal{N}\left(\big(\mathrm{sim}(\mathbf{v}_i, Q)\big)_{i=1}^{T}\right), \\ \mathbf{S}^f &= \mathcal{N}\left(\big(\mathrm{sim}(\mathbf{v}_i, Q_f)\big)_{i=1}^{T}\right), \\ S_i^{\mathrm{sem}} &= \sqrt{S_i^q S_i^f + \epsilon}. \end{aligned} \tag{3}$$

The geometric mean favors frames relevant to both text views.

Option discriminability. For each option, we construct $t_n = \mathrm{Concat}(Q, o_n)$ and compute $E_{n,i} = \mathrm{sim}(\mathbf{v}_i, t_n)$, yielding $E \in \mathbb{R}^{N \times T}$. Applying a shared normalization to the entire matrix produces $\widehat{E}$, preserving a common scale across options.

For frame i, let $\mathbf{p}_i = \mathrm{Softmax}(\widehat{E}_{:,i})$. We combine the margin between the top two matching scores with distribution concentration (Figure 2(c)):

$$\begin{aligned} m_i &= \widehat{E}_{(1),i} - \widehat{E}_{(2),i}, \\ u_i &= 1 - \frac{\mathcal{H}(\mathbf{p}_i)}{\log N}, \\ S_i^{\mathrm{ans}} &= \sqrt{[\mathcal{N}(\mathbf{m})]_i [\mathcal{N}(\mathbf{u})]_i + \epsilon}. \end{aligned} \tag{4}$$

Here, $\widehat{E}_{(1),i}$ and $\widehat{E}_{(2),i}$ are the largest and second-largest entries in column i, respectively. $\mathcal{H}$ denotes Shannon entropy, and $\mathbf{m}$ and $\mathbf{u}$ collect these statistics across frames. The margin measures the separation between the two strongest matches, while concentration accounts for all options. Both serve as proxies for discriminative value rather than answer correctness.

Visual and temporal cues. Fixed templates augment Q_f with object and action views, yielding normalized scores $\mathbf{S}^{\mathrm{obj}}$ and $\mathbf{S}^{\mathrm{act}}$. Their weighted average forms $\mathbf{S}^{\mathrm{vis}}$, with soft weights determined by question keywords (Figure 2(b)).

Temporal evidence combines visual changes between adjacent frames, action matching, and semantic relevance into a question-relevant response. We aggregate this response over a local window, combine it with the current frame's change score, and adjust its weight using temporal keywords to obtain $\mathbf{S}^{\mathrm{temp}}$ (Figure 2(d)). Templates, keyword weights, and the full temporal formulation are provided in Appendices B.2–B.4.

Multi-evidence fusion. The auxiliary cues reweight semantic relevance to produce frame utility:

$$\mathbf{R} = \mathcal{N}\left[\mathbf{S}^{\mathrm{sem}} \odot (\mathbf{1} + \mathbf{S}^{\mathrm{ans}}) \odot (\mathbf{1} + \mathbf{S}^{\mathrm{vis}}) \odot (\mathbf{1} + \mathbf{S}^{\mathrm{temp}})\right], \tag{5}$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1}$ is the length-T all-ones vector. The resulting scores guide both candidate construction and frame selection.

3.3 Diversity-Aware Selection

High-utility frames may contain redundant information. DAS constructs a candidate pool and iteratively balances utility against visual redundancy using Maximal Marginal Relevance (MMR), as shown in the right-hand module of Figure 2.

Candidate pool. We merge three sources: *Local Peaks*, retaining utility maxima within temporal neighborhoods of radius r; *Segment Top-1*, selecting the highest-utility frame from each of M approximately equal contiguous segments; and *Global Top-L*, retaining the L highest-utility frames overall:

$$\mathcal{C} = \mathcal{C}_{\mathrm{peak}} \cup \mathcal{C}_{\mathrm{segment}} \cup \mathcal{C}_{\mathrm{global}}. \tag{6}$$

Taking the union removes duplicates while combining local prominence, temporal coverage, and global priority. Specific parameter configurations are detailed in the Implementation Details.

Utility–redundancy trade-off. Let $\mathcal{Y}$ denote the selected index set and $\hat{\mathbf{v}}_i$ the normalized visual feature of frame i. For nonempty $\mathcal{Y}$, define

$$d(i, \mathcal{Y}) = \max\left\{0, \max_{j \in \mathcal{Y}} \hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_j\right\}, \tag{7}$$

with $d(i, \varnothing) = 0$. The maximum similarity penalizes repeated content, and clipping at zero prevents negative similarities from yielding an additional reward. At each step, we select

$$i^\star = \arg\max_{i \in \mathcal{A}_g(\mathcal{Y})} \left[\lambda R_i - (1-\lambda) d(i, \mathcal{Y})\right], \qquad \lambda = \frac{K}{K + \log(T+1)}. \tag{8}$$

The weight λ remains fixed during selection. For fixed T, smaller budgets place greater relative weight on redundancy reduction.

Table 1: Accuracy (%) across four video question-answering benchmarks. † marks our reproduced Uniform and BOLT results. – denotes an unreported result.

Model / selector	LLM size	Training-free	Video-MME (no subtitles)				LongVideoBench	MLVU	NExT-QA
			Overall	Short	Medium	Long			
Video-LLaVA(Lin et al., 2024)	7B	✘	39.9	45.3	38.0	36.2	39.1	47.3	–
VideoChat2(Li et al., 2025b)	7B	✘	39.5	48.3	37.0	33.2	39.3	44.5	–
ShareGPT4Video(Chen et al., 2024)	8B	✘	39.9	48.3	36.3	35.0	41.8	46.4	–
Chat-UniVi-V1.5(Jin et al., 2024)	7B	✘	40.6	45.7	40.3	35.8	–	–	–
VideoLLaMA2(Cheng et al., 2024)	7B	✘	47.9	56.0	45.4	42.1	–	–	–
Frame-Voyager(Yu et al., 2025)	8B	✘	57.5	67.3	56.3	48.9	–	65.6	73.9
LongVU(Shen et al., 2025)	7B	✘	60.9	64.7	58.2	59.5	–	65.4	–
8 frames									
Uniform†	7B	✔	53.9	65.0	52.0	44.7	54.2	58.5	77.4
BOLT(Liu et al., 2025b)†	7B	✔	56.1	68.0	53.4	46.8	54.1	63.9	77.4
DES	7B	✔	**58.1**	**68.4**	**57.0**	**48.8**	**55.7**	**67.2**	**77.6**
16 frames									
Uniform†	7B	✔	56.9	68.4	53.8	48.2	55.9	61.4	78.1
BOLT(Liu et al., 2025b)†	7B	✔	58.5	68.3	**58.0**	49.2	56.3	65.8	**78.3**
DES	7B	✔	**59.5**	**70.7**	57.9	**50.0**	**58.2**	**68.8**	**78.3**
32 frames									
Uniform†	7B	✔	58.4	**70.1**	56.4	48.6	56.7	63.0	**79.5**
BOLT(Liu et al., 2025b)†	7B	✔	59.4	69.2	**59.3**	49.8	58.0	66.6	78.7
DES	7B	✔	**60.1**	**70.8**	59.2	**50.2**	**59.5**	**68.3**	78.8

Table 2: Video-MME accuracy (%) across five downstream models with eight input frames and no subtitles.

