Q-GPS: QUERY-GUIDED PROGRESSIVE FRAME SELECTION FOR LONG VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Frame selection enables multimodal large language models (MLLMs) to process long videos within limited visual-token budgets by retaining only a small subset of frames. Highly relevant frames may concentrate within the same short interval, leaving other query-relevant moments unrepresented. Effective selection therefore requires assessing both candidate relevance and the additional temporal coverage each frame provides given previous selections. We propose Query-Guided Progressive Frame Selection (Q-GPS), a training-free, model-agnostic method that progressively selects frames according to relevance-calibrated gains in query-weighted temporal coverage. It initializes its residual state with a query-independent path-spectral temporal covariance that models temporal coupling, and uses this state as a surrogate for the temporal coverage remaining after previous selections. At each step, it **selects** the frame with the largest relevance-calibrated coverage gain, **conditions** the covariance on the selected frame's temporal location, and **rescores** the remaining candidates using the updated residual state. Across controlled evaluations with Qwen2.5-VL-7B and frame budgets of 8, 16, and 32, Q-GPS improves average accuracy over uniform sampling by 4.70%, 7.28%, and 11.05% on Video-MME, LongVideoBench, and MLVU, respectively. Code is available at https://anonymous.4open.science/r/Q-GPS-FEB4/.

1 INTRODUCTION

Long-video question answering enables users to retrieve information and interpret events in extended video content through natural-language queries. Answers may depend on brief events, temporal relations, or observations separated by long intervals. Modern multimodal large language models (MLLMs), such as Qwen2.5-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2025), and InternVL3 (Zhu et al., 2025), support this task by jointly processing visual content and language. Their practical visual-token and computation budgets, however, accommodate only a small fraction of a long video's frames. Selecting a compact visual context that preserves the information needed for the query is therefore central to effective and efficient long-video question answering.

Existing frame selectors combine query relevance with temporal coverage or set-level diversity. AKS (Tang et al., 2025) balances relevance with temporal coverage, MDP^3 (Sun et al., 2025) uses a query-conditioned determinantal objective, and WFS-SB (Chen et al., 2026) combines semantic segmentation with diversity-aware selection. These methods account for the selected set, but their different coverage and diversity objectives assign different values to adding a candidate frame.

A central difficulty arises when highly query-relevant frames concentrate in a short interval, while answering the query requires observations from several separated events. Neighboring frames can contain overlapping content, so repeated selections from one interval may add little context and leave another event unseen. For example, explaining why Tom flees after unmasking Jerry requires observations of the disguise, its removal, and the later chase by the real islanders (Fig. 1). Simply spreading selections farther apart is also insufficient: a wider temporal span can include frames unrelated to the question. The aim is to extend coverage where it matters for the query, while preserving the relevance of the selected frames.

The value of covering a location also changes as selection proceeds. Once a frame is selected, another candidate coupled to the same temporal structure may provide less additional coverage, even though

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.

its query relevance remains unchanged. A useful criterion should reflect this change while continuing to value relevant locations that remain underrepresented. Candidate relevance alone does not capture this dependence on what has already been selected. Within a set-aware framework, we therefore ask how to explicitly quantify the effect of selecting one frame on the remaining coverage needs across query-relevant locations. Our criterion combines a candidate's direct relevance with the reduction it induces in a query-weighted residual coverage objective over all currently unselected locations.

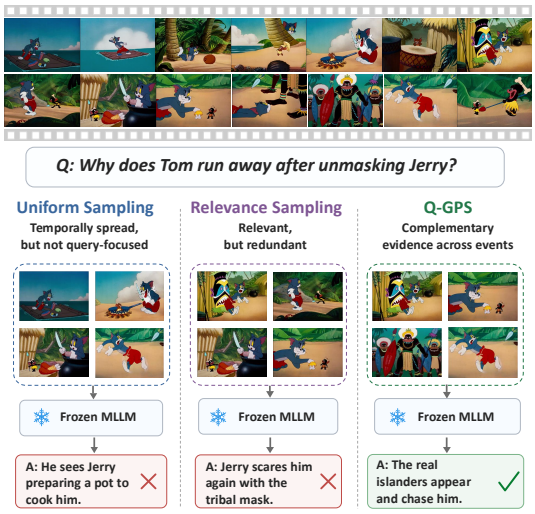


Figure 1: Qualitative frame-selection comparison using the same frozen MLLM. In this example, Uniform misses query-relevant events, while relevance-based selection concentrates on redundant evidence. Q-GPS selects complementary evidence across events and produces the correct answer.

We introduce *Query-Guided Progressive Frame Selection (Q-GPS)*, a training-free selector that implements this criterion through an evolving residual temporal covariance. Initialized from a query-independent path-spectral covariance, the residual state serves as a temporal surrogate for coverage remaining after previous selections. Query relevance determines both the importance of covering each location and the direct utility of selecting each candidate. At each step, Q-GPS selects the frame with the largest relevance-calibrated coverage gain, conditions the covariance on its location, and reevaluates the remaining candidates using the updated state. Selection uses fixed relevance scores and requires no feedback from the downstream MLLM, allowing integration with frozen downstream models.

We evaluate Q-GPS on Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025) with frame budgets of 8, 16, and 32. With Qwen2.5-VL-7B, its accuracy gains over uniform sampling average 4.70%, 7.28%, and 11.05% on the three benchmarks, respectively; gains over uniform sampling also persist with two additional MLLMs. Selection-behavior analysis connects these improvements to the motivating relevance–coverage trade-off: Q-GPS spans a wider temporal range than Top-K while retaining higher mean query relevance than AKS and MDP^3, whose selections span a larger range (Fig. 5). Ablation results further support the benefit of progressively updating the residual coverage state: disabling covariance updates reduces Video-MME accuracy by 3.33% under a 16-frame budget (Tab. 3).

Our contributions are summarized as follows:

- We introduce a query-aware selection criterion that combines direct candidate relevance with the conditioning-induced reduction of a query-weighted residual coverage objective over currently unselected locations.
- We realize this criterion with a path-spectral temporal covariance and an exact low-rank implementation, using a closed-form path basis to avoid numerical eigendecomposition and dense covariance storage.
- We demonstrate accuracy gains through controlled comparisons and cross-backbone evaluations, and use selection-behavior analysis and ablations to characterize the relevance–coverage trade-off and the benefit of progressive updates.

2 Related Work

Relevance and set-aware frame selection. Query-aware selectors combine relevance with coverage or diversity to construct compact video inputs. BOLT (Liu et al., 2025) and Q-Frame (Zhang et al., 2025c) guide sampling by query relevance, with Q-Frame additionally adapting resolution. Set-aware methods include AKS (Tang et al., 2025), MDP^3 (Sun et al., 2025), and WFS-SB (Chen et al., 2026), which use hierarchical temporal coverage, sequential determinantal diversity, and segment-wise

MMR (Carbonell & Goldstein, 1998), respectively. Threading Keyframes with Narratives (Fang et al., 2026) combines relevance–diversity selection with textual narratives from non-keyframes. These approaches define additional value through different coverage and diversity objectives. Q-GPS instead measures how selecting a candidate reduces query-weighted residual temporal coverage across all currently unselected locations. It computes this reduction by conditioning an evolving temporal covariance and calibrates the gain by the candidate's own relevance.

Adaptive and learned selection. T* (Ye et al., 2025) and DIG (Li et al., 2026b) adapt temporal search using query-grounded visual cues and query types, respectively; FOCUS (Zhu et al., 2026) and A.I.R. (Zou et al., 2026) guide exploration through confidence and iterative VLM feedback. Frame-Voyager (Yu et al., 2025), FrameOracle (Li et al., 2026a), and the M-LLM selector of Hu et al. (2025) learn query-dependent selection from supervision. Q-GPS operates on a fixed candidate pool and frozen relevance scores, adapting through residual-coverage updates without selector training or downstream MLLM feedback.

Covariance-based observation selection. Predictive-variance reduction in active learning (Cohn et al., 1996) and information objectives in Gaussian-process sensor placement (Krause et al., 2008) evaluate observations by their effects beyond the selected location. Q-GPS draws on covariance conditioning but constructs a deterministic temporal coverage surrogate from path modes and query-derived weights. Its updates depend only on selected locations, without observed response values; the covariance models residual temporal coverage rather than uncertainty in the MLLM's answer.

3 METHOD

Overview. Q-GPS selects a fixed-budget frame set by combining query relevance with the temporal coverage each candidate adds after previous selections (Fig. 2). Given ordered candidates and frozen query–frame scores, it first constructs a path-spectral covariance to model temporal coupling, then uses query relevance to weight the remaining coverage at unselected locations. It repeatedly selects the candidate with the largest relevance-calibrated coverage gain, conditions the covariance on that location, and rescores the remaining candidates. The selected frames are returned in chronological order for inference with a frozen MLLM.

3.1 PROBLEM FORMULATION

Given a video $V = \{v_t\}_{t=1}^{T}$, a natural-language query Q, and a frame budget B, frame selection seeks a compact visual context that preserves the evidence required by a frozen multimodal large language model (MLLM). We first uniformly sample an ordered candidate sequence $\mathcal{X} = \{x_i\}_{i=1}^{N}$, where $x_i = v_{t_i}$ and $t_1 < \cdots < t_N$. We assume $1 \le B \le N \le T$ and write $[N] = \{1, \ldots, N\}$. The selector constructs $\mathcal{I}_B \subseteq [N]$ with $|\mathcal{I}_B| = B$; the corresponding frames are restored to chronological order before downstream inference.

To measure the relevance of each candidate to the query, a frozen vision–language scorer computes, for each $i \in [N]$,

$$y_i = \phi(Q, x_i), \tag{1}$$

where larger y_i indicates stronger query–frame alignment. The scorer is used only to construct this relevance sequence. It is not trained or updated, and the downstream MLLM is never called during selection. The ordered scores $\{y_i\}_{i=1}^{N}$ and the candidate chronology form the inputs to Q-GPS.

We min–max normalize the raw scores within each query to obtain the relevance profile $\mathbf{r} = (r_i)_{i=1}^{N}$. With $y_{\min} = \min_i y_i$, $y_{\max} = \max_i y_i$, and a numerical threshold $\varepsilon > 0$, we set every entry of $\mathbf{r}$ to 1 when $y_{\max} - y_{\min} \le \varepsilon$; otherwise, its entries are

$$r_i = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}. \tag{2}$$

This fallback handles constant or numerically indistinguishable scorer outputs and avoids division by a zero or near-zero score range.

We formulate frame selection as a B-step process guided by relevance and residual temporal coverage. Given the current selected set $\mathcal{I}_b$, each unselected candidate is valued by its direct query relevance and its gain in query-weighted temporal coverage.

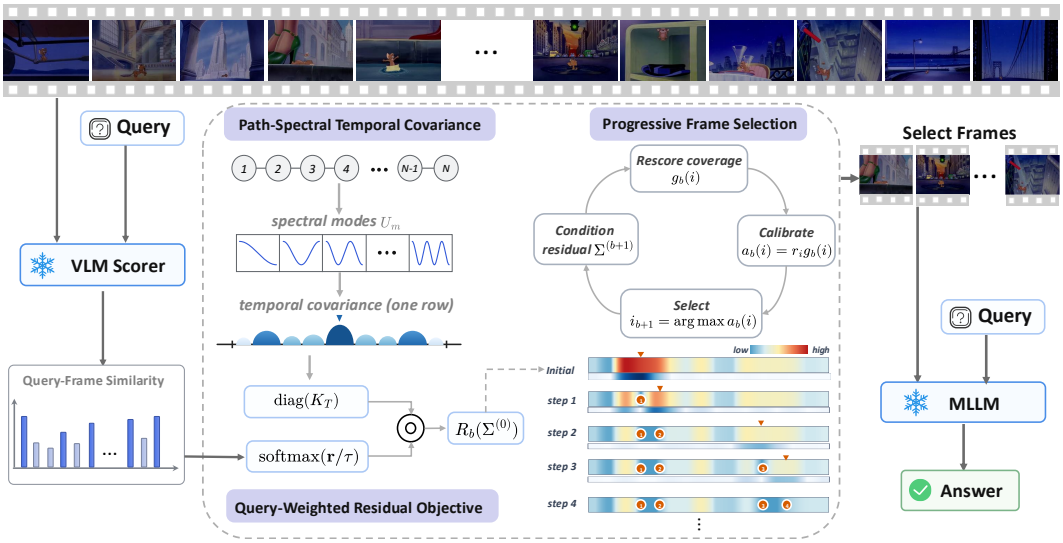


Figure 2: Overview of the Q-GPS framework. Given a long video and a query, a frozen VLM scorer produces frame–query relevance, while the retained modes of an ordered path construct a path-spectral temporal covariance. Its diagonal and the softmax-derived query weights initialize the query-weighted residual objective. Q-GPS then iteratively computes coverage gain, calibrates it by direct relevance, selects a frame, and conditions the residual covariance, with the strips illustrating the evolving residual state and selected locations. Finally, the selected frames are restored to temporal order and passed with the query to a frozen downstream MLLM for inference.

3.2 Query-Guided Progressive Frame Selection

Path-Spectral Temporal Covariance. Following standard spectral graph constructions (Chung, 1997; Shuman et al., 2013), we construct an unweighted path graph over the ordered candidate indices. With $\mathbf{1}_N$ denoting the all-ones vector, its adjacency is $A_{ij} = \mathbb{I}[|i-j| = 1]$ and its degree matrix is $D = \mathrm{diag}(A\mathbf{1}_N)$. The symmetric normalized Laplacian is

$$L = I_N - D^{-1/2} A D^{-1/2}. \tag{3}$$

For the degenerate $N = 1$ case, we set $L = [0]$ and $U_1 = [1]$. This graph depends only on temporal order: two nodes are connected exactly when their candidates are adjacent in the sampled sequence. Let $\{(\lambda_\ell, \mathbf{u}_\ell)\}_{\ell=1}^{N}$ be the orthonormal eigenpairs of L, ordered by ascending eigenvalue. Low-frequency eigenvectors describe smooth temporal patterns, whereas later modes represent increasingly rapid variation along the sequence.

