COMPASS: STRUCTURED EVIDENCE SELECTION FOR LONG-VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video understanding with Multimodal Large Language Models (MLLMs) remains challenging because raw video sequences contain far more visual information than can be accommodated within limited frame or token budgets. Existing question-driven frame selection methods typically rank frames independently based on query-frame similarity or temporal statistics, often selecting redundant frames, overlooking critical transitions, and failing to allocate visual evidence effectively across temporally complex videos. We propose **COMPASS** (Content-aware Multi-granularity Perception and Adaptive Semantic Selection), a deterministic and training-free framework for question-driven long-video understanding. COMPASS treats keyframe selection as a structured evidence allocation problem by jointly modeling query intent, video content structure, and temporal granularity. It derives the temporal scope and information requirements of a query, identifies coherent and informative regions, and performs set-level selection to balance query relevance, semantic coverage, and temporal continuity. A hierarchical allocation strategy further concentrates frame budgets on salient temporal windows while preserving global context. We evaluate COMPASS on LongVideoBench, Video-MME, LVBench, and MLVU under matched frame budgets. Across diverse visual encoders and downstream MLLMs, COMPASS consistently improves reasoning accuracy over existing selection strategies. On LongVideoBench with Qwen3-VL-8B, it achieves a 10.4% relative improvement over uniform sampling under identical frame budgets, with only 8.6 ms of CPU selection latency. These results demonstrate the effectiveness of structured, query-aware evidence allocation for efficient long-video understanding. Our code is available at anonymous.4open.science/r/compass-anonymous-source-C4CD.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have made substantial progress in video comprehension and reasoning (Maaz et al., 2024; Tang et al., 2025; Bai et al., 2025; Li et al., 2025; Team, 2025). However, processing long-form videos directly remains computationally prohibitive due to quadratic self-attention complexity and finite context windows (Tang et al., 2025; Sun et al., 2025; Song et al., 2024; Shu et al., 2025; Wu et al.; Wang et al., 2025; Nguyen et al., 2024). Question-driven keyframe selection has therefore emerged as an effective frontend approach to condense long videos into a compact frame budget (e.g., 32 or 64 frames) prior to multimodal inference (Liang et al., 2024; Hu et al., 2025; Liu et al., 2025; Zhang et al., 2025c).

Conventional query-aware selectors evaluate frame-query pairs independently using vision-language backbones such as CLIP (Radford et al., 2021) or BLIP (Li et al., 2022), followed by top-k filtering (Tang et al., 2025; Liu et al., 2025; Huang et al., 2025; Chen et al., 2025) or 1D signal modeling (Zhang et al., 2026; Xu et al., 2026). As illustrated in Figure 1(a), existing paradigms exhibit three primary limitations. First, scalar similarity statistics conflate question scope with background visual variance, causing global queries to collapse into isolated temporal clusters. Second, pointwise scoring ignores temporal feature geometry, yielding redundant frames in static scenes while missing critical physical transitions. Finally, single-granularity selection across raw sequences lacks coarse-to-fine adaptability, diluting frame density over transient yet informative events.

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

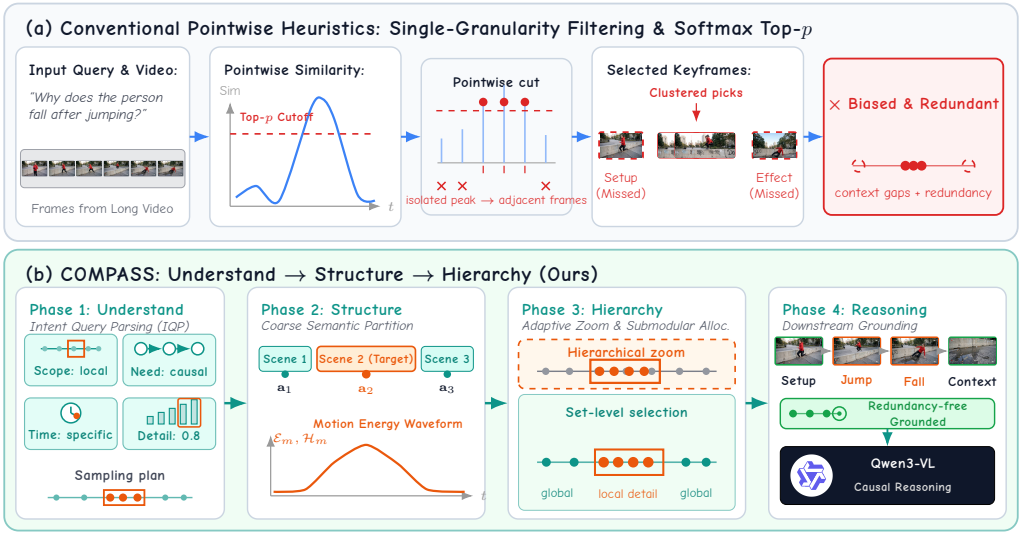

Figure 1: **Conceptual comparison of keyframe selection methods.** (a) Conventional pointwise filtering and 1D signal heuristics rank frames independently, leading to temporal collapse and redundant clusters. (b) COMPASS translates queries into structured plans, models video topology with a Content Structure Graph, and performs multi-scale submodular optimization.