Model	Size	Method	Overall	Short	Medium	Long
Video-LLaVA	7B	Uniform	37.6	42.7	37.1	33.0
		DES	**40.8**	**48.1**	**39.4**	**34.8**
LongVA	7B	Uniform	48.5	56.7	46.6	42.3
		DES	**52.4**	**62.1**	**51.0**	**44.2**
LLaVA-Video	7B	Uniform	56.0	67.8	53.6	46.7
		DES	**60.6**	**72.3**	**57.3**	**52.0**
Qwen2.5-VL-7B Instruct	7B	Uniform	53.8	62.0	52.3	47.1
		DES	**59.9**	**69.2**	**57.4**	**53.1**
Qwen3-VL-8B Instruct	8B	Uniform	56.0	64.2	53.6	50.1
		DES	**62.0**	**72.7**	**58.4**	**54.8**

Temporal spacing and iterative selection. To limit local temporal concentration, eligible candidates satisfy

$$\mathcal{A}_g(\mathcal{Y}) = \{i \in \mathcal{C} \setminus \mathcal{Y} : |i - j| \geq g,\ \forall j \in \mathcal{Y}\}\,, \tag{9}$$

where g is the minimum frame-index gap. At each iteration, we add $i^\star$ to $\mathcal{Y}$ and update the redundancy scores for the remaining candidates. If no eligible candidate remains before the budget is filled, we retain $\mathcal{Y}$ and progressively relax g, reducing it to zero if necessary. Once K frames have been selected, we set $\mathcal{I} = \mathcal{Y}$ and arrange the downstream input as described in Section 3.1.

4 Experiments

We evaluate DES on four video question-answering benchmarks across different frame budgets and downstream models. We examine the contributions of its components through ablations, parameter analyses, and frame-selection examples.

4.1 Experimental Setup

Models and baselines. Our main experiments use LLaVA-OneVision-7B Li et al. (2025a) and compare DES with uniform sampling and BOLT Liu et al. (2025b) at budgets of 8, 16, and 32 frames. To assess transfer across downstream models, we also evaluate Video-LLaVA Lin et al. (2024), LongVA Zhang et al. (2025b), LLaVA-Video Zhang et al. (2025d), Qwen2.5-VL-7B-Instruct Bai et al. (2025), and Qwen3-VL-8B-Instruct Team (2025) on Video-MME with eight frames. All downstream models remain frozen; no additional training is performed.

Benchmarks and metrics. We evaluate on Video-MME Fu et al. (2025), LongVideoBench Wu et al. (2024), MLVU Zhou et al. (2025), and NExT-QA Xiao et al. (2021) without subtitle input. Video-MME contains 900 videos and 2,700 multiple-choice questions; we report overall accuracy and results for its short, medium, and long video subsets. We use the LongVideoBench validation set (1,337 questions), the seven multiple-choice tasks in the MLVU Dev set (2,174 questions, excluding generation tasks), and the NExT-QA multiple-choice test split (8,564 questions). MLVU scores are computed as the unweighted mean of the accuracies across its seven tasks; for the other benchmarks, we report multiple-choice accuracy. All evaluations use lmms-eval 0.5.0 Zhang et al. (2025a). Details of the checkpoints and prompts are provided in Appendix C.

Implementation details. We sample an initial candidate sequence at 1 FPS and use pretrained CLIP-L/14 Radford et al. (2021) to extract visual and text embeddings and compute frame–text matching scores. DES selects K frames from this sequence based on utility and visual redundancy. Except in the temporal-spacing sensitivity analysis, the initial minimum spacing is 2 seconds and is progressively relaxed as needed to fill the frame budget. Experiments are run on two RTX 6000D GPUs.

4.2 Benchmark Results

Comparison with frame selection baselines. DES outperforms uniform sampling on Video-MME, LongVideoBench, and MLVU at all three frame budgets. Across the four benchmarks, it exceeds BOLT in 11 of the 12 overall comparisons and ties on NExT-QA with 16 frames. With eight frames, its gains on Video-MME, LongVideoBench, and MLVU are 4.2, 1.5, and 9.3 percentage points over uniform sampling, and 2.0, 1.6, and 3.9 points over BOLT, respectively. On Video-MME Medium, DES scores 0.1 points below BOLT with both 16 and 32 frames. On the Long subset, it exceeds BOLT by 0.4 points with 32 frames. These results support DES under fixed frame budgets, although its gains vary across duration subsets.

Performance under limited frame budgets. DES can outperform uniform sampling with fewer input frames. Compared with 32-frame uniform sampling, 16-frame DES improves accuracy by 1.1, 1.5, and 5.6 points on Video-MME, LongVideoBench, and MLVU, respectively. On MLVU, even eight-frame DES exceeds 32-frame uniform sampling by 4.8 points. The benefit depends on the benchmark: with 32 frames, DES scores 0.7 points below uniform sampling on NExT-QA. These comparisons show that DES can achieve higher accuracy with a smaller downstream frame budget on the first three benchmarks.

Results across downstream models. DES improves overall Video-MME accuracy for all five additional downstream models with eight frames and no subtitles. The gains are 3.2 points for Video-LLaVA, 3.9 for LongVA, 4.6 for LLaVA-Video, 6.1 for Qwen2.5-VL, and 6.0 for Qwen3-VL, with all other evaluation settings fixed within each model. These consistent improvements show that the benefits of DES extend beyond LLaVA-OneVision to the five additional models tested. Additional comparisons across models and frame budgets, along with detailed benchmark breakdowns, are provided in Appendix A.

4.3 Ablation Experiments and Analysis

We use LLaVA-OneVision-7B to examine the contributions of the scoring factors in DES, compare frame selection strategies, and assess sensitivity to the initial temporal gap.

Table 3: Scoring-cue ablation on Video-MME. Sem., Ans., Vis., and Temp. denote semantic relevance, answer discrimination, visual grounding, and temporal event evidence, respectively. A dash indicates that only the corresponding factor is removed.