The number of retained modes adapts to the output budget:

$$m = \min\{N, \lceil c_{\mathrm{rank}} B \rceil\}. \tag{4}$$

Here $c_{\mathrm{rank}} > 0$ controls the retained temporal resolution, and $U_m = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$ collects the first m eigenvectors. Linking m to B allows finer temporal variation to be represented as the selection budget grows. We assign equal spectral mass to the retained modes, with $p_\ell = 1/m$ for $\ell = 1, \ldots, m$. The resulting temporal covariance is

$$K_T = N U_m \,\mathrm{diag}(\mathbf{p}) U_m^\top = \frac{N}{m} U_m U_m^\top. \tag{5}$$

The path basis and spectral weights are defined independently of the query; equal weights avoid favoring particular retained modes. Because U_m has orthonormal columns, $K_T \succeq 0$ and $\mathrm{tr}(K_T) = N$. Here, K_T serves as a deterministic temporal coverage covariance: a large $|(K_T)_{ij}|$ indicates that locations i and j share retained temporal modes. Equal mass applies to the retained spectral modes, while query information enters through the residual objective and candidate relevance.

Query-Weighted Residual Objective. We convert the normalized relevance scores into target-location weights,

$$w_i = \frac{\exp(r_i/\tau)}{\sum_{j=1}^{N} \exp(r_j/\tau)}, \tag{6}$$

where $\tau > 0$ controls how strongly the target-location weights concentrate on highly relevant locations.

The three quantities play distinct roles: K_T defines the initial temporal coverage geometry, w_j specifies the utility of covering target location j, and r_i calibrates the direct utility of selecting candidate i.

We initialize the selected set as $\mathcal{I}_0 = \emptyset$ and the residual covariance as $\Sigma^{(0)} = K_T$. At step b, let $\mathcal{U}_b = [N] \setminus \mathcal{I}_b$ be the unselected locations. For a residual covariance Σ, we define the query-weighted residual objective as

$$R_b(\Sigma) = \sum_{j \in \mathcal{U}_b} w_j \Sigma_{jj}. \tag{7}$$

Each diagonal entry Σ_{jj} represents residual coverage demand at location j under the temporal model. The weighted sum therefore measures the remaining coverage demand at query-relevant locations; a smaller value indicates less residual demand. Restricting the sum to $\mathcal{U}_b$ removes selected locations from the objective without redistributing their weight.

Progressive Frame Selection. For each candidate $i \in \mathcal{U}_b$, we measure the reduction induced by conditioning using the same current unselected set $\mathcal{U}_b$ in both terms. With observation-noise variance $\nu > 0$, the hypothetically conditioned covariance is

$$\Sigma^{(b|i)} = \Sigma^{(b)} - \frac{\Sigma^{(b)}_{:,i} \Sigma^{(b)}_{i,:}}{\Sigma^{(b)}_{ii} + \nu}. \tag{8}$$

Conditioning is implemented as a deterministic, location-based coverage update that suppresses the residual temporal components shared with candidate i.

Using covariance symmetry, substituting Eq. (8) into Eq. (7) yields the conditioning-induced coverage gain over the current unselected set $\mathcal{U}_b$:

$$\begin{aligned} g_b(i) &= R_b\left(\Sigma^{(b)}\right) - R_b\left(\Sigma^{(b|i)}\right) \\ &= \sum_{j \in \mathcal{U}_b} w_j \left(\Sigma^{(b)}_{jj} - \Sigma^{(b|i)}_{jj}\right) \\ &= \frac{\sum_{j \in \mathcal{U}_b} w_j \left(\Sigma^{(b)}_{ji}\right)^2}{\Sigma^{(b)}_{ii} + \nu}. \end{aligned} \tag{9}$$

The gain $g_b(i)$ sums the conditioning-induced reduction across all currently unselected target locations, weighted by w_j. It therefore accounts for the candidate's effect on other relevant locations as well as its own. We recompute this gain from the updated residual covariance after every selection.

A candidate can reduce residual coverage at highly weighted locations without itself having high query relevance. The factor r_i therefore favors directly relevant frames, complementing the target-location weights w_j in the coverage gain. We select

$$i_{b+1} = \arg\max_{i \in \mathcal{U}_b} r_i g_b(i). \tag{10}$$

After selection, we set $\mathcal{I}_{b+1} = \mathcal{I}_b \cup \{i_{b+1}\}$. The conditioned covariance becomes the next residual covariance:

$$\Sigma^{(b+1)} = \Sigma^{(b|i_{b+1})}. \tag{11}$$

At the next iteration, each remaining candidate is rescored by its relevance-calibrated reduction of the updated query-weighted residual objective. A smaller ν more strongly suppresses candidates coupled to the selected location, whereas a larger ν preserves more residual covariance. After B iterations, Q-GPS returns the selected frames in their original temporal order. Algorithm 1 summarizes the complete selection procedure.

Efficient implementation. We implement the progressive updates through the exact factorization $\Sigma^{(b)} = U_m P_b U_m^\top$, avoiding explicit storage of the $N \times N$ residual covariance. The closed-form path basis also eliminates numerical eigendecomposition (Appendix A.1). The resulting B-step selector requires $\mathcal{O}(BNm^2)$ time and $\mathcal{O}(Nm + m^2)$ memory and is algebraically equivalent to the dense formulation above. The supplementary material provides the update rules and equivalence proof (Appendix A.2) and the detailed complexity analysis (Appendix A.4).

Algorithm 1: Query-Guided Progressive Frame Selection

Input: Ordered candidates $\mathcal{X} = \{x_i\}_{i=1}^{N}$, query Q, budget $1 \le B \le N$
Parameters: Frozen scorer ϕ; $c_{\text{rank}}, \tau, \nu, \varepsilon > 0$
Output: B selected candidate indices in chronological order

1. Path-spectral temporal covariance

$m \leftarrow \min\{N, \lceil c_{\text{rank}} B \rceil\}$ ▷ Eq. (4)
Form U_m from the first m analytic normalized-path modes; use $U_1 = [1]$ if $N = 1$
$K_T \leftarrow (N/m) U_m U_m^\top$ ▷ Eq. (5)

2. Query-weighted residual objective

$y_i \leftarrow \phi(Q, x_i)$ for all $i \in [N]$ ▷ Frozen scoring; Eq. (1)
Min–max normalize $\mathbf{y}$ to $\mathbf{r}$ (use $\mathbf{1}_N$ if $\max_i y_i - \min_i y_i \le \varepsilon$) ▷ Eq. (2)
$\mathbf{w} \leftarrow \mathrm{softmax}(\mathbf{r}/\tau)$ ▷ Fixed query weights; Eq. (6)

3. Progressive frame selection

$\mathcal{I}_0 \leftarrow \emptyset$, $\Sigma^{(0)} \leftarrow K_T$
for $b = 0, \ldots, B-1$ **do**
  $\mathcal{U}_b \leftarrow [N] \setminus \mathcal{I}_b$
  for $i \in \mathcal{U}_b$ **do**
    $g_b(i) \leftarrow \dfrac{\sum_{j \in \mathcal{U}_b} w_j \left(\Sigma^{(b)}_{ji}\right)^2}{\Sigma^{(b)}_{ii} + \nu}$ ▷ Rescore coverage; Eq. (9)
    $a_b(i) \leftarrow r_i g_b(i)$ ▷ Calibrate by direct relevance
  $i_{b+1} \leftarrow \arg\max_{i \in \mathcal{U}_b} a_b(i)$ ▷ Select; lexicographic ties in Appendix A.3
  $\mathcal{I}_{b+1} \leftarrow \mathcal{I}_b \cup \{i_{b+1}\}$
  $\Sigma^{(b+1)} \leftarrow \Sigma^{(b)} - \dfrac{\Sigma^{(b)}_{:,i_{b+1}} \Sigma^{(b)}_{i_{b+1},:}}{\Sigma^{(b)}_{i_{b+1} i_{b+1}} + \nu}$ ▷ Condition; Eq. (8)
return $\mathrm{sort}(\mathcal{I}_B)$ ▷ Restore chronological order

4 Experiments

We evaluate frame-selection accuracy under a controlled protocol, examine transfer to additional MLLMs, and analyze the components and computational cost of Q-GPS.

4.1 Experimental Setup and Details

Benchmarks and evaluation. We evaluate on three long-video multiple-choice QA benchmarks: Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2025). We use the full splits in our evaluation protocol: 2,700 Video-MME questions, 1,337 LongVideoBench validation questions, and 2,174 MLVU development questions. Subtitles are excluded throughout. We report sample-level accuracy for Video-MME and LongVideoBench and the unweighted mean over MLVU's seven multiple-choice tasks. All deltas and aggregate differences are computed from unrounded full-split accuracies; displayed cells are rounded independently.

Implementation details. For our primary controlled evaluation, we use frame budgets $B \in \{8, 16, 32\}$ under the shared protocol specified in Tab. 1. Unless otherwise specified, we uniformly sample up to 512 ordered candidate frames per video. For these controlled methods, frame relevance is the positive-class softmax probability from the frozen BLIP-2-ITM ViT-g head (Li et al., 2023). We use a single configuration, $c_{\text{rank}} = 2$, $\nu = 0.05$, and $\tau = 0.15$, across all datasets, frame budgets, scorers, and downstream MLLMs. Tab. 4 examines sensitivity around this shared configuration. Downstream MLLM inference is conducted on NVIDIA RTX 4090 GPUs with 24 GB of memory. We report accuracy measured by the LMMs-Eval toolkit (Zhang et al., 2025b). Further implementation details are provided in the supplementary material (Appendix B.3).

4.2 Comparison with State-of-the-Art Methods

Main results. Tab. 1 compares Q-GPS, Uniform sampling, AKS (Tang et al., 2025), MDP^3 (Sun et al., 2025), WFS-SB (Chen et al., 2026), and DIG (Li et al., 2026b) across three benchmarks and

Table 1: Frame-selection comparison with Qwen2.5-VL-7B. † indicates results reproduced under a controlled setting using the same frozen BLIP-2-ITM ViT-g model for query–frame scoring, the same candidate sequence, and the same inference model as Q-GPS; only the frame-selection rule varies. ‡ indicates results reproduced using the official implementation with LLM-based frame scoring and candidate frames sampled at 2 fps. We report accuracy (%); Δ is measured relative to Uniform. Bold and underlined values indicate the highest and second-highest values, with ties retained.

Method	Venue	Frames	Video-MME (w/o subs)					LongVideoBench						MLVU	
			Short	Med.	Long	Overall	Δ	15s	60s	600s	3600s	Overall	Δ	Overall	Δ
Uniform	–	8	60.0	51.3	45.1	52.15	–	60.8	61.6	50.7	44.9	51.08	–	53.65	–
AKS†	CVPR'25	8	65.4	55.0	49.7	56.70	+4.56	66.7	64.5	**60.7**	50.9	57.89	+6.81	62.46	+8.81
MDP^3,†	ICCV'25	8	65.0	55.9	**51.3**	57.41	+5.26	64.0	67.4	53.4	47.0	54.00	+2.92	61.91	+8.26
WFS-SB†	CVPR'26	8	66.9	55.3	49.4	57.22	+5.07	61.9	65.1	56.8	49.3	55.42	+4.34	63.77	+10.12
DIG‡	CVPR'26	8	63.4	54.7	47.2	55.11	+2.96	63.0	64.5	53.4	49.8	54.67	+3.59	58.57	+4.92
Q-GPS	–	8	**67.2**	**56.7**	49.7	**57.85**	**+5.70**	**68.3**	**70.9**	59.0	**55.1**	**60.21**	**+9.12**	**64.16**	**+10.51**
Uniform	–	16	66.3	55.3	48.1	56.59	–	67.7	66.3	53.4	46.6	54.23	–	56.09	–
AKS†	CVPR'25	16	68.3	59.0	50.9	59.41	+2.81	68.8	68.0	58.0	52.0	58.26	+4.04	63.39	+7.31
MDP^3,†	ICCV'25	16	66.9	58.8	51.3	59.00	+2.41	69.3	68.6	55.3	47.3	55.65	+1.42	62.43	+6.34
WFS-SB†	CVPR'26	16	**70.9**	59.6	51.3	60.59	+4.00	68.3	68.0	57.5	54.4	59.09	+4.86	65.41	+9.32
DIG‡	CVPR'26	16	68.1	56.9	**51.4**	58.81	+2.22	65.6	66.9	56.8	52.7	57.59	+3.37	63.23	+7.14
Q-GPS	–	16	70.2	**61.6**	**51.4**	**61.07**	**+4.48**	**71.4**	**71.5**	**58.5**	**54.6**	**60.36**	**+6.13**	**66.99**	**+10.90**
Uniform	–	32	70.9	58.8	51.9	60.52	–	72.0	72.7	53.6	49.5	56.92	–	58.85	–
AKS†	CVPR'25	32	72.4	60.8	53.0	62.07	+1.56	**73.0**	73.3	57.0	51.2	58.94	+2.02	66.50	+7.65
MDP^3,†	ICCV'25	32	72.8	63.0	**54.8**	63.52	+3.00	71.4	**75.6**	59.2	52.0	59.99	+3.07	67.34	+8.49
WFS-SB†	CVPR'26	32	72.3	62.9	53.3	62.85	+2.33	71.4	**75.6**	60.0	54.6	61.33	+4.41	69.39	+10.54
DIG‡	CVPR'26	32	70.2	61.4	53.0	61.56	+1.04	68.8	69.8	60.7	52.7	59.61	+2.69	66.75	+7.90
Q-GPS	–	32	**75.2**	**63.8**	54.3	**64.44**	**+3.93**	**73.0**	73.3	**63.8**	**57.1**	**63.50**	**+6.58**	**70.58**	**+11.73**

three frame budgets. AKS, MDP^3, WFS-SB, and Q-GPS share the controlled protocol specified in the caption, isolating the frame-selection method; DIG is included under its distinct official protocol. Q-GPS achieves the highest reported overall accuracy in all nine benchmark–budget settings, with positive gains over Uniform at every budget. The mean accuracy gain over competitive WFS-SB across these settings is statistically significant (Appendix C.4).