To address these limitations, we introduce COMPASS (Content-aware Multi-granularity Perception & Adaptive Semantic Selection), a deterministic, training-free framework that aligns query intent with video topology (Figure 1(b)). COMPASS coordinates three complementary components. Intent-guided Query Planning (IQP) first parses natural language questions into explicit execution parameters $\mathcal{P} = (\alpha_{\text{scope}}, \mathcal{C}_{\text{need}}, \tau_{\text{time}}, \delta_{\text{detail}})$. Hierarchical Zoom (HZ) then identifies query-relevant temporal windows with content-adaptive budgeting based on information entropy and motion dynamics. Finally, Content Structure Graph Submodular Selection (CSG-SS) models the visual manifold, formulating frame selection as a monotone submodular facility-location problem with a theoretical $(1 - 1/e)$ guarantee for each fixed-window surrogate optimization.

Evaluations across four demanding benchmarks (LongVideoBench (Wu et al.), Video-MME (Fu et al., 2025), LVBench (Wang et al., 2025), and MLVU (Zhou et al., 2025)) demonstrate that COMPASS consistently outperforms prior heuristic, stochastic, and signal-based selectors. In summary, our contributions are threefold. First, we propose COMPASS, a training-free framework that replaces pointwise heuristics with structured query planning and topology-aware submodular optimization with a $(1 - 1/e)$ bound for fixed-window greedy selection. Second, we introduce a coarse-to-fine Hierarchical Zoom mechanism with content-aware budgeting to balance macro-timeline coverage with dense sampling on critical event intervals. Finally, comprehensive experiments across four benchmarks and four MLLM families show that COMPASS delivers up to a $4\times$ visual token reduction with only 8.6 ms of CPU selection latency.

2 Related Work

2.1 Long-Form Video Understanding with MLLMs

Recent MLLMs advance video understanding via dynamic resolution patching and joint spatial-temporal pretraining (Maaz et al., 2024; Tang et al., 2025; Li et al., 2025; Zhang et al., 2025d; Bai et al., 2025; Team, 2025; Zhu et al., 2025). However, processing hour-scale videos directly remains computationally prohibitive due to context limits and quadratic self-attention complexity (Tang et al., 2025; Song et al., 2024). While internal token compression and memory caches (Shen et al., 2025; Song et al., 2024; Shu et al., 2025) mitigate in-network activation costs, models remain bounded by token limits. Question-driven keyframe selection (Liang et al., 2024; Tang et al., 2025) provides

an orthogonal, model-agnostic frontend solution, condensing raw sequences into informative visual evidence before multimodal inference.

2.2 Question-Driven Keyframe Selection

Pointwise and diversity heuristics. Early selectors score frame-query pairs independently using vision-language models such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) via top-k ranking (Radford et al., 2021) or clip retrieval (Huang et al., 2025; Chen et al., 2025). Because pointwise scoring neglects temporal continuity, subsequent works enforce diversity through recursive binning (Tang et al., 2025), determinantal point processes (Sun et al., 2025), stochastic sampling (Liu et al., 2025; Zhang et al., 2025c), or active bandit search (Zhu et al., 2026; Zou et al., 2025). While mitigating redundancy, these heuristics rely on fixed temporal intervals or randomized perturbations, limiting their adaptability to complex causal questions.

Temporal signal and variance modeling. Recent approaches treat similarity trajectories as temporal signals. AdaQ (Zhang et al., 2026) dynamically scales the softmax temperature using score variance to generate adaptive quasi-Gaussian sampling intervals. RIDGE (Xu et al., 2026) identifies event boundaries and peaks through numerical derivatives and curvature filtering. However, both methods operate primarily on 1D scalar projections. Because scalar statistics capture distribution variance rather than feature manifold geometry, these approaches struggle to track non-linear scene transitions and often over-allocate frame budgets to visually volatile backgrounds.

Structured set-level optimization. In contrast to pointwise ranking and 1D derivative filters, COMPASS unifies structured query intent with video manifold topology. By explicitly deriving temporal scope via Intent-guided Query Planning (IQP), structuring candidate frames on a Content Structure Graph (CSG), and optimizing set-level facility-location coverage with submodular guarantees, COMPASS achieves multi-granularity, redundancy-free evidence allocation in a single deterministic pass.

3 Methodology

We present COMPASS (Content-aware Multi-granularity Perception & Adaptive Semantic Selection), a deterministic, training-free framework for question-driven long-video keyframe selection. As illustrated in Figure 2, COMPASS coordinates query intent with video topology across three stages. First, Intent-guided Query Planning (IQP) translates the input question into structured sampling priors. Next, the Content Structure Graph (CSG) models the visual manifold to propose query-relevant temporal windows through Hierarchical Zoom (HZ). Finally, Submodular Selection (CSG-SS) optimizes intra-window frame subsets under a $(1 - 1/e)$ approximation guarantee before downstream reconciliation. The overall procedural flow of COMPASS is formally summarized in Algorithm 1 in the Appendix.