Sem.	Ans.	Vis.	Temp.	Acc. (%)	Acc. drop
✓	✓	✓	✓	**58.1**	–
✓	✓	✓	–	57.7	0.4
✓	✓	–	✓	57.1	1.0
✓	–	✓	✓	57.0	1.1
✓	–	–	–	56.4	1.7

Ablation of scoring cues. Table 3 evaluates the scoring cues on Video-MME with eight input frames. Removing any auxiliary factor reduces accuracy, with larger drops when answer discrimination or visual grounding is removed than when temporal event evidence is removed. Using semantic relevance alone lowers accuracy by 1.7 percentage points. These results show that combining auxiliary scoring cues with semantic relevance improves downstream question-answering accuracy in this setting.

Comparison with alternative frame selection strategies. Table 5 compares 16-frame selection strategies on Video-MME alongside a no-video baseline. Top-K, ITS, and DAS use the same MEUE scores. DAS exceeds Top-K by 2.19 percentage points and ITS by 0.30 points. Uniform + Shuffle and Random achieve lower accuracy than Uniform by 0.60 and 1.70 points, respectively. These results support the full DAS selection pipeline over direct score ranking and indicate that temporal order also matters for downstream question answering.

Ablation of frame scoring and selection. Table 4 compares query-only scoring and MEUE, each paired with Top-K or DAS, on Video-MME without subtitles using LLaVA-OneVision-7B and eight frames. DAS improves accuracy over Top-K under both scoring functions, with gains of 1.04 and 2.91 percentage points for query-only scoring and MEUE, respectively. Replacing query-only scoring with MEUE slightly lowers accuracy under Top-K (55.19% vs. 55.33%) but improves it by 1.73 points under DAS. The complete DES framework achieves 58.10% accuracy, exceeding query-only scoring with Top-K by 2.77 points. These results support the complementary roles of MEUE and DAS in DES: multi-evidence scoring improves accuracy when paired with diversity-aware selection, demonstrating the effectiveness of jointly designing frame scoring and selection under the tested frame budget.

Table 4: Frame scoring and selection on Video-MME without subtitles

Scoring	Selection	Acc. (%)	DAS gain (pp)
Query-only	Top-K	55.33	–
Query-only	DAS	56.37	+1.04
MEUE	Top-K	55.19	–
MEUE	DAS	**58.10**	+2.91

Initial temporal gap. Figure 3 shows the effect of varying the initial temporal gap from 1 to 5 seconds with eight input frames, while keeping all other settings and the gap-relaxation procedure fixed. Accuracy varies over ranges of 0.9, 0.7, and 0.8 percentage points on Video-MME, LongVideoBench, and MLVU, respectively. Larger gaps do not consistently improve results, and the default $g = 2$ seconds remains within 0.5 points of each benchmark's best observed accuracy. These results indicate limited sensitivity to the initial gap within the tested range and support the default setting across the three benchmarks.

4.4 Visualization of Selected Frames

Figure 4 compares Query-only and DES under the same eight-frame budget, using LLaVA-OneVision as the downstream model. In the first case, DES retains ranking cues that support the

Table 5: Comparison of training-free frame selection strategies on Video-MME.

Frame selection	MEUE scores	Acc. (%)
Blind Test*	–	40.85
Uniform*	×	56.85
Uniform + Shuffle*	×	56.25 (−0.60)
Random*	×	55.15 (−1.70)
Top-K	✓	57.22 (+0.37)
ITS	✓	59.11 (+2.26)
DAS (DES)	✓	**59.41 (+2.56)**

Note. * marks our reproduced baselines. Top-K, ITS, and DAS use the same MEUE scores. Parentheses report changes relative to Uniform in percentage points. The highest accuracy is shown in bold.

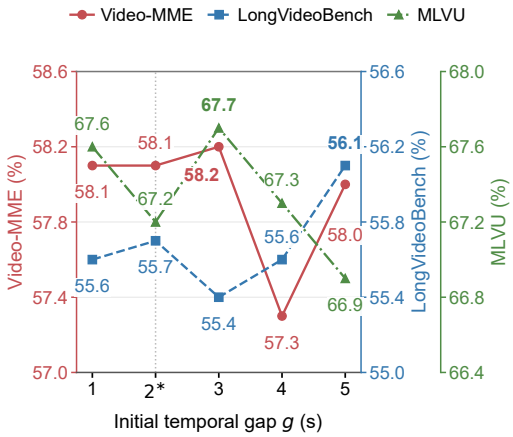

Figure 3: Sensitivity to the initial temporal gap g using LLaVA-OneVision-7B with eight input frames.

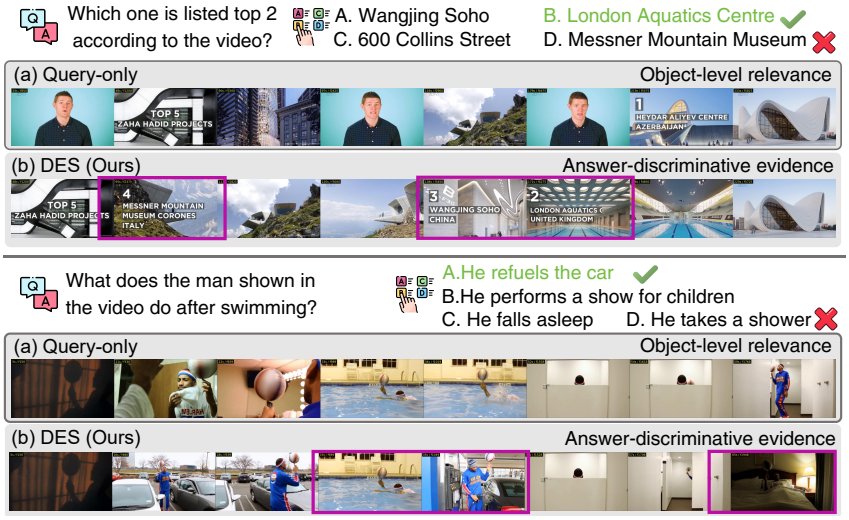

Figure 4: Qualitative comparison of frame selection. Compared with Query-only, DES captures answer-discriminative ranking evidence that helps the model rule out distractors and identify the correct answer.

correct answer and rule out distractors. In the second, it captures swimming and subsequent candidate actions, providing temporal context for distinguishing among the options. These examples illustrate how DES retains answer-discriminative evidence beyond question-relevant content.

5 Conclusion

We introduced DES, which combines multi-cue frame utility estimation with diversity-aware selection to select visual inputs for frozen video-language models under a fixed frame budget. Across benchmarks and downstream models, DES generally improves question-answering accuracy. Our ablations show that better frame scores alone are insufficient: multi-cue scoring does not improve results with direct Top-K selection but is more effective when paired with diversity-aware selection. Gains vary across datasets and budgets, and some 32-frame settings do not outperform the baseline. Future work will examine this variation and measure the inference overhead of frame selection.