Backbone transfer. Tab. 2 evaluates Q-GPS with LLaVA-OV (Li et al., 2025) and InternVL3 (Zhu et al., 2025). Fig. 4 further shows that the gains over Uniform persist across all three frame budgets and benchmarks for both backbones. In the six settings displayed in Tab. 2, the gains are $6.3\%/7.2\%/8.4\%$ on Video-MME/LongVideoBench/MLVU with LLaVA-OV and $2.0\%/5.9\%/5.9\%$ with InternVL3. Q-GPS also achieves the best LongVideoBench result among the compared selectors with both backbones. Together, these results indicate that the effectiveness of Q-GPS is not tied to a specific backbone.

4.3 Ablation Studies and Analysis

Candidate density. Fig. 3 examines the effect of candidate count at a fixed output budget of $B = 16$. Increasing the number of uniformly sampled candidates from 64 to 512 improves accuracy by $2.26\%/2.69\%/3.63\%$ on Video-MME/LongVideoBench/MLVU, respectively, demonstrating the benefit of denser candidate sampling while keeping the number of frames passed to the downstream MLLM unchanged. Doubling the candidate count from 512 to 1024 changes accuracy by at most 0.18% across the three benchmarks, indicating diminishing returns and supporting the default choice of $N = 512$.

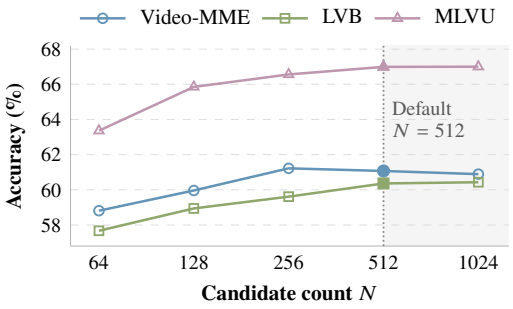


Figure 3: Q-GPS accuracy versus uniformly sampled candidate count ($B = 16$, Qwen2.5-VL-7B).

Table 2: Cross-backbone transfer under source-reported evaluation settings (accuracy, %). For each row, Base is the corresponding Uniform result and Δ is the gain over that Base. ∗ denotes previously reported results. ⋄ denotes additional textual narratives. # denotes a training-based method, and § denotes an iterative method using at most 32 frames.

MLLM	Method	Venue	Frames	V-MME (w/o subs)			LongVideoBench			MLVU		
				Base	Overall	Δ	Base	Overall	Δ	Base	Overall	Δ
LLaVA-OV-7B	Frame-Voyager∗#	ICLR'25	8	53.3	57.5	+4.2	–	–	–	58.5	65.6	+7.1
	AKS∗	CVPR'25	8	54.1	58.2	+4.1	54.2	58.4	+4.2	58.6	62.9	+4.3
	MDP³∗	ICCV'25	8	53.6	59.6	+6.0	54.2	59.0	+4.8	59.3	69.8	**+10.5**
	Nar-KFC∗⋄	ICLR'26	8	53.3	57.8	+4.5	54.5	56.5	+2.0	58.5	66.2	+7.7
	WFS-SB∗	CVPR'26	8	54.1	59.3	+5.2	54.2	59.8	+5.6	58.6	67.2	+8.6
	FrameOracle∗#	ICML'26	8	53.8	57.5	+3.7	54.3	56.0	+1.7	58.4	62.9	+4.5
	Q-GPS	–	8	53.7	60.0	**+6.3**	54.2	61.3	**+7.2**	59.4	67.9	+8.4
InternVL3-8B	AKS∗	CVPR'25	32	65.6	66.3	+0.7	58.5	61.5	+3.0	68.4	74.2	+5.8
	MDP³∗	ICCV'25	32	65.6	66.8	+1.2	58.3	60.9	+2.6	68.4	74.0	+5.6
	A.I.R.∗	ICLR'26	$\leq 32^{\S}$	65.6	68.2	**+2.6**	58.3	62.8	+4.5	68.4	74.5	+6.1
	WFS-SB∗	CVPR'26	32	65.6	67.4	+1.8	58.5	62.9	+4.4	68.4	74.8	**+6.4**
	Q-GPS	–	32	65.5	67.6	+2.0	57.7	63.7	**+5.9**	68.2	74.1	+5.9

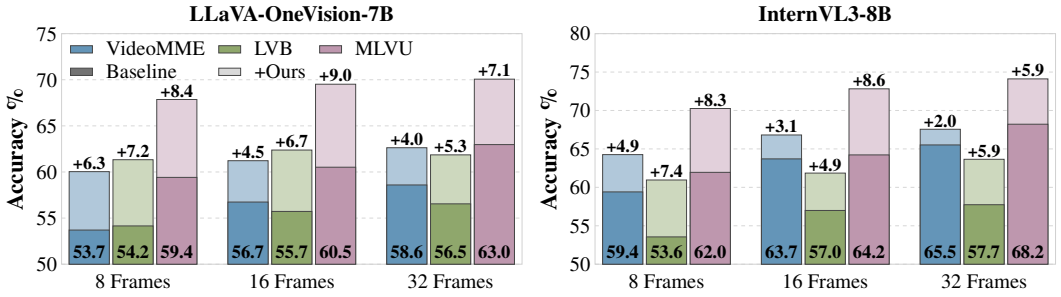


Figure 4: Cross-backbone transfer across frame budgets. Dark bars show Uniform accuracy, while light extensions show the improvement from Q-GPS; labels report the corresponding gains. LVB denotes LongVideoBench.

Component ablation. Tab. 3 evaluates each component through controlled ablations. The *w/o Temporal Ordering* variant permutes the candidate order before constructing the path-spectral covariance in Eq. (5). The *w/o Query Weighting* variant replaces the target weights with $w_j = 1/N$ in Eq. (7), retaining relevance calibration. The *w/o Covariance Update* variant fixes $\Sigma^{(b)} = K_T$ instead of applying Eq. (11). The *w/o Relevance Calibration* variant removes r_i from Eq. (10), retaining the query-weighted coverage gain. The *w/o Cross-location Gain* variant retains only the $j = i$ term in Eq. (9), keeping relevance calibration and covariance updates unchanged. The *w/o Coverage Gain (Top-K)* variant removes $g_b(i)$ from Eq. (10) and selects frames by relevance alone. All variants score below the full method across the three benchmarks. Further ablations (Appendices C.1–C.3) show that visual and hybrid graphs do not consistently improve accuracy over the temporal path across benchmarks. These results support retaining the temporal path, which admits a closed-form spectral basis and requires no additional visual-feature extraction.

Table 3: Component ablation with Qwen2.5-VL-7B at $B = 16$ (accuracy, %). Darker red marks larger drops from full Q-GPS.

Variant	V-MME	LVB	MLVU
Full Q-GPS	**61.07**	**60.36**	**66.99**
Path-spectral temporal covariance			
w/o Temporal Ordering	58.93	59.91	65.82
Query-weighted residual objective			
w/o Query Weighting	59.63	59.09	66.23
Progressive frame selection			
w/o Covariance Update	57.74	58.94	65.96
w/o Relevance Calibration	59.26	57.14	63.38
w/o Cross-location Gain	60.37	59.84	66.09
w/o Coverage Gain (Top-K)	58.44	59.54	65.05

Hyperparameter and scorer sensitivity. Tab. 4 examines c_{rank}, ν, and τ, which control temporal resolution, conditioning strength, and the concentration of query weights, respectively. Q-GPS maintains stable performance under moderate parameter changes. Increasing c_{rank} beyond 2 yields no consistent gain. The best settings differ across benchmarks, while the shared configuration remains competitive. Across nine frozen scorers, Q-GPS also improves over Uniform in all 27 scorer–benchmark combinations (Tab. 5). Thus, the gains extend beyond the default BLIP-2 scorer.

Table 4: Hyperparameter sensitivity with Qwen2.5-VL-7B at $B = 16$ (accuracy, %). Boxes mark defaults. Green/red indicates gains/losses from defaults.

(a) Spectral-rank multiplier

c_{rank}	0.5	1	**2**	3	4	8
V-MME	60.5	60.4	61.1	61.1	60.2	60.6
MLVU	67.3	66.5	67.0	67.6	67.2	67.0

(b) Observation noise

ν	0.001	0.01	**0.05**	0.1	0.5	1
V-MME	60.9	61.1	61.1	60.6	60.8	60.0
MLVU	68.1	66.8	67.0	66.7	66.9	67.4

(c) Relevance temperature

τ	0.05	0.10	**0.15**	0.20	0.30	0.50
V-MME	59.6	60.9	61.1	61.0	59.9	59.9
MLVU	67.0	67.2	67.0	66.9	67.1	67.4

Table 5: Frozen-scorer robustness at $B = 16$ with Qwen2.5-VL-7B (accuracy, %). Green bars show accuracy gains over Uniform (full width: 13 percentage points). The shaded row marks the default scorer; all other settings are held fixed.

Frozen scorer	V-MME	LVB	MLVU
Uniform	56.59	54.23	56.09
CLIP-B/16 (Radford et al., 2021)	60.63	58.19	65.90
CLIP-L/14 (Radford et al., 2021)	60.70	59.31	68.52
LongCLIP-L/14 (Zhang et al., 2025a)	60.30	**60.36**	67.53
SigLIP-SO400M (Zhai et al., 2023)	60.07	58.56	**68.66**
SigLIP2-224 (Tschannen et al., 2025)	59.89	57.59	66.36
SigLIP2-384 (Tschannen et al., 2025)	58.56	57.59	65.85
BLIP-ITM-B (Li et al., 2022)	**61.48**	59.76	67.49
BLIP-ITM-L (Li et al., 2022)	61.15	**60.36**	67.81
BLIP-2-ITM (Li et al., 2023)	61.07	**60.36**	66.99

Selection behavior. Fig. 5 summarizes the selected sets under the shared candidate pool and frozen BLIP-2 scorer. Temporal span measures the distance between the earliest and latest selected candidate locations; mean relevance averages the selected frames' raw scorer outputs. We normalize span by Uniform and mean relevance by Top-K within each benchmark, then average the normalized values equally across the three benchmarks. On this aggregate, Q-GPS retains 80.0% of Uniform's span and 91.2% of Top-K's mean relevance. It spans a wider temporal range than Top-K, while preserving higher relevance than AKS and MDP^3, which spread their selections more broadly. The complete benchmark-level comparisons and additional coverage statistics are provided in the supplementary material (Appendix D.1).

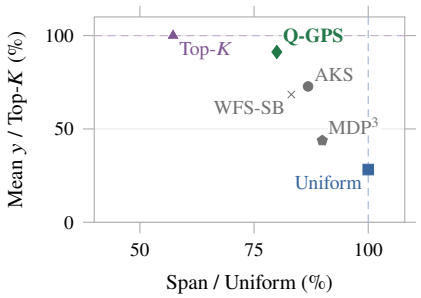


Figure 5: Mean selection behavior across three benchmarks. Each benchmark averages $B \in \{8, 16, 32\}$; span and mean relevance are normalized to Uniform and Top-K, respectively, before equal-weight averaging.

4.4 Efficiency Analysis

Selector latency is measured on CPU from precomputed relevance scores; it excludes video decoding, frozen scoring, and downstream MLLM inference. Fig. 6 compares selector-only latency and accuracy as the frame budget increases from 8 to 16 to 32 with Qwen2.5-VL-7B. Accuracy is the equal-weight mean of the three benchmark-level overall scores in Tab. 1; latency averages the three benchmark-level medians, each computed from three full-split runs. The supplementary material describes the measurement environment and reports the total end-to-end pipeline cost (Appendix C.5).

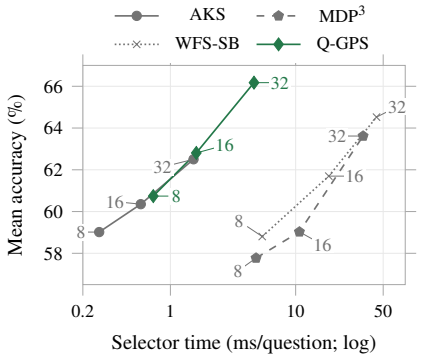


Figure 6: Accuracy–latency trajectories with Qwen2.5-VL-7B. Point labels denote frame budgets $B = 8, 16, 32$. Both axes average three benchmarks; time uses a log scale.

5 Conclusion

We presented Q-GPS, a training-free frame selector for long-video question answering. It combines direct relevance with query-weighted coverage gains over unselected locations. A path-spectral covariance initializes a residual temporal coverage surrogate, with exact low-rank updates enabling progressive candidate reassessment. Controlled comparisons across three benchmarks show accuracy gains over competing selectors, while improvements over uniform sampling persist across budgets and downstream MLLMs. These results support combining query relevance with an explicit residual temporal coverage state for fixed-budget frame selection.

AI USE STATEMENT

We used generative AI tools to assist with writing and polishing the manuscript, including improving clarity, grammar, and readability. We also used these tools for information retrieval and literature discovery, including identifying potentially relevant prior work. These uses supported manuscript preparation and literature search. The authors take responsibility for the final text, citations, and scientific content of this paper.

REPRODUCIBILITY STATEMENT

We provide an anonymous code repository at https://anonymous.4open.science/r/Q-GPS-FEB4/ to support reproducibility. Sec. 3 and Algorithm 1 describe the frame-selection method, while Appendix A provides supporting derivations and implementation details. Sec. 4 and Appendix B document the experimental settings, including compute resources, dataset splits, candidate-frame construction, frozen relevance scoring, and downstream evaluation. Appendix E provides the inference prompt templates.

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.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. Wavelet-based frame selection by detecting semantic boundary for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24052–24061, June 2026.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. doi: 10.1613/jair.295.

Bo Fang, Yuxin Song, Haoyuan Sun, Qiangqiang Wu, Wenhao Wu, and Antoni B. Chan. Threading keyframe with narratives: MLLMs as strong long video comprehenders. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kyLS9EhPhY.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712, June 2025.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. URL http://jmlr.org/papers/v9/krause08a.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. 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://arxiv.org/abs/2510.03584.

Jialuo Li, Bin Li, Jiahao Li, and Yan Lu. Divide, then ground: Adapting frame selection to query types for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11369–11380, June 2026b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3318–3327, June 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013. doi: 10.1109/MSP.2012.2235192.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101, October 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, June 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Henaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/329ad516cf7a6ac306f29882e9c77558-Paper-Datasets_and_Benchmarks_Track.pdf.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591, June 2025.

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*, 2025. URL https://openreview.net/forum?id=LNL7zKvm7e.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 310–325, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72983-6.