AI USE STATEMENT

In this work, we used generative AI tools to write execution scripts, check code formatting, and analyze code errors. We did not use generative AI tools for core research conceptualization, experimental data generation, or figure and chart creation. Crowdsourced data collection and human subject studies are not applicable to this work. We also used generative AI tools to refine English grammar, rewrite sentences, and assist with LaTeX table formatting. We have reviewed all AI-assisted work. Specifically, all AI-polished text and formatted tables underwent line-by-line semantic proofreading to ensure scientific accuracy and fidelity to our original intent, and all AI-assisted scripts were executed locally and tested for logical correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We document our methodology and empirical evaluation in the manuscript and supplementary materials to support reproducibility. The complete source code, including video frame selection algorithms, execution scripts, and evaluation pipelines, is provided as an anonymous downloadable file in the supplementary materials and will be fully open-sourced upon acceptance. All dataset descriptions, data preprocessing steps, hyper-parameter settings, hardware specifications, and software dependencies are thoroughly documented in the paper and appendix. All baseline comparisons strictly follow standardized benchmark evaluation protocols and use publicly available pre-trained checkpoints.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433, 2015.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. *CoRR*, abs/2502.13923, 2025.

Jaime G. Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (eds.), *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pp. 335–336, 1998.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

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.

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, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118, 2025.

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, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13702–13712, 2025.

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, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13700–13710, 2024.

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. *Trans. Mach. Learn. Res.*, 2025, 2025a.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: chat-centric video understanding, 2025b.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, pp. 323–340, 2024.

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 Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5971–5984, 2024.

Huabin Liu, Filip Ilievski, and Cees G. M. Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 3262–3271, 2025a.

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, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 3318–3327, 2025b.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

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, 2026.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602, 2024.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18221–18232, 2024.

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, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 24090–24101, 2025.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXX*, volume 15138 of *Lecture Notes in Computer Science*, pp. 58–76, 2024.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 33784–33793, 2026.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. A training-free framework for long video understanding via video-query-options similarity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9777–9786, 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13204–13214, 2024.

Baixuan Xu, Yinyui Xu, Tianshi Zheng, Zhaowei Wang, Weiqi Wang, Haochen Shi, Jiayu Liu, Qing Zong, Xiyu Ren, Xinyu Geng, Zhitao He, and Yangqiu Song. Finding the right evidence: Factor-guided coarse-to-fine reasoning for long videos, 2026.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See-Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning, 2024.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

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 Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 881–916, 2025a.

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

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22056–22065, 2025c.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025d.

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, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13691–13701, 2025.

A ADDITIONAL QUANTITATIVE COMPARISONS

A.1 COMPARISONS ACROSS MODELS AND FRAME BUDGETS

Results across downstream models. Table 6 compares Uniform, BOLT, and DES on Qwen3-VL-8B and Qwen2.5-VL-7B across three benchmarks and three frame budgets. DES outperforms uniform sampling in all 18 evaluated configurations, with accuracy gains ranging from 1.58 to 10.64 percentage points. The gains are particularly large on MLVU, ranging from 5.95 to 10.64 points across the two models and three budgets. These results show that the improvements over uniform sampling extend across both tested models and persist as the frame budget increases.

Effect of frame budget. DES accuracy increases as the frame budget rises from 8 to 32 frames on all three benchmarks for both models. On Video-MME, however, its advantage over uniform sampling narrows: the gains decrease from 5.96 to 1.59 points for Qwen3-VL-8B and from 6.13 to 1.58 points for Qwen2.5-VL-7B. This pattern suggests that frame selection is especially useful under the smaller input budget on this benchmark. On MLVU, DES with eight frames exceeds uniform sampling with 32 frames by 2.14 points on Qwen3-VL-8B and 1.91 points on Qwen2.5-VL-7B. In these two comparisons, DES achieves higher accuracy using one quarter of the downstream input frames.

Comparison with BOLT. DES achieves higher accuracy than BOLT in 16 of the 18 configurations. The exceptions are LongVideoBench with Qwen3-VL-8B at eight frames, where DES scores 0.71 points lower, and Video-MME with Qwen2.5-VL-7B at 32 frames, where it scores 0.22 points lower. The margins also vary substantially: on Qwen3-VL-8B, DES improves Video-MME accuracy by 3.06 points with eight frames, whereas its gain on Qwen2.5-VL-7B is only 0.05 points on MLVU with 32 frames. These comparisons therefore support improvements over BOLT in most evaluated settings, with the gains depending on the model, benchmark, and frame budget.

A.2 COMPARISON BY DURATION AND QUESTION TYPE ON LONGVIDEOBENCH

We evaluate LLaVA-OneVision-7B on the LongVideoBench validation set (1337 questions) using three frame selectors at three frame budgets. Uniform samples frames evenly over time; BOLT uses CLIP-L inverse-transform sampling; DES is our query-conditioned selector. Table 7 reports accuracy by duration group and L1/L2 question type, along with Overall accuracy. Each row corresponds to a subset; within each frame budget, the highest accuracy among Uniform, BOLT, and DES is shown in bold red, including ties. DES achieves the highest Overall accuracy at every frame budget: 55.7 / 58.2 / 59.5 with 8/16/32 frames, compared with 54.2 / 55.9 / 56.7 for Uniform and 54.1 / 56.3 / 58.0 for BOLT. The gain over Uniform increases with the frame budget ($+1.5$, $+2.3$, $+2.8$). All methods improve as more frames are provided, but query-based selection benefits more: from 8 to 32 frames, Uniform's accuracy increases by only $+2.5$, while DES's increases by $+3.8$. BOLT performs slightly worse than Uniform at 8 frames (-0.1) and outperforms it only from 16 frames onward. The overall gains are concentrated in long videos. For videos in $(180, 3600]$ (976 questions), DES achieves accuracies of 51.1 / 53.9 / 56.7, corresponding to gains of $+1.8$ / $+3.9$ / $+5.3$ over Uniform; BOLT follows the same trend but remains lower (50.5 / 52.7 / 55.0). For short videos in $(8, 60]$ (361 questions), Uniform achieves the highest accuracy at 16/32 frames (71.7 / 70.9), BOLT shows the largest drop (66.2 / 65.9), and DES falls between them (69.8 / 67.0). Thus, similarity-based selection reallocates frames toward the queried moment: it benefits hour-scale videos but reduces coverage when the video is already short. By question type, DES's gains are concentrated mainly in L1 Perception (65.9 / 68.3 / 68.0 vs. Uniform 61.0 / 62.9 / 64.2). L2 Relation remains more challenging for all methods. DES does not outperform Uniform on L2 at 8 or 16 frames (46.8 / 49.3 vs. 48.3 / 49.7) and takes the lead only at 32 frames (52.0 vs. 50.1). SSS has the lowest accuracy in all nine runs (21.6–34.0), indicating that sparse selected frames still fail to capture global scene order.

Conclusion. DES is a more effective inference-time frame selector than both Uniform and BOLT on the LongVideoBench validation set. It improves Overall accuracy at 8/16/32 frames, with gains that increase as the frame budget grows. The gains come from long videos and L1 referring-

Table 6: Accuracy and gains over uniform sampling on three video question-answering benchmarks using Qwen3-VL-8B and Qwen2.5-VL-7B with 8, 16, and 32 input frames. Acc. denotes accuracy (%), and Δ denotes the gain over Uniform in percentage points for the same model, frame budget, and benchmark. The highest accuracy among the three selectors is shown in bold red within each comparison group. Video-MME uses no subtitles, and MLVU reports mean accuracy across seven tasks.

Frames	Selector	Video-MME (no subtitles)		LongVideoBench		MLVU	
		Acc. (%)	Δ (pp)	Acc. (%)	Δ (pp)	Acc. (%)	Δ (pp)
Qwen3-VL-8B							
8	Uniform	56.00	–	54.80	–	55.30	–
	BOLT	58.90	+2.90	**57.40**	+2.60	62.90	+7.60
	DES	**61.96**	+5.96	56.69	+1.89	**65.94**	+10.64
16	Uniform	60.50	–	57.70	–	58.90	–
	BOLT	62.70	+2.20	58.90	+1.20	67.70	+8.80
	DES	**64.26**	+3.76	**61.18**	+3.48	**67.94**	+9.04
32	Uniform	64.30	–	60.40	–	63.80	–
	BOLT	65.70	+1.40	62.20	+1.80	70.30	+6.50
	DES	**65.89**	+1.59	**62.75**	+2.35	**71.00**	+7.20
Qwen2.5-VL-7B							
8	Uniform	53.80	–	53.20	–	57.10	–
	BOLT	57.40	+3.60	55.10	+1.90	63.10	+6.00
	DES	**59.93**	+6.13	**56.17**	+2.97	**64.41**	+7.31
16	Uniform	58.80	–	56.10	–	59.90	–
	BOLT	60.90	+2.10	57.50	+1.40	66.20	+6.30
	DES	**61.74**	+2.94	**58.27**	+2.17	**67.24**	+7.34
32	Uniform	62.20	–	58.60	–	62.50	–
	BOLT	**64.00**	+1.80	60.00	+1.40	68.40	+5.90
	DES	63.78	+1.58	**60.21**	+1.61	**68.45**	+5.95

perception questions. Two limitations remain: lower accuracy than Uniform on short videos at 16/32 frames, and weak L2 relational reasoning—especially SSS—until the budget reaches 32 frames.

A.3 Task-wise Comparison on MLVU

Overall comparison. Table 8 compares Uniform, BOLT, and DES on the MLVU dev set using LLaVA-OneVision-7B. DES achieves the highest M-Avg at all three frame budgets, reaching 67.8%, 68.6%, and 68.6% with 8, 16, and 32 frames, respectively. The corresponding gains are 9.3, 7.2, and 5.6 percentage points over Uniform, and 3.9, 2.8, and 2.0 points over BOLT. These results establish a consistent overall advantage for DES under matched model and frame-budget settings, with the largest gains at eight frames.

Task-specific results. DES performs best on ER, PQA, AO, and AC at every frame budget, but its advantage does not extend to all tasks. With eight frames, DES improves over BOLT by 12.7 points on AO and 7.0 points on PQA. Its NQA gain is 20.3 points over Uniform but only 0.6 points over BOLT, indicating that much of the improvement on this task is shared by both selectors. Uniform leads on TR and AR at every budget, while BOLT achieves the highest NQA accuracy with 16 frames. The accuracy deficit on AR is most pronounced at 32 frames, where DES trails Uniform and BOLT by 6.5 and 4.5 points, respectively. Thus, DES offers stronger performance on several detail-oriented tasks, while holistic understanding remains a limitation relative to uniform sampling.

Effect of the frame budget. With eight frames, DES achieves an M-Avg that exceeds those of Uniform and BOLT with 32 frames by 4.8 and 1.2 points, respectively. Increasing DES's budget from 16 to 32 frames, however, leaves M-Avg unchanged at 68.6% at the reported precision. Ac-

Table 7: LongVideoBench validation accuracy (%) across duration groups and question categories with 8, 16, and 32 input frames. All methods use LLaVA-OneVision-Qwen2-7B-OV and are evaluated on 1,337 validation questions. Uniform uses evenly spaced frames. BOLT uses CLIP-L inverse-transform sampling (1 fps, power 2.5). DES is our query-conditioned frame selector. The highest accuracy among the three selectors is shown in bold red within each row and frame budget, including ties. Overall reports accuracy on the full validation set.

Group	Subset	8 frames			16 frames			32 frames		
		Uniform	BOLT	DES	Uniform	BOLT	DES	Uniform	BOLT	DES
Duration (seconds)	(8, 15]	**68.3**	64.0	**68.3**	**70.4**	66.1	67.2	**70.4**	63.0	64.0
	(15, 60]	66.9	63.4	**68.0**	**73.3**	66.3	72.7	**71.5**	69.2	70.3
	(180, 600]	52.7	53.4	**54.9**	52.4	54.1	**59.0**	54.9	58.3	**59.5**
	(900, 3600]	46.8	**48.4**	**48.4**	48.2	**51.6**	50.2	48.9	52.7	**54.6**
L1 Perception	S2E	63.4	**72.0**	66.7	**68.8**	66.7	67.7	**72.0**	**72.0**	69.9
	S2O	52.8	54.2	**59.7**	54.2	51.4	**59.7**	54.2	**61.1**	59.7
	S2A	63.6	70.5	**76.1**	76.1	71.6	**78.4**	75.0	79.5	**81.8**
	E2O	66.2	63.1	**69.2**	61.5	72.3	**73.8**	69.2	67.7	**72.3**
	O2E	63.2	**65.5**	**65.5**	64.4	69.0	**72.4**	59.8	**70.1**	**70.1**
	T2E	60.0	**63.1**	60.0	**61.5**	58.5	58.5	56.9	**63.1**	60.0
	T2O	**63.2**	53.9	61.8	55.3	56.6	**63.2**	**64.5**	57.9	63.2
	T2A	54.4	59.5	**65.8**	57.0	62.0	**69.6**	58.2	**63.3**	**63.3**
L2 Relation	E3E	61.7	**63.8**	60.6	64.9	62.8	**66.0**	**70.2**	67.0	**70.2**
	O3O	50.0	43.9	**53.0**	48.5	47.0	**50.0**	51.5	53.0	**54.5**
	SSS	**30.9**	21.6	25.8	29.9	29.9	**32.0**	32.0	**34.0**	**34.0**
	SOS	60.5	**65.4**	59.3	**69.1**	63.0	65.4	59.3	60.5	**69.1**
	SAA	54.2	51.4	**55.6**	51.4	56.9	**58.3**	50.0	50.0	**51.4**
	T3E	45.2	**49.3**	42.5	42.5	39.7	**49.3**	49.3	47.9	**53.4**
	T3O	**48.6**	37.8	41.9	45.9	**54.1**	44.6	**54.1**	47.3	47.3
	TOS	34.2	37.0	**38.4**	**41.1**	35.6	38.4	35.6	38.4	**43.8**
	TAA	**50.0**	45.1	46.3	53.7	**58.5**	40.2	**48.8**	**48.8**	43.9
Overall		54.2	54.1	**55.7**	55.9	56.3	**58.2**	56.7	58.0	**59.5**

curacy on TR, ER, and PQA improves, whereas accuracy on AR and AC declines by 1.5 and 3.8 points, respectively. Despite the decline on AC, DES still achieves the highest AC accuracy among the three methods at 32 frames. These results show that DES can achieve higher overall accuracy with fewer input frames, while increasing the frame budget does not benefit every task.