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*, pp. 881–916, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.51. URL https://aclanthology.org/2025.findings-naacl.51/.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

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

APPENDIX CONTENTS

A Additional Method Details **15**

A.1 Closed-Form Path-Spectral Construction 15

A.2 Exact Low-Rank Implementation 16

A.3 Tie-Breaking and Numerical Safeguards 18

A.4 Computational Complexity 18

B Controlled Evaluation Details **19**

B.1 Experimental Compute Resources 19

B.2 Dataset Details 19

B.3 Implementation Details 20

C Additional Experiments and Analysis **20**

C.1 Graph Structure for the Initial Covariance 21

C.2 Temporal Graph Construction 21

C.3 Uniform versus Query-Conditioned Spectrum 21

C.4 Statistical Significance Analysis 22

C.5 End-to-End Pipeline Cost 24

D Quantitative Analysis and Qualitative Visualization **25**

D.1 Quantitative Analysis for Frame Selection 25

D.2 Qualitative Visualization for Frame Selection 27

E Inference Prompt Templates **35**

A Additional Method Details

This section derives the temporal-path basis, proves the exact low-rank implementation, and specifies tie-breaking and numerical safeguards. It then derives the time and memory complexity of the complete spectral-coordinate selection procedure. We follow the notation of the main paper. Here N is the number of ordered candidates, B is the output frame budget, $\mathcal{I}_b$ and $\mathcal{U}_b = [N] \setminus \mathcal{I}_b$ are the selected and unselected locations after b selections, and $\Sigma^{(b)}$ is the residual covariance. The vectors $\mathbf{r}$ and $\mathbf{w}$ denote normalized candidate relevance and softmax target-location weights, respectively, and $\nu > 0$ is the observation-noise variance.

A.1 Closed-Form Path-Spectral Construction

For $N \geq 2$, let $A_{ij} = \mathbb{I}[|i-j| = 1]$ be the adjacency of the unweighted temporal path and let $D = \mathrm{diag}(d_1, \ldots, d_N)$, where $d_1 = d_N = 1$ and $d_i = 2$ for $2 \leq i \leq N-1$. As in the main paper, the symmetric normalized Laplacian is

$$L = I_N - D^{-1/2} A D^{-1/2}. \tag{12}$$

To construct the retained path basis U_m without numerical eigendecomposition, we first solve the eigenproblem of L in closed form.

Proposition A.1.1 (Closed-form eigendecomposition of the normalized path)**.** *For the normalized Laplacian L in Eq. (12) and $\ell = 1, \ldots, N$, define*

$$\theta_\ell = \frac{(\ell-1)\pi}{N-1}. \tag{13}$$

The corresponding eigenvalue is $\lambda_\ell = 1 - \cos\theta_\ell$. The normalized eigenvector associated with λ_ℓ has entries

$$\begin{aligned} [\mathbf{u}_\ell]_i &= \frac{\sqrt{d_i}}{c_\ell} \cos\big((i-1)\theta_\ell\big), \\ c_\ell^2 &= \sum_{j=1}^{N} d_j \cos^2\big((j-1)\theta_\ell\big). \end{aligned} \tag{14}$$

For every ℓ, these quantities satisfy

$$L\mathbf{u}_\ell = \lambda_\ell \mathbf{u}_\ell.$$

The eigenvectors $\{\mathbf{u}_\ell\}_{\ell=1}^N$ form an orthonormal basis, with $0 = \lambda_1 < \cdots < \lambda_N = 2$, and are therefore ordered from low to high temporal frequency.

Proof. Write $\mathbf{u} = D^{1/2}\mathbf{v}$ to remove the degree-normalization factors from the eigenvalue equation. Substitution gives

$$\begin{aligned} L\mathbf{u} = \lambda\mathbf{u} &\iff D^{1/2}\mathbf{v} - D^{-1/2}A\mathbf{v} = \lambda D^{1/2}\mathbf{v} \\ &\iff A\mathbf{v} = (1-\lambda)D\mathbf{v}. \end{aligned} \tag{15}$$

Since the spectrum of a normalized Laplacian is contained in $[0, 2]$, $\mu = 1 - \lambda \in [-1, 1]$; write $\mu = \cos\theta$ with $\theta \in [0, \pi]$.

At an interior location $2 \leq i \leq N-1$, the ith row of Eq. (15) becomes

$$v_{i-1} + v_{i+1} = 2\mu v_i. \tag{16}$$

For $\theta \in (0, \pi)$, the real solutions of Eq. (16) are $v_i = \alpha\cos((i-1)\theta) + \beta\sin((i-1)\theta)$. The left endpoint condition $v_2 = \mu v_1$ gives $\beta\sin\theta = 0$, and hence $\beta = 0$. The endpoint cases $\theta \in \{0, \pi\}$ directly give the same cosine-form modes. After fixing the arbitrary scale, we may therefore take $v_i = \cos((i-1)\theta)$. The right endpoint condition $v_{N-1} = \mu v_N$ then reduces to

$$\sin\big((N-1)\theta\big)\sin\theta = 0. \tag{17}$$

Within $[0, \pi]$, the roots of $\sin((N-1)\theta) = 0$ are precisely $\theta_\ell = (\ell-1)\pi/(N-1)$ for $\ell = 1, \ldots, N$; the additional factor $\sin\theta$ introduces no other roots. Thus the boundary conditions admit exactly the N angles in Eq. (13).

For these angles, $\lambda_\ell = 1 - \cos\theta_\ell$. Since $1 - \cos\theta$ is strictly increasing on $[0, \pi]$, $0 = \lambda_1 < \cdots < \lambda_N = 2$. The eigenvalues are distinct, so the corresponding eigenvectors of the real symmetric matrix L are orthogonal. Transforming back through $\mathbf{u} = D^{1/2}\mathbf{v}$ and applying the normalization in Eq. (14) gives N orthonormal eigenvectors and hence a basis of $\mathbb{R}^N$.

To identify its frequency order, expand Eq. (12) over the path edges. For any $\mathbf{x} \in \mathbb{R}^N$, this gives

$$\mathbf{x}^\top L\mathbf{x} = \sum_{i=1}^{N-1} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_{i+1}}{\sqrt{d_{i+1}}} \right)^2 \geq 0.$$

This is the total squared normalized variation between adjacent locations and also shows $L \succeq 0$. For a unit eigenvector it equals $\mathbf{u}_\ell^\top L\mathbf{u}_\ell = \lambda_\ell$; hence the increasing eigenvalue order is precisely the low-to-high path-frequency order. The closed form gives the same order. Indeed, θ_ℓ increases with ℓ, so its cosine oscillates increasingly rapidly along the ordered path. □

Because the modes are ordered from low to high frequency, $\mathrm{span}(U_m)$ contains every path mode up to the cutoff λ_m. Increasing m therefore adds finer temporal variation without removing the coarser modes already represented. We tie m to the output budget B, so larger budgets admit finer temporal resolution. The following corollary uses the scaled orthogonal projector onto this subspace as the initial temporal covariance.

Corollary A.1.2 (Uniform path-spectral covariance)**.** *Let*

$$\begin{aligned} m &= \min\{N, \lceil c_{\mathrm{rank}} B \rceil\}, \\ U_m &= [\mathbf{u}_1, \ldots, \mathbf{u}_m]. \end{aligned} \tag{18}$$

Assigning equal mass $p_\ell = 1/m$ to the retained modes gives

$$\begin{aligned} K_T &= N U_m \,\mathrm{diag}(\mathbf{p}) U_m^\top \\ &= \frac{N}{m} U_m U_m^\top, \end{aligned} \tag{19}$$

where $K_T \succeq 0$ and $\mathrm{tr}(K_T) = N$.

Proof. For any $\mathbf{z} \in \mathbb{R}^N$,

$$\mathbf{z}^\top K_T \mathbf{z} = \frac{N}{m} \left\| U_m^\top \mathbf{z} \right\|_2^2 \geq 0.$$

Thus $K_T \succeq 0$ and is a valid covariance. Every linear combination has nonnegative variance, with equality exactly when $U_m^\top \mathbf{z} = \mathbf{0}$, so $\mathrm{range}(K_T) = \mathrm{span}(U_m)$, $\mathrm{rank}(K_T) = m$, and directions outside the retained subspace receive no variance. Orthonormality also gives $\mathrm{tr}(K_T) = (N/m)\,\mathrm{tr}(U_m^\top U_m) = N$, so the average initial marginal variance is one. □

Eq. (19) gives

$$(K_T)_{ij} = \frac{N}{m} \sum_{\ell=1}^{m} [\mathbf{u}_\ell]_i [\mathbf{u}_\ell]_j.$$

Thus the retained-mode inner product determines how strongly two locations are temporally coupled. Under the conditioning rule in Appendix A.2, selecting one location discounts the residual coverage of other coupled locations and thereby suppresses repeated coverage of the same smooth temporal structure. With equal mass, retaining all N modes would instead give $K_T = I_N$ and eliminate cross-location coupling, so the low-pass truncation is a modeling choice rather than only a computational approximation. Equal mass among the retained modes introduces no further spectral preference.

A.2 Exact Low-Rank Implementation

The dense covariance in the main paper makes the progressive update explicit. In implementation, however, the residual state need not be stored as an $N \times N$ matrix. For any residual covariance supported on the retained subspace $\mathrm{span}(U_m)$, define its spectral-coordinate state as

$$P_b = U_m^\top \Sigma^{(b)} U_m. \tag{20}$$

The matrix $P_b \in \mathbb{R}^{m\times m}$. Because $U_m^\top U_m = I_m$, this representation is unique and satisfies $\Sigma^{(b)} = U_m P_b U_m^\top$. Thus P_b is the covariance of the retained path-mode coefficients and compactly represents $\Sigma^{(b)}$.

Proposition A.2.1 (Exact spectral-coordinate conditioning)**.** *The initial spectral-coordinate state is*

$$P_0 = \frac{N}{m} I_m, \tag{21}$$

where I_m is the $m \times m$ identity matrix. Hence $\Sigma^{(0)} = U_m P_0 U_m^\top = K_T$.

Suppose that $\Sigma^{(b)} = U_m P_b U_m^\top$. For candidate location i, let $\mathbf{e}_i \in \mathbb{R}^N$ be the ith standard basis vector, and write $\mathbf{h}_i = U_m^\top \mathbf{e}_i \in \mathbb{R}^m$ for its coordinates in the retained path basis. Conditioning on candidate i updates the spectral-coordinate state as

$$P_{b|i} = P_b - \frac{P_b \mathbf{h}_i \mathbf{h}_i^\top P_b}{\mathbf{h}_i^\top P_b \mathbf{h}_i + \nu}. \tag{22}$$

The corresponding dense covariance is exactly $\Sigma^{(b|i)} = U_m P_{b|i} U_m^\top$. Consequently, if $s = i_{b+1}$ is selected, let $\mathbf{q}_b = P_b \mathbf{h}_s$ denote the residual cross-covariance between the retained mode coefficients and location s. The next spectral-coordinate state is

$$P_{b+1} = P_b - \frac{\mathbf{q}_b \mathbf{q}_b^\top}{\mathbf{h}_s^\top \mathbf{q}_b + \nu}. \tag{23}$$

Proof. Substitute $\Sigma^{(b)} = U_m P_b U_m^\top$ into the dense conditioning rule and use $\mathbf{h}_i = U_m^\top \mathbf{e}_i$. Expanding each substitution in order gives

$$\begin{aligned}
\Sigma^{(b|i)} &= \Sigma^{(b)} - \frac{\Sigma^{(b)}_{:,i} \Sigma^{(b)}_{i,:}}{\Sigma^{(b)}_{ii} + \nu} \\
&= U_m P_b U_m^\top - \frac{\left(U_m P_b U_m^\top \mathbf{e}_i\right)\left(\mathbf{e}_i^\top U_m P_b U_m^\top\right)}{\mathbf{e}_i^\top U_m P_b U_m^\top \mathbf{e}_i + \nu} \\
&= U_m \left(P_b - \frac{P_b \mathbf{h}_i \mathbf{h}_i^\top P_b}{\mathbf{h}_i^\top P_b \mathbf{h}_i + \nu} \right) U_m^\top,
\end{aligned} \tag{24}$$

which proves Eq. (22). The base case follows from Eqs. (19) and (21); induction over b therefore shows that every conditioned covariance remains in the retained subspace and that Eq. (20) is well defined at every step. Eq. (23) is the selected-candidate case. Thus the spectral-coordinate calculation is algebraically equivalent to dense conditioning. □

Corollary A.2.2 (Low-rank gain evaluation)**.** *Let $\mathbf{1}_{\mathcal{U}_b}$ be the indicator vector of the current unselected set. Define*

$$\begin{aligned}
M_b &= \operatorname{diag}\left(\mathbf{w} \odot \mathbf{1}_{\mathcal{U}_b}\right), \\
G_b &= U_m^\top M_b U_m.
\end{aligned} \tag{25}$$

Here M_b contains the target weights on the unselected locations, and G_b is its $m \times m$ compression in the retained path basis. Then, for every $i \in \mathcal{U}_b$,

$$g_b(i) = \frac{\mathbf{h}_i^\top P_b G_b P_b \mathbf{h}_i}{\mathbf{h}_i^\top P_b \mathbf{h}_i + \nu}. \tag{26}$$

Proof. Because M_b contains the target weights on $\mathcal{U}_b$ and zeros elsewhere, and because P_b is symmetric, the numerator of the dense gain is

$$\begin{aligned}
\sum_{j \in \mathcal{U}_b} w_j \left(\Sigma^{(b)}_{ji}\right)^2 &= \left(\Sigma^{(b)}_{:,i}\right)^\top M_b \Sigma^{(b)}_{:,i} \\
&= \left(U_m P_b \mathbf{h}_i\right)^\top M_b \left(U_m P_b \mathbf{h}_i\right) \\
&= \mathbf{h}_i^\top P_b U_m^\top M_b U_m P_b \mathbf{h}_i \\
&= \mathbf{h}_i^\top P_b G_b P_b \mathbf{h}_i,
\end{aligned} \tag{27}$$

which gives Eq. (26). The numerators for all locations are the diagonal of $U_m P_b G_b P_b U_m^\top$, and their marginal variances are the diagonal of $U_m P_b U_m^\top$. □

Proposition A.2.1 gives the same residual covariance when both implementations condition on the same candidate, while Eq. (26) gives the same gain for every remaining candidate. With the same deterministic tie-breaking rule and initial state $\Sigma^{(0)} = K_T$, induction over b yields identical selected sets and residual states. Thus, the dense and low-rank implementations are exactly equivalent.