Interpretation and limitations. The gains on PQA and AO are consistent with DES's emphasis on question-relevant and complementary evidence. The semantic, option-discrimination, and action-related cues may help retain useful frames, while temporal coverage and diversity-aware selection may preserve evidence from different stages of a video. Conversely, prioritizing high-scoring local evidence may exclude broader context useful for TR and AR. Visual redundancy suppression may also discard similar frames depicting distinct action repetitions. However, AC accuracy also declines from 16 to 32 frames for Uniform and BOLT, so this pattern cannot be attributed to DES's MMR step alone. These observations suggest a trade-off between evidence selection and contextual coverage, but task accuracy alone does not establish the underlying causes.

A.4 Category-wise Comparison on Video-MME

Overall comparison. Table 9 compares Uniform, BOLT, and DES across six Video-MME categories. DES achieves the highest overall accuracy at all three frame budgets, reaching 58.07%, 59.52%, and 60.07% with 8, 16, and 32 frames, respectively. The gains over Uniform are 4.18, 2.71, and 1.70 percentage points, while those over BOLT are 2.00, 1.00, and 0.63 points. These results demonstrate a consistent overall advantage under matched model and frame-budget settings, with larger improvements when fewer input frames are available.

Table 8: Task-wise accuracy (%) on MLVU dev using LLaVA-OneVision-7B with different frame selection methods. Bold indicates the best result among the three methods within each frame budget. M-Avg is the unweighted mean accuracy across the seven tasks.

Method	Holistic		Single Detail			Multi Detail		M-Avg
	TR	AR	NQA	ER	PQA	AO	AC	
8 frames								
Uniform	**85.9**	**65.5**	58.9	61.6	57.0	42.1	38.3	58.5
BOLT	85.6	63.5	78.6	65.9	65.7	45.2	42.7	63.9
DES	84.4	65.0	**79.2**	**67.0**	**72.7**	**57.9**	**48.1**	**67.8**
16 frames								
Uniform	**87.1**	**69.0**	71.3	63.6	62.3	42.1	34.5	61.4
BOLT	84.0	64.5	**80.8**	67.6	72.4	48.3	42.7	65.8
DES	85.2	64.5	80.0	**68.5**	**75.0**	**58.3**	**48.5**	**68.6**
32 frames								
Uniform	**89.0**	**69.5**	73.5	65.1	66.0	45.2	33.0	63.0
BOLT	86.7	67.5	78.9	67.9	74.6	52.1	38.8	66.6
DES	87.5	63.0	**79.7**	**70.5**	**77.2**	**57.9**	**44.7**	**68.6**

Task definitions. TR: Topic Reasoning; AR: Anomaly Recognition; NQA: Needle QA; ER: Ego Reasoning; PQA: Plot QA; AO: Action Order; AC: Action Count. The evaluation covers 2,174 multiple-choice questions from MLVU dev, with 263, 200, 355, 352, 539, 259, and 206 questions in these tasks, respectively.

Table 9: Category-wise accuracy (%) on Video-MME without subtitles using LLaVA-OneVision-7B. Bold indicates the best result among the three frame selection methods within each frame budget. Overall reports accuracy across all 2,700 questions.

Method	Category						Overall
	Knowledge	Film & Television	Sports Competition	Artistic Performance	Life Record	Multilingual	
8 frames							
Uniform	53.95	55.28	53.78	55.28	52.06	55.56	53.89
BOLT	56.42	58.61	50.89	58.06	**57.46**	51.11	56.07
DES	**59.51**	**59.72**	**55.33**	**59.17**	56.83	**56.67**	**58.07**
16 frames							
Uniform	57.90	58.06	53.11	60.28	55.71	**54.44**	56.81
BOLT	58.52	61.39	55.11	**63.06**	57.94	50.00	58.52
DES	**60.99**	**62.78**	**56.67**	58.33	**60.00**	48.89	**59.52**
32 frames							
Uniform	60.49	59.72	56.22	58.61	57.14	52.22	58.37
BOLT	58.89	**63.06**	57.33	**63.33**	58.25	53.33	59.44
DES	**60.99**	62.22	**57.56**	60.83	**59.68**	**55.56**	**60.07**

Note. Knowledge, Film & Television, Sports Competition, Artistic Performance, Life Record, and Multilingual contain 810, 360, 450, 360, 630, and 90 questions, respectively. Overall is computed over all questions rather than as an unweighted average of the six category accuracies.

Category-specific results. DES achieves the highest accuracy in 13 of the 18 category–budget combinations, including Knowledge and Sports Competition at every frame budget. With eight frames, it improves over BOLT by 3.09 and 4.44 points in these two categories, respectively. On Life Record, DES trails BOLT by 0.63 points at eight frames but leads by 2.06 and 1.43 points at 16 and 32 frames. The most pronounced deficit occurs in Artistic Performance, where DES trails BOLT by 4.73 and 2.50 points at 16 and 32 frames. DES also falls below both baselines on Multilingual at 16 frames and below BOLT on Film & Television at 32 frames. Thus, its overall gains extend

across most categories, while performance on artistic content and multilingual questions remains less consistent.

Effect of the frame budget. Increasing DES's frame budget improves overall accuracy with diminishing returns: increasing the budget from 8 to 16 frames adds 1.45 points, whereas increasing it from 16 to 32 adds 0.55 points. With only eight frames, DES exceeds Uniform with 16 frames by 1.26 points and falls just 0.30 points below Uniform with 32 frames. Category-level trends are less uniform. Film & Television and Life Record peak at 16 frames, while Multilingual falls from 56.67% to 48.89% before recovering to 55.56%. The latter category contains only 90 questions, so a change in one answer's correctness shifts its accuracy by approximately 1.11 points. These results support the effectiveness of DES under limited frame budgets, while showing that additional frames do not consistently benefit every category.

Interpretation and limitations. The larger gains at eight frames are consistent with DES prioritizing question-relevant and complementary evidence. Multi-cue scoring may retain useful frames that question–frame similarity alone overlooks, while segment-wise candidate coverage and MMR may help preserve evidence from different stages of a video. However, visually similar frames can contain distinct temporal or fine-grained information. Redundancy suppression and temporal spacing may make such evidence less likely to be retained, potentially contributing to weaker results on Artistic Performance. Errors in text–image matching may also affect frame ranking, although fluctuations in Multilingual accuracy alone cannot establish this explanation. The results therefore support DES's overall effectiveness while suggesting a trade-off between reducing redundant frames and retaining task-specific evidence. Identifying the causes of individual category deficits requires frame-level analysis or targeted ablations.

Overall conclusions and future directions. The detailed results across the three benchmarks show that DES improves overall accuracy over uniform sampling, while its benefits vary across content categories and reasoning requirements (Tables 7, 8, and 9). The gains on longer videos and perception questions in LongVideoBench, plot QA and action ordering in MLVU, and Knowledge and Sports Competition in Video-MME are consistent with DES's emphasis on relevant and complementary evidence. Semantic, option-discrimination, and visual grounding cues may help prioritize informative frames, while temporal cues, segment-wise candidate coverage, and MMR may help retain evidence from different stages of a video. However, the deficits on short videos and some relational questions in LongVideoBench, topic reasoning and anomaly recognition in MLVU, and Artistic Performance in Video-MME suggest that high frame utility and visual diversity do not fully capture every task's evidence requirements. Option-based matching may favor visually plausible distractors, and visual redundancy penalties or temporal spacing may suppress similar or adjacent frames that carry distinct relational or temporal information. These are possible explanations, not component effects established by the benchmark comparisons; both selection errors and downstream reasoning limitations may contribute. Future work could therefore adapt cue weighting and the MMR trade-off to the question's evidence requirements, allocate part of the fixed frame budget to global context when needed, and preserve short frame sequences for questions requiring event continuity or repetition. Targeted ablations and selected-frame error analysis would help determine whether these changes address the observed limitations. Together, the results support combining relevance and complementarity for frame selection, while motivating more explicit modeling of contextual coverage and temporal relationships.

B Detailed Method Formulation

This appendix supplements Section 3.2 with details on normalization, text templates, question-dependent weights, and temporal cue computation. Following the notation in the main text, T denotes the number of video frames, C denotes the number of answer options, S_i^x denotes the scalar score of frame i, and $\mathbf{S}^x$ denotes the corresponding score vector over all frames.

B.1 Robust Normalization

For a score vector $\mathbf{x}$ of length T, let $q_{0.05}(\mathbf{x})$ and $q_{0.95}(\mathbf{x})$ denote its 5th and 95th percentiles, respectively. We define normalization as

$$[\mathcal{N}(\mathbf{x})]_i = \begin{cases} 0, & q_{0.95}(\mathbf{x}) - q_{0.05}(\mathbf{x}) < \epsilon_n, \\ \text{clip}\left(\dfrac{x_i - q_{0.05}(\mathbf{x})}{q_{0.95}(\mathbf{x}) - q_{0.05}(\mathbf{x})}, 0, 1\right), & \text{otherwise}, \end{cases} \tag{10}$$

where $\epsilon_n = 10^{-6}$. This operation scales scores from different sources to $[0, 1]$ and uses percentile clipping to limit the influence of extreme values on the scaling range.

For the option-matching matrix $E \in \mathbb{R}^{C \times T}$, we compute the percentiles jointly over all CT entries and apply the same scaling and clipping to each entry to obtain $\widehat{E}$. Thus, normalization is performed over the entire matrix, rather than row-wise or column-wise.

The Softmax used to compute the option distribution operates along the option dimension with a temperature of 1. The margin vector $\mathbf{m}$ and the distribution concentration vector $\mathbf{u}$ defined in the main text are normalized separately across frames before the option-discriminability scores are computed.

B.2 Construction of Text Views

All text views are constructed from the question and candidate answers using fixed templates, without an additional language model for rewriting or decomposition. The text views are defined as follows:

- **Question text.** We use the question Q directly.
- **Full context.** We append `Options:` to the question, followed by all labeled answer options in order.
- **Single-option text.** We combine the question with one candidate answer using the following two lines:

```
Question: {question}
Candidate answer: {option}
```

- **Object/scene text.** We prepend the following fixed instruction to the full context:

```
Find frames that show important objects, people,
scenes, text, attributes, and visual details needed
to answer the question: {full_query}
```

- **Action/interaction text.** We prepend the following fixed instruction to the full context:

```
Find frames that show key actions, movements,
interactions, temporal changes, and event transitions
needed to answer the question: {full_query}
```

Here, `{full_query}` denotes the full context containing the question and all candidate answers. We compute matching scores between each text view and the video-frame features, then normalize the scores as described in the main text.

B.3 Question-Dependent Weights and Visual Cue Fusion

We determine the auxiliary cue weights using deterministic keyword rules. We first convert the question to lowercase and then check whether the resulting string contains any of the following keywords:

- **Object keywords:** `what`, `which`, `who`, `where`, `color`, `number`, `many`, `wearing`, `text`, and `object`.
- **Action keywords:** `doing`, `move`, `put`, `take`, `open`, `close`, `walk`, `run`, `hold`, and `interact`.

- **Temporal keywords:** `before`, `after`, `then`, `first`, `last`, `start`, `end`, `next`, and `finally`.

The implementation uses substring matching. Let b_{obj}, b_{act}, and b_{temp} denote the three matching indicators, each taking a value of 1 if a match is found and 0 otherwise. The weights are set as follows:

$$\gamma_{obj} = \begin{cases} 1.0, & b_{obj} = 1, \\ 0.5, & b_{obj} = 0, \end{cases} \qquad \gamma_{act} = \begin{cases} 1.0, & b_{act} = 1 \text{ or } b_{temp} = 1, \\ 0.5, & \text{otherwise}, \end{cases} \tag{11}$$

and

$$\gamma_{temp} = \begin{cases} 1.0, & b_{temp} = 1, \\ 0.3, & b_{temp} = 0. \end{cases} \tag{12}$$

These weights adjust the relative strengths of the auxiliary cues; each cue is retained even when no corresponding keyword is matched. The object and action scores are fused as

$$S_i^{vis} = \frac{\gamma_{obj} S_i^{obj} + \gamma_{act} S_i^{act}}{\gamma_{obj} + \gamma_{act} + \epsilon}, \tag{13}$$

where $\epsilon = 10^{-6}$. The weights are identical across all frames associated with a given question.