A.3 Tie-Breaking and Numerical Safeguards

The primary selection score remains $a_b(i) = r_i g_b(i)$ as defined in the main paper. Its maximizer need not be unique when multiple candidates attain exactly the same score, so we specify the deterministic tie-breaking and numerical safeguards used by the implementation. These conventions do not alter the primary selection objective.

Definition A.3.1 (Deterministic lexicographic tie-breaking)**.** Let $\delta_N = \max\{N-1, 1\}$ denote the temporal-distance scale. For each $i \in \mathcal{U}_b$, define the temporal tie-breaking score

$$s_b(i) = \begin{cases} -\left|i - (N+1)/2\right| / \delta_N, & b = 0, \\ \min\limits_{k \in \mathcal{I}_b} |i-k| / \delta_N, & b > 0. \end{cases} \tag{28}$$

For each $i \in \mathcal{U}_b$, let t_i be the original frame index associated with candidate location i, and define

$$\kappa_b(i) = \big(a_b(i), g_b(i), r_i, s_b(i), -t_i\big). \tag{29}$$

Q-GPS selects the candidate whose key $\kappa_b(i)$ is lexicographically largest.

Because $a_b(i)$ is the first component of $\kappa_b(i)$, this rule returns a maximizer of the primary selection score. Subsequent components order exact ties by coverage gain, direct relevance, temporal separation, and the earlier original frame. Before any selection, the temporal tie-breaker favors the center of the candidate sequence; thereafter, it favors maximal distance from $\mathcal{I}_b$. Candidates are traversed in increasing candidate order, which resolves any remaining equality.

Boundary and numerical cases. The relevance normalization follows the main paper. When $y_{\max} - y_{\min} \le \varepsilon$, the score profile contains no location-specific relevance information. The target weights are uniform, and selection is determined by coverage gain and the remaining deterministic terms. The implementation uses $\varepsilon = 10^{-8}$ as the numerical threshold and evaluates softmax weights after subtracting their maximum input. Denominators in the conditioning and gain expressions are protected by the same numerical floor, and P_b is symmetrized after each rank-one update to suppress roundoff asymmetry. Finally, because candidate locations follow temporal order, selected locations are restored to increasing candidate index, equivalently increasing t_i, before their frames are passed to the downstream MLLM.

A.4 Computational Complexity

The exact equivalence established in Proposition A.2.1 confines all residual-covariance operations to the retained m-dimensional path subspace. We next quantify the resulting time and memory requirements of the complete selection procedure.

Proposition A.4.1 (Complexity of exact spectral-coordinate selection)**.** *Assume that $c_{\mathrm{rank}} > 0$ is fixed and $1 \le B \le N$. After the raw frame–query scores $\{y_i\}_{i=1}^N$ have been obtained, the exact spectral-coordinate implementation requires $\mathcal{O}(BNm^2)$ time and $\mathcal{O}(Nm + m^2)$ memory, where m is defined in Eq. (18). No dense $N \times N$ residual covariance is stored explicitly.*

Proof. The initialization constructs the closed-form retained path basis U_m in $\mathcal{O}(Nm)$ time, normalizes the raw relevance scores and computes the target-location weights in $\mathcal{O}(N)$ time, and initializes P_0 in $\mathcal{O}(m^2)$ time.

At selection step b, the dominant operation is the simultaneous evaluation of the coverage gains in Eq. (26). Because M_b is diagonal, constructing $G_b = U_m^\top M_b U_m$ requires $\mathcal{O}(Nm^2)$ time. Forming $P_b G_b P_b$ requires $\mathcal{O}(m^3)$ time. The numerator and denominator of $g_b(i)$ are quadratic forms in $\mathbf{h}_i \in \mathbb{R}^m$, so evaluating them for all remaining candidates $i \in \mathcal{U}_b$ requires $\mathcal{O}(Nm^2)$ time. The complete gain evaluation therefore costs $\mathcal{O}(Nm^2 + m^3)$ at one step.

The dependence on the selection index b enters through the temporal tie-breaking score. Directly computing $s_b(i)$ compares each remaining candidate with at most b selected locations, after which lexicographic maximization requires one scan over $\mathcal{U}_b$. Their combined cost is $\mathcal{O}(N(b+1))$. Once a candidate has been selected, the rank-one update of P_b requires $\mathcal{O}(m^2)$ time. Hence

$$T_b = \mathcal{O}\big(Nm^2 + m^3 + N(b+1) + m^2\big)\,. \tag{30}$$

Because $m \le N$, the $\mathcal{O}(m^3)$ term is absorbed by $\mathcal{O}(Nm^2)$. Moreover, for fixed $c_{\mathrm{rank}} > 0$ and $B \le N$, Eq. (18) implies $B = \mathcal{O}(m)$. Summing the temporal tie-breaking and selection costs over all selections gives

$$\sum_{b=0}^{B-1} \mathcal{O}\big(N(b+1)\big) = \mathcal{O}(NB^2) = \mathcal{O}(BNm),$$

which is also absorbed by $\mathcal{O}(BNm^2)$. The initialization cost, the accumulated rank-one updates, and the final $\mathcal{O}(B \log B)$ temporal ordering are lower-order terms. The complete selection procedure therefore requires $\mathcal{O}(BNm^2)$ time.

The implementation stores the retained basis U_m together with a constant number of $m \times m$ spectral-coordinate matrices. The relevance vectors, candidate indices, and selected and unselected index sets require only $\mathcal{O}(N)$ additional memory, which is absorbed because $m \ge 1$. The total memory requirement is therefore $\mathcal{O}(Nm + m^2)$. □

Tab. 6 summarizes the initialization, per-step, and postselection costs established above.

Table 6: Asymptotic time complexity of the exact spectral-coordinate selection procedure. Raw frame–query scores $\{y_i\}_{i=1}^{N}$ are treated as inputs.

Phase	**Operation**	**Time complexity**
Initialization	Relevance normalization and target weights	$\mathcal{O}(N)$
	Closed-form path basis U_m	$\mathcal{O}(Nm)$
	Initial state P_0	$\mathcal{O}(m^2)$
Per selection step	Coverage gains for remaining candidates	$\mathcal{O}(Nm^2 + m^3)$
	Temporal tie-breaking and lexicographic selection	$\mathcal{O}(N(b+1))$
	Rank-one update of P_b	$\mathcal{O}(m^2)$
Postselection	Temporal ordering of selected frames	$\mathcal{O}(B \log B)$
Overall	**Complete B-step selection**	$\boldsymbol{\mathcal{O}(BNm^2)}$

Since $m \le N$, the m^3 term is absorbed by Nm^2.

B Controlled Evaluation Details

This section describes the compute environment, benchmark settings, and implementation shared by all controlled frame-selection methods in the main paper. Within each benchmark and frame budget, all controlled methods share the candidate pool, frozen query–frame scores when applicable, and downstream evaluation protocol, so that only the frame-selection rule changes.

B.1 Experimental Compute Resources

Frozen relevance scoring and downstream MLLM inference are performed on NVIDIA RTX 4090 GPUs with 24 GB of memory, using up to eight GPUs to run experiments in parallel. Q-GPS selection latency is measured on an Intel Xeon Platinum 8336C with 16 parallel worker processes. The same hardware configuration is used across all controlled accuracy comparisons.

B.2 Dataset Details

Video-MME. Video-MME contains 900 videos spanning six visual domains and provides three expert-annotated multiple-choice questions for each video. We evaluate all 2,700 test questions

without subtitles. The reported result is sample-level multiple-choice accuracy, and generation allows at most 16 new tokens.

LongVideoBench. It evaluates long-form video-language understanding with temporally grounded multiple-choice questions over long web videos. We use all 1,337 questions in the validation split without subtitles. We report sample-level multiple-choice accuracy and allow at most 32 new tokens.

MLVU. MLVU is a multi-task long-video benchmark covering diverse video genres and reasoning requirements. We evaluate all 2,174 development questions from its seven multiple-choice tasks without subtitles. The reported score is the unweighted mean of the seven task accuracies, and generation allows at most 16 new tokens.

The dataset splits and metrics above are fixed across all controlled runs.

B.3 Implementation Details

Each method maps the shared candidate pool to a selected index set $\mathcal{I}_B$ for downstream evaluation, using the frozen relevance scores when required.

B.3.1 Candidate construction

For a video with F indexed source frames, the actual candidate count in the default controlled setting is

$$N = \min\{F, 512\}. \tag{31}$$

We take N distinct integer frame indices uniformly over the closed interval from the first to the last frame. Videos with at least 512 frames therefore use the shared 512-candidate setting, while shorter videos use every available frame index. All methods receive the same ordered candidate sequence for a given video.

B.3.2 Frozen relevance scoring

The frozen scorer is `Salesforce/blip2-itm-vit-g`. For each multiple-choice item, the question and all answer options are concatenated in their original order to form one text query. Every candidate is paired with this query and evaluated by the BLIP-2 image–text matching head. This implements the frozen scorer ϕ defined in the main paper. If $\mathbf{z}_i \in \mathbb{R}^2$ contains the non-match and match logits for candidate i, its raw score is

$$y_i = \phi(Q, x_i) = \big[\mathrm{softmax}(\mathbf{z}_i)\big]_2. \tag{32}$$

The scorer uses its native image processor with padding and text truncation. All candidates are scored in half precision on CUDA before selection. The resulting raw-score vector $\mathbf{y} = (y_i)_{i=1}^N$ is shared by methods that use query–frame scores.

B.3.3 Downstream inference

Once a method returns $\mathcal{I}_B$, we restore the selected source frames to chronological order and encode them as a single video input. All controlled methods are evaluated with the same Qwen2.5-VL-7B-Instruct model through LMMs-Eval. The benchmark task implementation determines the text prompt, answer extraction, and scoring rule. The corresponding prompt templates are listed in Appendix E. Only the selected visual input changes across methods, while the downstream inference and evaluation procedure remain fixed.

Each selected frame is processed with `min_pixels` $= 25{,}088$ and `max_pixels` $= 100{,}352$. The model is loaded in bfloat16 with FlashAttention-2 and evaluated with batch size one. Decoding is greedy, with temperature zero and sampling disabled, and follows the dataset-specific token limits in Appendix B.2.

C Additional Experiments and Analysis

This section evaluates three design choices in Q-GPS, tests the statistical significance of its gains over the baselines in Tab. 1 of the main paper, and reports the end-to-end cost of the complete pipeline.

Unless otherwise stated, all comparisons follow the controlled protocol in Appendix B, and accuracy is reported in %.

C.1 Graph Structure for the Initial Covariance

Q-GPS initializes $\Sigma^{(0)} = K_T$, whose off-diagonal entries encode how conditioning on a selected location reduces residual coverage at coupled unselected locations. The graph used for K_T determines how selection affects other locations. Candidate order and visual appearance provide two plausible sources. A temporal path represents local continuity, whereas DINOv2 (Oquab et al., 2024) similarity can connect repeated scenes across distant locations.

The temporal graph is the unit-weight path over the ordered candidate locations. For the visual graph, a frozen DINOv2 image encoder maps each candidate frame to an ℓ_2-normalized, query-independent appearance descriptor. We connect each candidate to its eight nearest neighbors using locally scaled similarity weights and symmetrize the graph. The combined graph adds the temporal and visual adjacencies. Each graph is converted to a symmetric normalized Laplacian, and its first m eigenvectors are used in the same uniform-spectrum retained-mode covariance construction and selection procedure. The candidate frames, BLIP-2 relevance scores, retained rank, downstream model, and evaluation splits remain fixed, so only the graph used to construct $\Sigma^{(0)}$ changes.

Relative to the temporal path, the visual graph changes budget-averaged accuracy by -0.59%, $+0.72\%$, and -0.77% on Video-MME, LongVideoBench, and MLVU, respectively. Links between distant but visually similar frames help only on LongVideoBench.

Adding the visual edges to the path changes the three dataset averages by -0.31%, $+0.22\%$, and -0.01%. The combined graph outperforms the path in three of the nine cells, matches it in one, and is lower in five. Its overall mean is 63.21%, compared with 63.24% for the temporal path and 63.03% for the visual graph. Thus, adding visual-similarity edges does not consistently improve upon the temporal path. We therefore use the temporal path to construct K_T because it preserves the closed-form basis in Proposition A.1.1 and avoids both DINOv2 feature extraction and a general-graph eigendecomposition.

C.2 Temporal Graph Construction

Having selected the temporal path, we test chronological alignment and neighborhood radius. Random uses a fixed deterministic permutation of candidates, preserving the one-hop topology and spectrum but breaking temporal alignment.

The chronological constructions connect locations within one, two, or three hops. For radius ρ, candidate locations at distance $\ell \le \rho$ are connected with inverse-distance weights normalized over the retained hops,

$$A_{ij}^{(\rho)} = \sum_{\ell=1}^{\rho} \frac{\ell^{-1}}{\sum_{a=1}^{\rho} a^{-1}} \mathbb{I}[|i-j| = \ell]. \tag{33}$$

Aligning the one-hop path with chronological candidate order improves all nine dataset–budget cells over Random, with gains from 0.45% to 3.44%. The overall mean rises from 61.50% to 63.24%. Because Random keeps the same graph and spectrum, its lower accuracy shows that graph neighbors should follow temporal order.

Expanding the chronological neighborhood produces smaller changes. Relative to one hop, the largest absolute change is 0.60% for two hops and 0.90% for three hops. The one-hop, two-hop, and three-hop variants average 63.24%, 63.10%, and 62.92%, respectively, over all nine cells. Two- and three-hop graphs provide local gains at low budgets on Video-MME and MLVU, while the one-hop graph is highest on all three datasets at $B = 32$. Extending each location beyond its immediate neighbors therefore yields small, dataset-dependent changes without a systematic gain.

C.3 Uniform versus Query-Conditioned Spectrum

With the graph source and temporal topology fixed, the remaining question is whether the query should reshape the spectral mass within U_m. Q-GPS uses $p_\ell = 1/m$, assigning equal mass to all

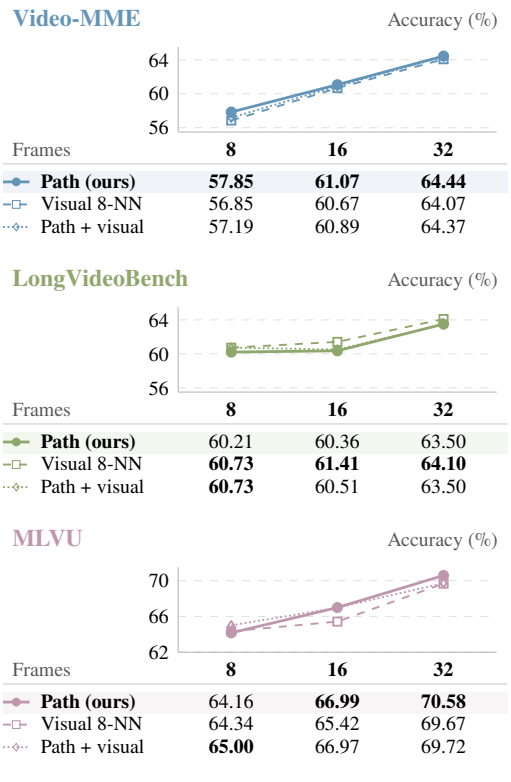


Figure 7: Graph structure for the initial covariance (accuracy, %). The visual graph uses symmetric 8-nearest-neighbor links with locally scaled similarity weights on normalized DINOv2 descriptors. Solid lines and shaded rows mark the temporal path; bold values mark column maxima, including ties.

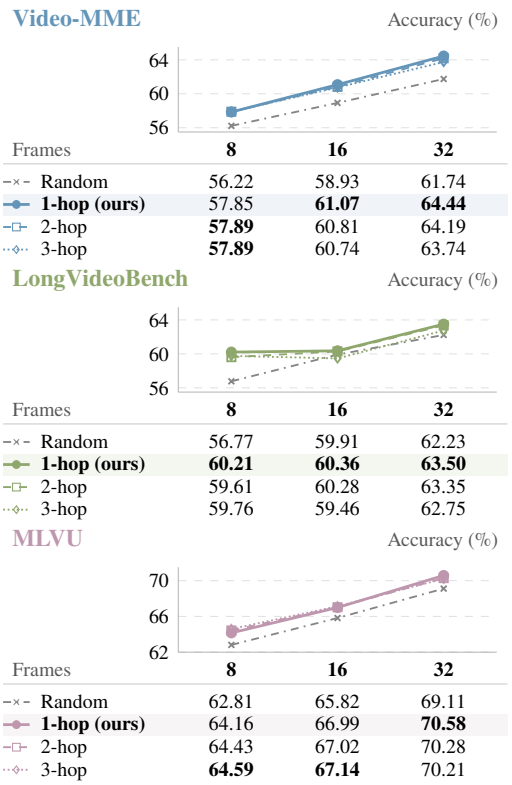


Figure 8: Temporal-topology comparison (accuracy, %). Random permutes the candidate-to-path assignment; chronological variants use one-, two-, or three-hop neighborhoods. Solid lines and shaded rows mark the one-hop default; bold values mark column maxima, including ties.

retained modes and the same temporal covariance to every query. The target weights $\mathbf{w}$ and candidate relevance $\mathbf{r}$ instead introduce query guidance during selection. The comparison variant projects the relevance profile onto the same basis and forms a query-conditioned spectrum,

$$e_\ell(q) = \left(\mathbf{u}_\ell^\top \mathbf{r}(q)\right)^2,$$
$$p_\ell(q) = \frac{e_\ell(q) + \eta_s}{\sum_{a=1}^{m}\left(e_a(q) + \eta_s\right)}, \tag{34}$$

where $\eta_s = 10^{-4}$ was the stabilizing floor used by that variant. Both constructions retain the same low-frequency subspace. They differ only in whether the initial variance is equal across retained modes or reshaped by the query. The target weights $\mathbf{w}$, candidate relevance $\mathbf{r}$, selection updates, and evaluation settings remain fixed, so the comparison isolates the effect of additionally using the query to reshape K_T.

The query-conditioned spectrum is higher in six of the nine cells, but its 63.35% overall mean is only $+0.11\%$ above the uniform spectrum's 63.24%. Thus, query-conditioned spectral mass changes individual cells but provides negligible aggregate gain. It uses relevance in both K_T and the selection score, so query information enters twice. We use the uniform spectrum so that query guidance enters only through $\mathbf{w}$ and $\mathbf{r}$ during selection.

C.4 Statistical Significance Analysis

We jointly analyze the nine dataset–budget settings in Tab. 1 of the main paper for every reported baseline. For method m, let $c_{d,b,i}^{(m)} \in \{0, 1\}$ denote whether question i is answered correctly on dataset

Table 7: Uniform versus query-conditioned spectra (accuracy, %). Uniform is the shared default; the query-conditioned variant uses $\eta_s = 10^{-4}$. Green/red cells indicate higher/lower accuracy than the corresponding default. Bold scores mark paired maxima.

Dataset	Spectrum	Frame budget $B = 8$	$B = 16$	$B = 32$
Video-MME	Uniform (default)	**57.85**	61.07	**64.44**
	Query-conditioned	57.41	**61.26**	64.26
LongVideoBench	Uniform (default)	60.21	60.36	63.50
	Query-conditioned	**60.28**	**60.88**	**64.40**
MLVU	Uniform (default)	**64.16**	66.99	70.58
	Query-conditioned	63.55	**67.27**	**70.81**

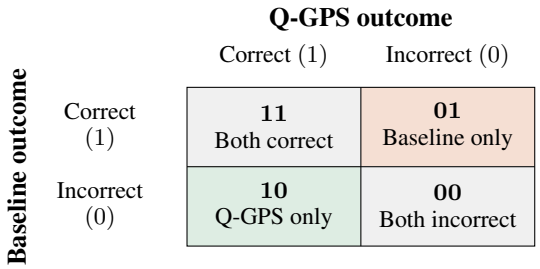


Figure 9: Paired correctness outcomes for Q-GPS and a baseline.

d with frame budget b. The Q-GPS and baseline outcomes for the same question form one paired observation, summarized in Fig. 9 with the Q-GPS outcome first.

The 11 and 00 cells are concordant pairs and do not favor either method. The discordant cells 10 and 01 denote a Q-GPS-only correct answer and a baseline-only correct answer, respectively. For a Video-MME or LongVideoBench dataset–budget setting with n questions, the paired accuracy difference is $100(n_{10} - n_{01})/n$ in %. Pairing controls for question difficulty, and only the discordant outcomes affect the accuracy difference.

Video-MME and LongVideoBench average the paired correctness difference over all questions. MLVU first computes the paired difference within each of its seven task types and then macro-averages the task-level differences, matching its reported accuracy metric. This gives a cell-level accuracy difference $\Delta_{m,d,b}$ between Q-GPS and baseline m. We define the joint effect as the equal-weight average over the three datasets and three budgets,

$$\overline{\Delta}_m = \frac{1}{9} \sum_{d \in \mathcal{D}} \sum_{b \in \{8,16,32\}} \Delta_{m,d,b}. \tag{35}$$

The 95% confidence intervals use 100,000 nonparametric video-cluster bootstrap resamples, with each sampled video carrying all of its questions and frame-budget outcomes. For MLVU, every resample recomputes the seven task-level differences and averages them equally, using the same bootstrap count for each video throughout. We test the null hypothesis that neither method has a systematic paired advantage using 100,000 video-level sign flips. Each repetition assigns one random sign to all differences from the same video before recomputing Eq. (35). Both procedures preserve within-video and cross-budget dependence, and the resulting p-values are Holm-corrected across the five baselines.

Tab. 8 shows that Q-GPS improves over every baseline on average, from 1.57% over WFS-SB to 7.68% over Uniform. All confidence intervals exclude zero, $p < 0.001$ after Holm correction for every comparison, and $n_{10} > n_{01}$ throughout.

Table 8: Video-clustered paired comparison across all nine dataset–budget settings. Panel (a) reports baseline accuracy, the mean Q-GPS gain and its 95% CI in %, and Holm-adjusted p; the dotted zero line denotes no gain. Panel (b) groups the descriptive counts by agreement and disagreement, with Q-GPS correctness first in each code.

(a) Effect estimates and uncertainty

Baseline	Baseline acc. (%)	Gain [95% CI] (%)	Gain (%) 0 4 8	Holm p
Uniform	55.56	+7.68 [6.76, 8.60]		< 0.001
AKS	60.63	+2.62 [1.96, 3.28]		< 0.001
MDP^3	60.14	+3.10 [2.27, 3.94]		< 0.001
WFS-SB	61.68	+1.57 [0.81, 2.32]		< 0.001
DIG	59.54	+3.70 [2.83, 4.56]		< 0.001

(b) Descriptive paired outcomes

	Agreement		Disagreement	
Baseline	Both correct 11	Both incorrect 00	Q-GPS only 10	Baseline only 01
Uniform	9,080	5,381	2,807	1,365
AKS	10,259	5,602	1,628	1,144
MDP^3	9,884	5,254	2,003	1,492
WFS-SB	10,251	5,377	1,636	1,369
DIG	9,719	5,262	2,168	1,484

Table 9: BLIP2-ITM frontend scaling. Times are in seconds per question. The shaded $N = 512$ row is the reference for reduction; gray bars encode complete frontend time on a common zero-based scale. The lowest time and largest reduction are bold.

Workload		Frontend time (s/question)			
Candidates N	Scored pairs	Decoding	Scoring	Total frontend	Reduction (%)
512 (ref.)	1,380,447	1.435	1.200	2.639	–
256	691,131	0.869	0.568	1.449	45.1
128	345,600	0.546	0.257	0.813	69.2
64	172,800	0.326	0.122	**0.459**	**82.6**

C.5 End-to-End Pipeline Cost

All timing results use the full Video-MME split. The scorer frontend and Qwen downstream are measured on NVIDIA RTX 4090 GPUs. The scorer uses three workers per GPU and a per-worker batch size of 128. Q-GPS selection is measured with precomputed relevance scores and 16 CPU workers. Frontend time covers candidate decoding through score serialization.

Tab. 9 first isolates how candidate-pool size affects the BLIP2-ITM frontend. Reducing N from 512 to 64 lowers frontend time from 2.639 to 0.459 seconds per question, a reduction of 82.6%. Although the number of scored pairs grows nearly linearly with N, throughput rises from 142.94 to 193.98 image–text pairs per second as larger pools improve batch utilization. Frontend latency therefore grows sublinearly with N.

Tab. 10 next decomposes the pipeline at $N = 512$ and $B = 32$ and compares three frozen scorers. With BLIP2-ITM, Q-GPS selection takes 0.005 seconds out of the 4.466-second total, or approximately 0.1% of the pipeline time. SigLIP2 reduces scoring time by 70.8% relative to BLIP2-ITM, but total latency falls by only 14.6% because decoding and Qwen inference account for most of the remaining time.

Fig. 10 combines these measured components to compare the end-to-end trade-off across N and B. End-to-end time sums the BLIP2-ITM frontend in Tab. 9, Q-GPS selection, and budget-specific Qwen inference. Each curve fixes B and varies $N = 64, 128, 256, 512$ from left to right, while the outlined points mark $(N, B) = (64, 32)$ and $(512, 8)$.

Table 10: Pipeline wall time per Video-MME question (seconds). Shading marks the default scorer; the Q-GPS stage is highlighted separately. Gray bars encode total time on a common zero-based scale, and the lowest total is bold.

	Frontend		Selection	Inference	End-to-end
Frozen scorer	**Decoding**	**Scoring**	**Q-GPS**	**Qwen**	**Total (s)**
BLIP2-ITM ViT-G	1.435	1.200	0.005	1.826	4.466
SigLIP2 Base/16@224	1.635	0.350	0.005	1.826	**3.816**
CLIP-L/14@336	1.495	0.985	0.005	1.826	4.311

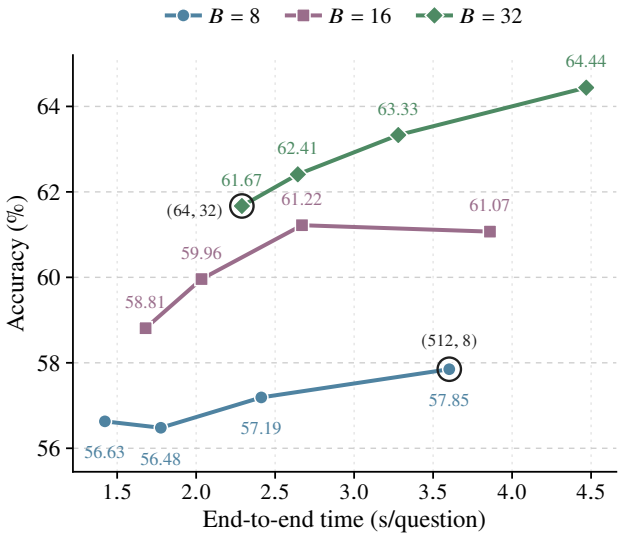


Figure 10: End-to-end accuracy–latency trade-off across candidate-pool sizes N and output budgets B on Video-MME. Point labels report accuracy.

Accuracy increases with B at every candidate-pool size, whereas increasing N at fixed B is not consistently beneficial. The configuration $(N, B) = (64, 32)$ is both faster and more accurate than $(512, 8)$, reaching 61.67% in 2.289 seconds per question instead of 57.85% in 3.602 seconds. The highest accuracy, 64.44%, is obtained at $(N, B) = (512, 32)$, showing that a larger candidate pool is most useful when paired with a sufficient output budget. Thus, N and B should be chosen together: N controls candidate density and frontend cost, while B controls the visual evidence passed to Qwen and its downstream cost.