B.4 Full Computation of Temporal Cues

Visual changes between adjacent frames. We apply L_2 normalization to the visual features to obtain $\hat{\mathbf{v}}_i$. For $T \geq 2$, the change between adjacent frames is defined as

$$\delta_i = 1 - \hat{\mathbf{v}}_i^\top \hat{\mathbf{v}}_{i-1}, \qquad i = 2, \ldots, T. \tag{14}$$

To make the change sequence the same length as the frame sequence, we set $\delta_1 = \delta_2$. When $T = 1$, we set $\delta_1 = 0$. The normalized change scores are then

$$\mathbf{s}^{chg} = \mathcal{N}(\boldsymbol{\delta}), \qquad \boldsymbol{\delta} = (\delta_1, \ldots, \delta_T). \tag{15}$$

Question-relevant seed scores. We combine the action-matching and visual-change scores with equal weights, then weight the result by semantic relevance:

$$\mathbf{z} = \mathcal{N}\left(\mathbf{S}^{sem} \odot \frac{\mathbf{S}^{act} + \mathbf{s}^{chg}}{2}\right). \tag{16}$$

Here, $\odot$ denotes element-wise multiplication. Incorporating semantic relevance allows the seed responses to account for the current question, rather than depend solely on action matching or the magnitude of visual changes.

Local temporal context aggregation. For each frame, we construct a window with radius $r \geq 1$:

$$\mathcal{W}_r(i) = \{j : 1 \leq j \leq T,\ |j - i| \leq r\}. \tag{17}$$

The window is truncated at the video boundaries to include only valid frames. We combine the mean and maximum seed scores within the window with equal weights:

$$h_i = \frac{1}{2|\mathcal{W}_r(i)|} \sum_{j \in \mathcal{W}_r(i)} z_j + \frac{1}{2} \max_{j \in \mathcal{W}_r(i)} z_j. \tag{18}$$

We then normalize across frames to obtain the local temporal context scores:

$$\mathbf{S}^{event} = \mathcal{A}_r(\mathbf{z}) = \mathcal{N}\big((h_1, \ldots, h_T)\big). \tag{19}$$

Here, $\mathbf{S}^{event}$ represents aggregated local responses and does not correspond to explicit semantic event segmentation.

Table 10: Models, lmms-eval classes, and vision encoders used in the experiments.

Model	lmms-eval class	Vision encoder
LLaVA-OneVision-7B	`llava_onevision`	SigLIP-SO400M/14 @ 384
Qwen3-VL-8B	`qwen3_vl` (chat wrapper)	Qwen3-VL ViT, patch size 16
Video-LLaVA-7B	`video_llava`	CLIP-ViT-L/14 @ 224
LongVA-7B	`longva`	CLIP-ViT-L/14 @ 336
LLaVA-Video-7B	`llava_vid`	SigLIP-SO400M/14 @ 384
Qwen2.5-VL-7B	`qwen2_5_vl` (chat wrapper)	Qwen2.5-VL ViT, patch size 14

Temporal cue fusion. We combine the current frame's change score with its local temporal context and apply the question-dependent temporal weight:

$$S_i^{temp} = \gamma_{temp} \sqrt{s_i^{chg} S_i^{event} + \epsilon}. \tag{20}$$

This yields the temporal cue vector $\mathbf{S}^{temp} = (S_1^{temp}, \ldots, S_T^{temp})$, which is used in the multi-cue fusion formula in the main text.

C Experimental Details

This appendix documents the model configurations, benchmark settings, and prompts used in our evaluation. It specifies the lmms-eval interfaces and vision encoders for each model, the dataset splits and evaluation metrics, and the text prompts used for the four benchmarks.

C.1 Models and Evaluation Interfaces

Table 10 summarizes the models used in our experiments, their corresponding lmms-eval classes, and their vision encoders. The Qwen3-VL and Qwen2.5-VL models use chat wrappers in the evaluation framework.

C.2 Benchmarks and Evaluation Settings

We evaluate on Video-MME, LongVideoBench, MLVU, and NExT-QA. Table 11 lists the dataset identifiers, evaluation splits, and effective question counts. Table 12 provides the corresponding lmms-eval task and metric identifiers.

Subtitle configuration. No subtitle text is supplied in any of the four benchmark evaluations. For Video-MME, we use `videomme` rather than `videomme_w_subtitle`. For LongVideoBench, we use the video-only task `longvideobench_val_v` rather than the subtitle-interleaved task `longvideobench_val_i`. The MLVU and NExT-QA configurations used here do not include a subtitle channel.

Evaluation splits and question counts. Video-MME contains 2,700 questions associated with 900 videos, with three questions per video. We evaluate on the LongVideoBench validation split, which contains 1,337 questions. For MLVU, we use the `sy1998/MLVU_dev` dataset and its `test` split, covering 2,174 questions across seven multiple-choice task categories. For NExT-QA, we use the Hugging Face `MC` configuration and its `test` split, which contains 8,564 questions with answer options labeled A–E.

C.3 Evaluation Prompts

The following templates define the evaluation prompts for the four benchmarks. Placeholders enclosed in braces are replaced with the corresponding question and answer options. We retain the benchmark-specific option labels and response instructions shown below.

Table 11: Dataset identifiers, evaluation splits, and effective question counts. MLVU uses the `test` split of the `MLVU_dev` dataset.

Benchmark	Dataset	Split	Questions
Video-MME	`lmms-lab/Video-MME`	test	2,700
LongVideoBench	`longvideobench/LongVideoBench`	validation	1,337
MLVU (dev)	`sy1998/MLVU_dev`	test	2,174
NExT-QA MC	`lmms-lab/NExTQA`, config `MC`	test	8,564

Table 12: lmms-eval task and metric identifiers used for each benchmark.

Benchmark	Task identifier	Metric identifier
Video-MME	`videomme`	`videomme_perception_score`
LongVideoBench	`longvideobench_val_v`	`lvb_acc`
MLVU (dev)	`mlvu_dev`	`mlvu_percetion_score`
NExT-QA MC	`nextqa_mc_test`	`exact_match`

Video-MME. The prompt asks the model to select an answer based on the video and return only the corresponding option letter.

```
You are a helpful assistant.

Select the best answer to the following
multiple-choice question based on the video.  Respond
with only the letter (A, B, C, or D) of the correct
option.
{question}
{options}

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

LongVideoBench. The template presents the question followed by labeled answer options and an instruction to return the option letter directly.

```
You are a helpful assistant.
{question}
A. {option0}
B. {option1}
C. {option2}
D. {option3}
E. {option4}
Answer with the option's letter from the given choices
directly.
```

MLVU. The template uses parenthesized option labels and ends with `Best option:` to prompt the model's response.

```
You are a helpful assistant.
{question}
(A) {option A}
(B) {option B}
(C) {option C}
```

```
(D) {option D}
Only give the best option.
Best option:
```

NExT-QA MC. The prompt contains the question and five options labeled A–E, without an additional instruction after the options.

```
You are a helpful assistant.
{question}
A. {a0}
B. {a1}
C. {a2}
D. {a3}
E. {a4}
```