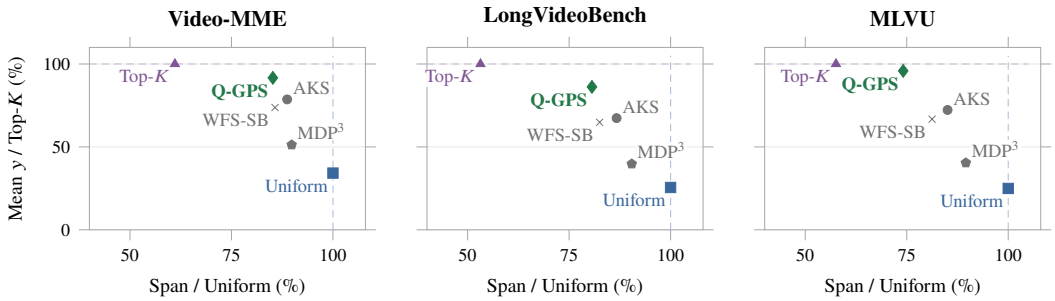
D Quantitative Analysis and Qualitative Visualization

D.1 Quantitative Analysis for Frame Selection

We provide the full statistics underlying the selection-behavior comparison in Fig. 5 of the main paper. We analyze the selected sets produced by Uniform, AKS, MDP^3, WFS-SB, and Q-GPS under the controlled protocol of Tab. 1 in the main paper, and include frozen BLIP-2-ITM Top-K as a relevance-only reference. DIG is excluded because its official protocol uses a different 2-fps candidate pool and LLM-based frame scoring, so its selections do not use the same candidate sequence. Every method uses the same candidate construction with a cap of 512 frames. To compare all selections on the same temporal scale, each selected frame is mapped to its nearest location in the corresponding candidate pool. For a selected set $\mathcal{I}_B$, let $i_{(1)} < \cdots < i_{(B)}$ denote its candidate locations in temporal order. We define its normalized temporal span as

$$\mathrm{Span}(\mathcal{I}_B) = 100\frac{i_{(B)} - i_{(1)}}{N - 1}. \tag{36}$$

Table 11: Temporal coverage and query alignment under the protocol of Tab. 1 in the main paper. In the plots, span is relative to Uniform and mean relevance is relative to Top-K; dashed lines mark the respective 100% references, and gray points denote AKS, MDP^3, and WFS-SB. Q-GPS retains 74.1–85.2% of Uniform's span and 86.2–95.8% of Top-K's mean relevance. The table reports original values; bold query values are best among non-Top-K methods.



Dataset	Method	Temporal distribution (Uniform reference) Span (%)	Deciles	Close (%)	Query alignment (Top-K reference) Mean y	Top-B (%)
Video-MME	Uniform	100.00	9.33	0.00	0.237	4.03
	Top-K	61.06	4.69	61.52	0.693	100.00
	AKS	88.72	8.59	10.17	0.545	40.82
	MDP^3	89.83	8.71	8.68	0.355	11.92
	WFS-SB	85.73	6.96	25.47	0.511	28.82
	Q-GPS	85.20	7.50	15.80	**0.635**	**50.14**
LongVideoBench	Uniform	100.00	9.33	0.00	0.122	5.02
	Top-K	53.17	4.12	68.35	0.478	100.00
	AKS	86.70	8.39	12.89	0.322	38.82
	MDP^3	90.43	8.66	9.73	0.190	12.43
	WFS-SB	82.52	6.46	33.74	0.310	28.71
	Q-GPS	80.63	6.71	24.57	**0.412**	**53.27**
MLVU	Uniform	100.00	9.33	0.00	0.126	3.70
	Top-K	57.57	4.47	64.93	0.505	100.00
	AKS	85.05	8.26	14.83	0.365	43.24
	MDP^3	89.56	8.68	9.40	0.204	13.04
	WFS-SB	81.25	6.31	38.37	0.337	32.19
	Q-GPS	74.13	6.10	38.82	**0.484**	**70.15**

For Deciles, we map each selected location to the normalized coordinate $u_i = (i-1)/(N-1)$, partition $[0,1]$ into ten equal temporal bins, and count the number of occupied bins. Close measures the fraction of consecutive selected locations whose normalized separation is below 1%:

$$\mathrm{Close}(\mathcal{I}_B) = \frac{100}{B-1} \sum_{k=1}^{B-1} \mathbb{I}\left[\frac{i_{(k+1)} - i_{(k)}}{N-1} < 0.01\right]. \tag{37}$$

Mean y is the average frozen BLIP-2-ITM relevance of the selected frames,

$$\mathrm{MeanY}(\mathcal{I}_B) = \frac{1}{B} \sum_{i \in \mathcal{I}_B} y_i. \tag{38}$$

Let $\mathcal{T}_B$ denote the B candidates with the highest frozen relevance for the same query. Their overlap with the selected set is

$$\mathrm{TopB}(\mathcal{I}_B) = 100 \frac{|\mathcal{I}_B \cap \mathcal{T}_B|}{B}. \tag{39}$$

Span, Close, and Top-B are reported in %. Each entry in Tab. 11 averages per-question measurements over $B \in \{8, 16, 32\}$.

The comparison shows two extremes. Uniform spans the complete candidate order by construction, occupies 9.33 deciles, and has no adjacent selections within the 1% threshold. Its mean relevance

ranges from 0.122 to 0.237, and its Top-B overlap ranges from 3.70% to 5.02%. Top-K shows the opposite extreme. It has the highest direct relevance and 100% Top-B overlap by construction, but covers only 4.12 to 4.69 deciles and places 61.52% to 68.35% of adjacent selections within 1% of the candidate pool. MDP^3, AKS, and WFS-SB make different trade-offs between temporal coverage and query relevance.

Q-GPS combines broad temporal coverage with high query relevance. It retains 74.1%–85.2% of Uniform's temporal span and 86.2%–95.8% of Top-K's mean relevance. It retains 50.14% to 70.15% of the exact Top-B set and has the highest mean y and Top-B overlap among the non-Top-K methods on every dataset. The path covariance therefore does not force the selected frames to spread uniformly over time. Compared with Top-K, its selected sets span a wider temporal range while retaining high relevance.

D.2 Qualitative Visualization for Frame Selection

We next visualize how the selection behavior summarized in Tab. 11 translates into evidence available to the downstream model. All six cases use an $N = 64$ uniformly sampled candidate pool, a frame budget of $B = 8$, and Qwen2.5-VL-7B-Instruct as the answer model. Each reported answer is generated from the frames selected by the corresponding method. In panels showing the progressive Q-GPS updates, the residual heat strip encodes $w_j \Sigma_{jj}^{(b)}$ at each unselected candidate j, while the suppression strip shows the reduction induced by the next selection. Each color scale is shared across steps within a case. The displayed $R(b)$ is the sum of the residual-strip values after b selections divided by their initial sum. Thus, $R(0) = 1$, and lower values indicate less remaining query-weighted residual.

D.2.1 Multi-Event Questions on *Mouse in Manhattan*

We consider two complementary queries over the first eight minutes of *Mouse in Manhattan*. One is a multi-event multiple-choice question, and the other is an open-ended generation question. The compact candidate pool allows every candidate and all eight selection steps to be displayed at readable resolution. Frames selected by Uniform, Top-K, and Q-GPS are supplied to the model in chronological order.

The next two pages show the complete cases. Each question appears in the shaded header. Panel A shows all 64 candidates in an 8×8 grid. Green S-badges indicate Q-GPS selection order, blue U-badges mark Uniform, and red T-badges mark Top-K. In the multiple-choice case, Panel B shows all eight progressive updates, including the residual state, suppression strip, next selection, and selected frame. In the open-ended case, Panel B instead shows the raw Uniform, Top-K, and Q-GPS answers, while the shaded header gives the reference answer. Green text marks reference-supported facts, red text marks incorrect or conflicting claims, and the remaining narrative stays black.

Fig. 11 asks for the sequence linking Jerry's shoe-stand mishap, hotel hiding place, and later jewelry. The decisive visual evidence appears at 106 seconds (black shoe polish), 228 seconds (a black top hat), and 412 seconds (a white pearl necklace). Uniform retains the necklace but misses the first two facts. Top-K allocates most of its budget to the street interval from 138 to 176 seconds. It retains the hotel hat but misses the shoe-polish and jewelry evidence. Across successive selections, residual conditioning suppresses already covered neighborhoods, and the Q-GPS set reaches all three decisive intervals.

Fig. 12 tests whether an open-ended response can combine evidence about a street obstacle, recovery from a rooftop fall, and a later party. The relevant frames show giant green high heels at 152 seconds, Jerry catching and climbing the red candle at 282 seconds, and the party activity at 320–358 seconds. Uniform covers the street and party intervals but omits the candle recovery. Top-K covers the candle and party intervals but omits the street obstacle. Q-GPS covers all three intervals. Correspondingly, the model substitutes a cat and omits the recovery action from the Uniform context, while its Top-K response describes the candle recovery but does not identify the high heel or dancing. The Q-GPS response mentions elements associated with all three parts of the query: the high heel, candle climbing, and dancing.

D.2.2 BENCHMARK CASE STUDIES

We further examine four benchmark cases with official answer labels from Video-MME and LongVideoBench. Each figure shows the complete selection process. Panel A marks the three selected sets in the candidate pool, Panel B shows the eight Q-GPS residual updates, and the caption reports the downstream answers.

Video-MME action recognition. Fig. 13 asks which acrobatic pose the pair performs. Uniform and Top-K retain nearby but different poses, causing the downstream model to choose the lifting and shoulder-balancing alternatives. Q-GPS first anchors the relevant acrobatics segment. Conditioning then lowers the coverage gain of nearby redundant candidates, and a later selection captures the distinct handstand pose. The resulting context provides the direct evidence needed to distinguish the correct answer from the visually similar alternatives.

Video-MME counting. Fig. 14 requires following a collision through its outcome rather than counting an intermediate state. The downstream model answers "two" from both the Uniform and Top-K contexts. The Q-GPS context spans the collision and its outcome, including a later frame showing the three resulting bodies. This evidence supports the correct post-collision count of three.

Video-MME temporal perception. Fig. 15 asks which planet is visited for the longest time, so the answer requires comparing evidence across several temporally separated segments. The Top-K context concentrates on the Mars–Jupiter portion of the video, while the Uniform context does not retain enough observations to determine which visit lasts the longest. The Q-GPS context contains observations from the Venus visit together with the Mars and Jupiter segments, enabling the downstream model to compare the visits and identify Venus as the longest.

LongVideoBench event relation.
Fig. 16 asks what appears after a "How to Draw Lewis Structures" slide. The question requires both the referenced slide and the state that follows it. Q-GPS retains the subsequent frame at 1041 seconds, which directly shows a light-green box labeled CH_3OH. From the Uniform and Top-K contexts, the downstream model instead chooses the yellow and pink alternatives, respectively. This evidence resolves both the temporal relation and the box color.

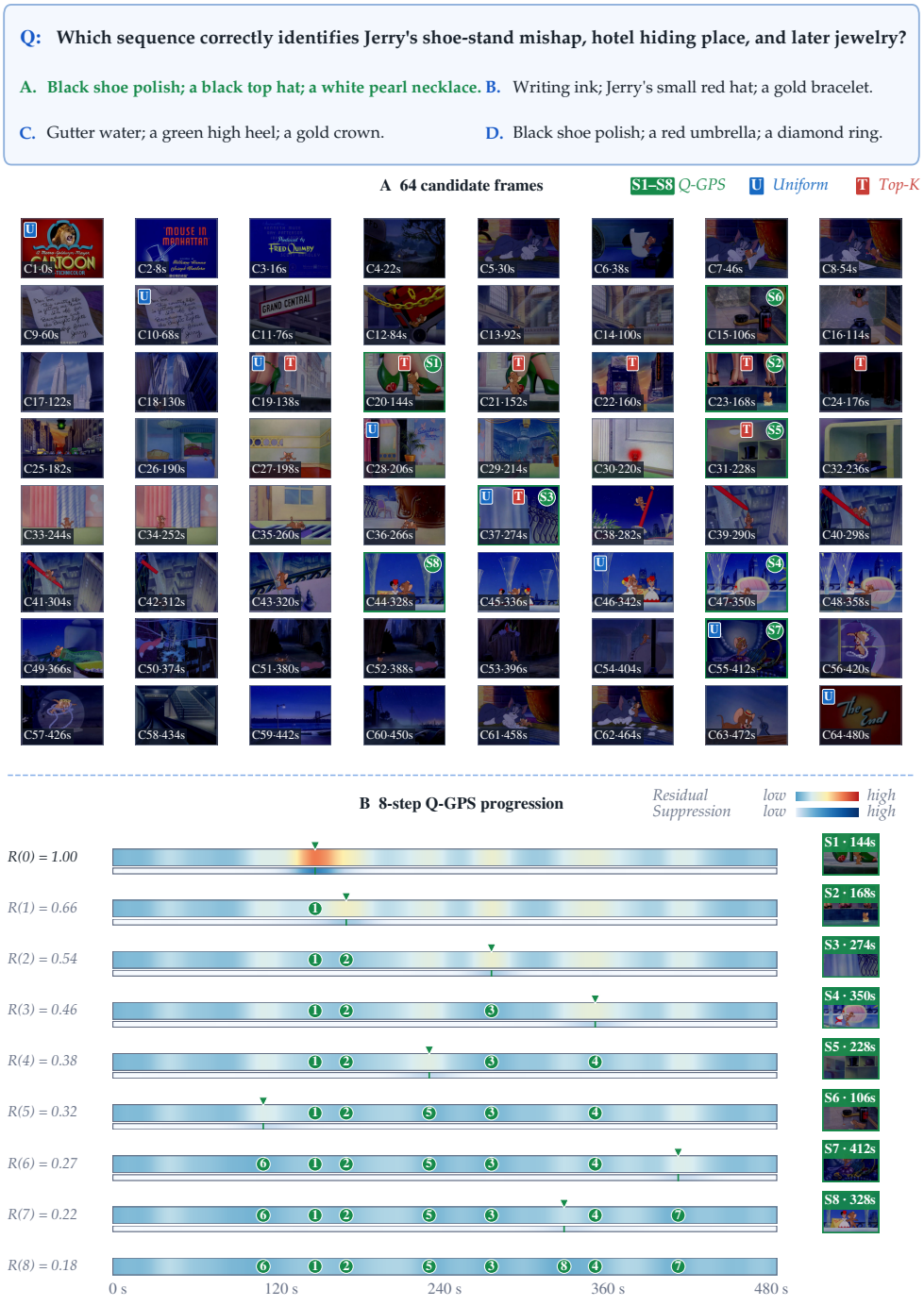


Figure 11: The model answers B ("writing ink; Jerry's small red hat; a gold bracelet") from the Uniform context, C ("gutter water; a green high heel; a gold crown") from the Top-K context, and A ("black shoe polish; a black top hat; a white pearl necklace") from the Q-GPS context. Only Q-GPS retains the shoe-polish, hotel-hat, and pearl-necklace evidence needed to verify the complete sequence.

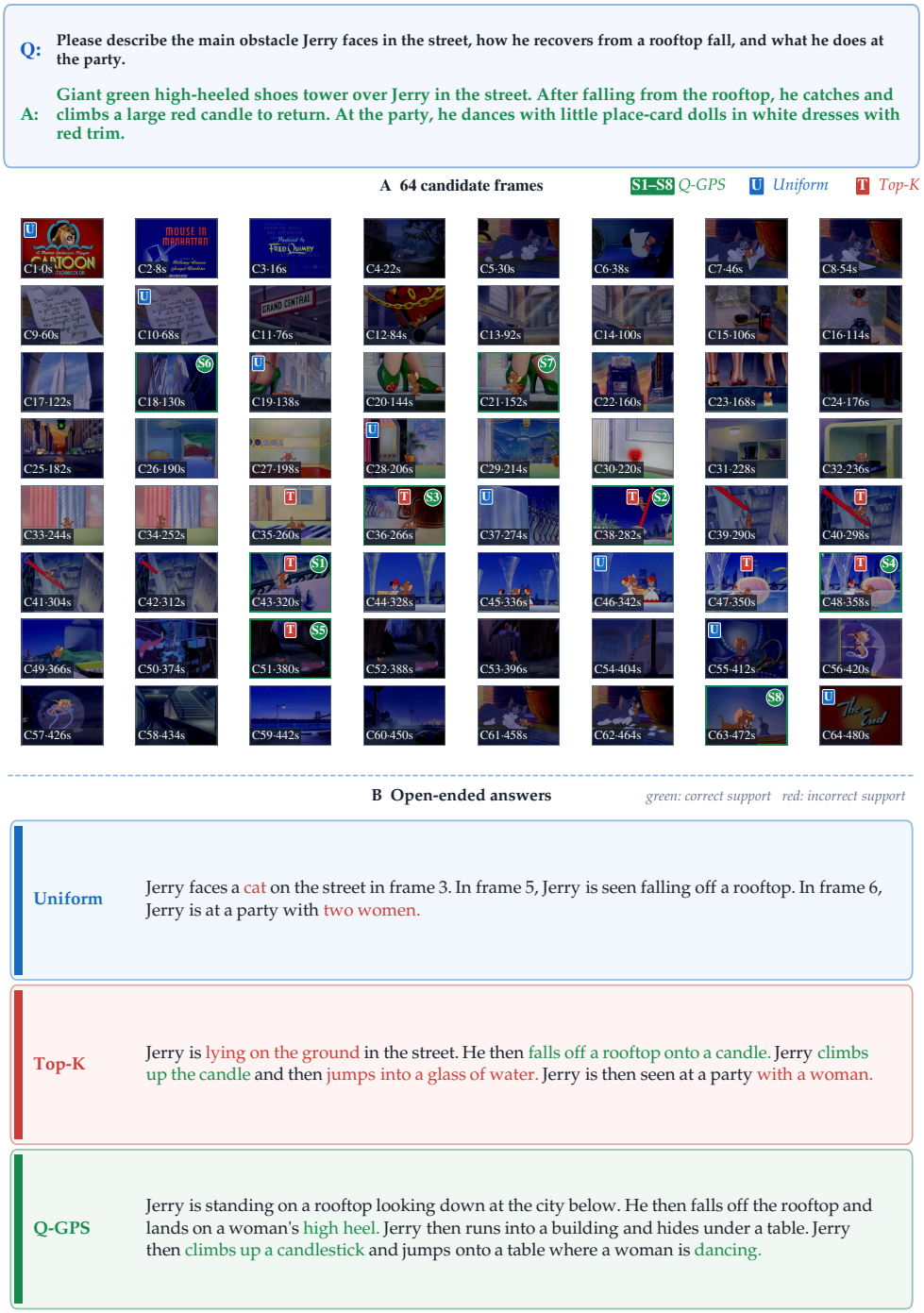


Figure 12: Open-ended *Mouse in Manhattan* case. The query concerns a street obstacle, recovery from a rooftop fall, and a later party. Uniform omits the candle-recovery interval, while Top-K omits the street-obstacle interval. Q-GPS retains frames from all three relevant intervals, and its response mentions the high heel, candle climbing, and dancing.



Figure 13: The model answers D ("the man lifts the woman with one hand") from the Uniform context, A ("the woman balances on the man's shoulders") from the Top-K context, and C ("the man handstands on the woman's stomach") from the Q-GPS context. The baseline contexts contain other acrobatic poses, while Q-GPS retains the queried pose itself.

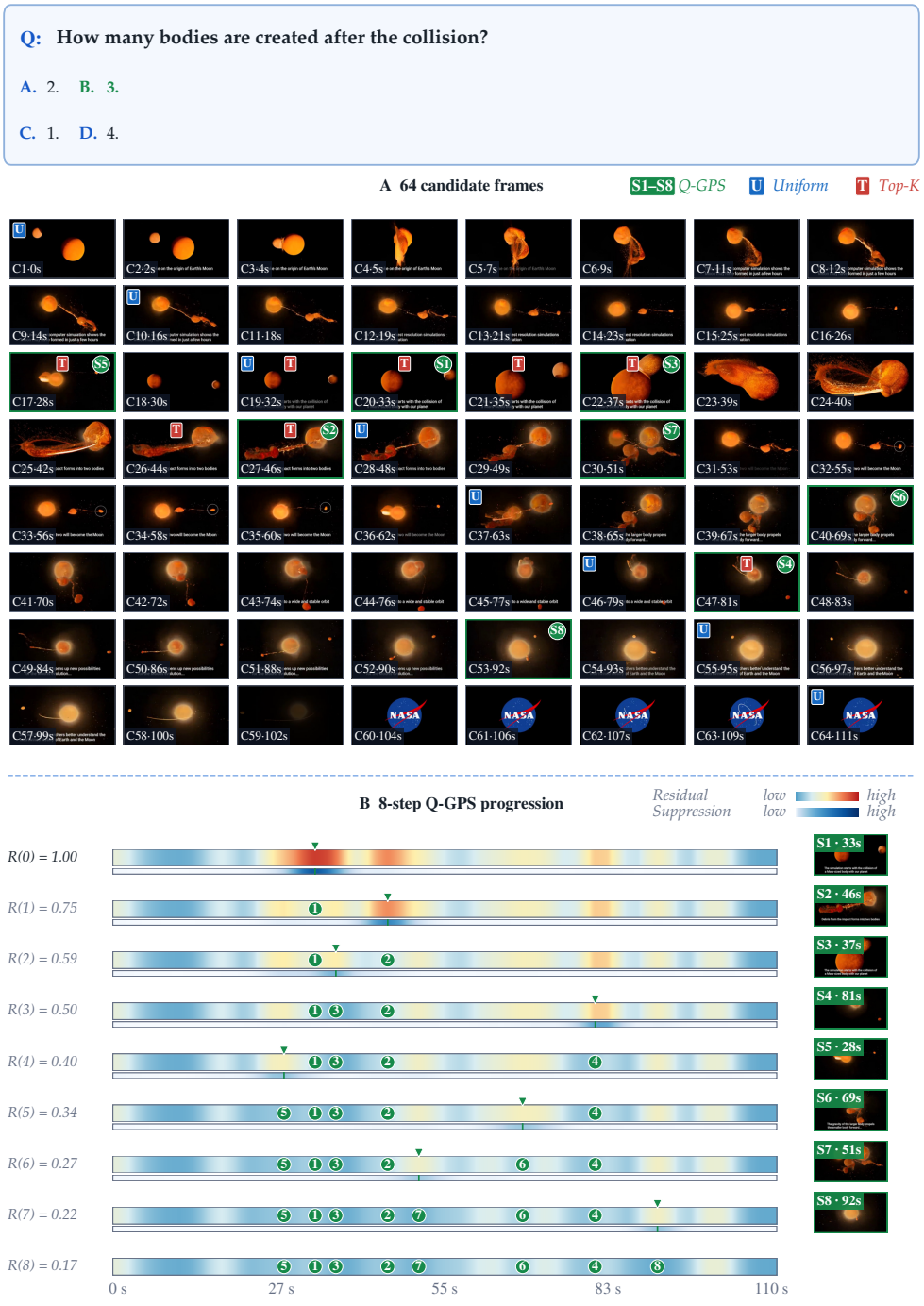


Figure 14: The model answers A (“2”) from both the Uniform and Top-K contexts and B (“3”) from the Q-GPS context. The baseline contexts lead to an undercount, while Q-GPS spans the collision and the resulting bodies needed for the correct post-collision count.



Figure 15: The model answers A ("Jupiter") from both the Uniform and Top-K contexts and C ("Venus") from the Q-GPS context. Top-K concentrates on the Mars–Jupiter portion of the video, whereas Q-GPS retains temporally separated observations across the Venus visit and the competing planetary segments needed to identify the longest visit.

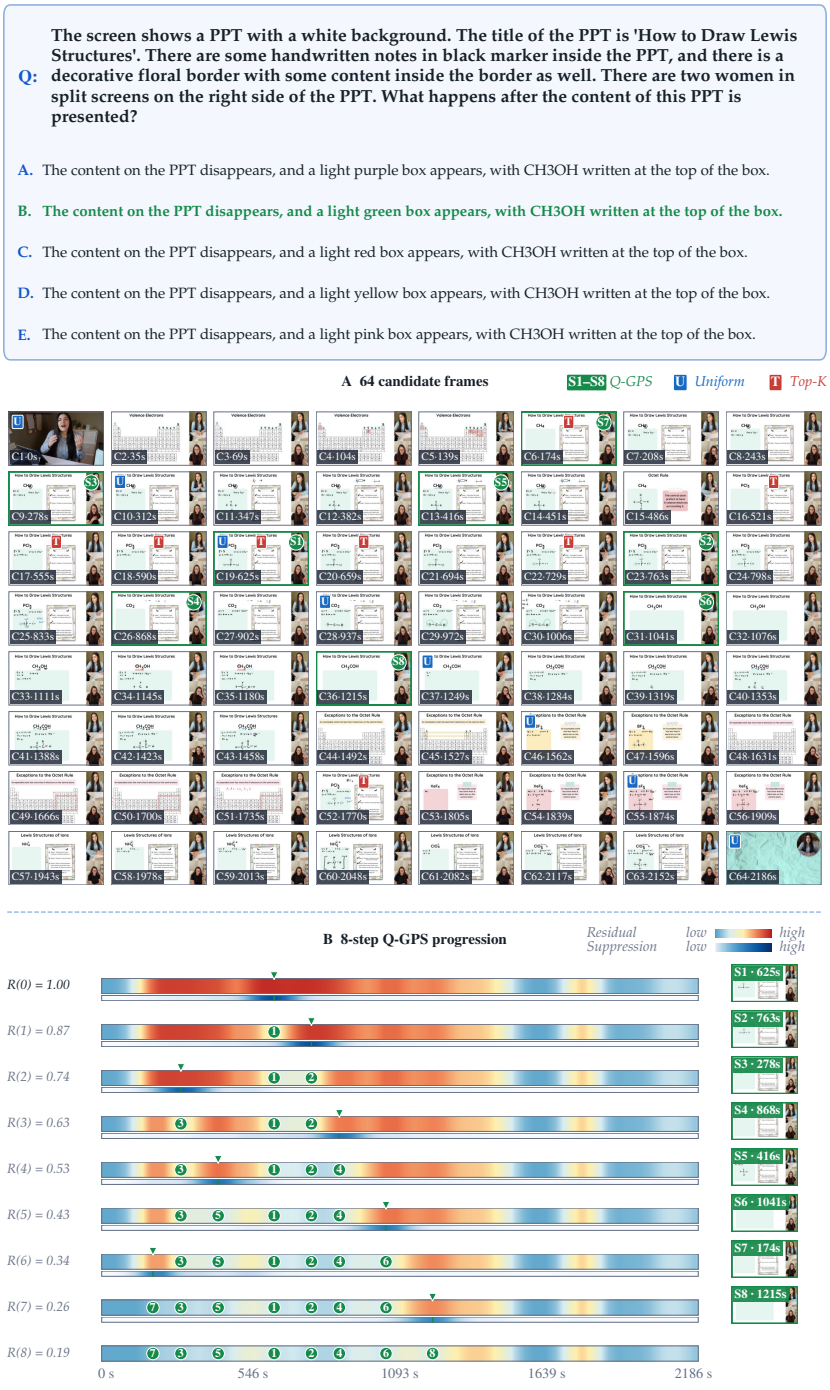


Figure 16: LongVideoBench event-relation case. The model answers D ("light yellow") from the Uniform context, E ("light pink") from the Top-K context, and B ("light green") from the Q-GPS context. Q-GPS retains the subsequent frame at 1041 seconds that directly shows the light-green box labeled CH_3OH, while both baseline contexts omit this state.

E INFERENCE PROMPT TEMPLATES

The following benchmark-specific user prompts are used by the Qwen2.5-VL evaluation adapter in the released repository, with the selected frames supplied as visual inputs. The adapter uses the system prompt `You are a helpful assistant.` Within each benchmark, the user prompt remains fixed across frame-selection methods and output budgets. Placeholders enclosed in braces are replaced with the content of each evaluation sample, and the fifth LongVideoBench option is included only when available. The original Video-MME prompt mentions subtitles, although no subtitles are provided in our experiments. For MLVU, {question} contains the question and its four labeled answer choices.

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}
A. {option_1}
B. {option_2}
C. {option_3}
D. {option_4}
Answer with the option's letter from the given choices directly.
```

LongVideoBench

```
{question}
A. {option_1}
B. {option_2}
C. {option_3}
D. {option_4}
E. {option_5_if_available}
Answer with the option's letter from the given choices directly.
```

MLVU

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