3.1 Preliminaries and Problem Formulation

Let a video sequence be $\mathcal{V} = \{f_t\}_{t=1}^{N}$. Using a frozen vision-language backbone such as CLIP, candidate frames and text query Q are projected onto the unit sphere $\mathbb{S}^{D-1}$ as ℓ_2-normalized embeddings $\mathbf{f}_t$ and $\mathbf{g}_Q$, respectively. Pointwise cosine similarities $s_t = \langle \mathbf{f}_t, \mathbf{g}_Q \rangle \in [-1, 1]$ are mapped to min-max normalized relevance weights $w_t = (s_t - \min_j s_j)/(\max_j s_j - \min_j s_j + \epsilon)$, where $\epsilon = 10^{-12}$.

Problem formulation. Given a frame budget constraint $K_{\min} \leq |S^*| \leq K \ll N$, our objective is to select a keyframe subset $S^* \subseteq \mathcal{U} \triangleq \{1, \ldots, N\}$ to maximize the downstream reasoning likelihood $\mathbb{P}(Y^* \mid \mathcal{M}(\mathcal{V}_{S^*}, Q))$. Because direct optimization through frozen MLLM decoders is non-differentiable and computationally intractable, COMPASS optimizes a set-level submodular surrogate $F(S)$ (Section 3.4) that provides theoretical approximation guarantees. The selected keyframe sequence $\mathcal{V}_{S^*}$ is provided chronologically to $\mathcal{M}$, while importance weights $\boldsymbol{\gamma} = \{\gamma_t\}_{t \in S^*}$ serve as frame-level attribution weights. Conventional heuristic selectors assume pointwise independence with $p_t \propto \exp(s_t/\tau)$, which conflates query scope with background visual variance, often over-selecting frames from static scenes while missing critical physical transitions. COMPASS addresses this limitation by steering set-level selection with explicit query intent.

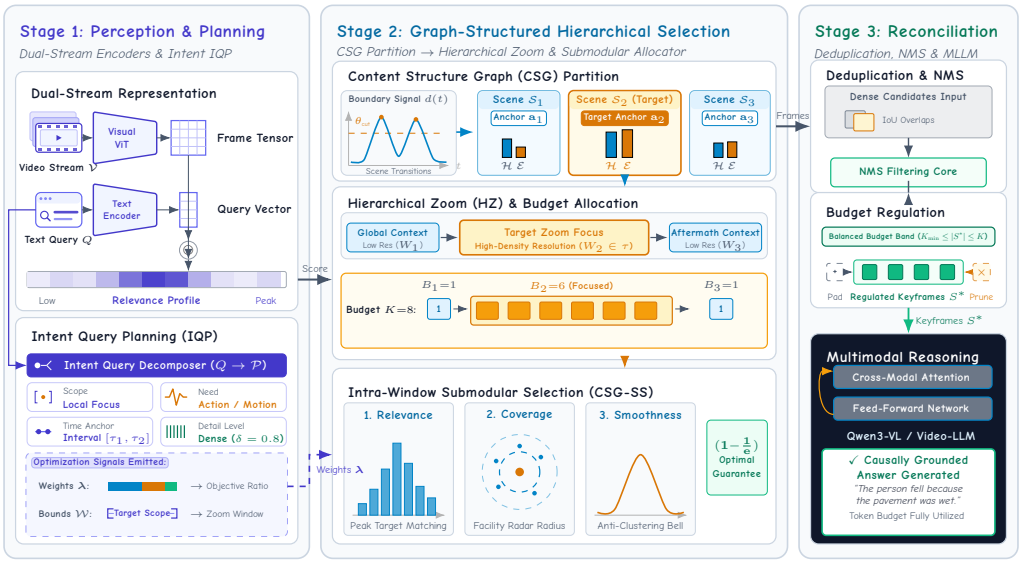

Figure 2: **Overview of the COMPASS framework.** Stage 1 translates input questions into structured execution plans via IQP. Stage 2 constructs a Content Structure Graph to guide coarse-to-fine Hierarchical Zoom (HZ) and content-aware budgeting. Stage 3 performs set-level submodular selection within candidate windows, followed by temporal NMS and budget regulation for downstream MLLM reasoning.

3.2 Intent-Guided Query Planning (IQP)

Rather than inferring temporal scope post-hoc from similarity distributions, IQP parses the query Q into an executable plan $\mathcal{P} = \Pi(Q) = (\alpha_{\text{scope}}, \mathcal{C}_{\text{need}}, \tau_{\text{time}}, \delta_{\text{detail}})$. Here, $\alpha_{\text{scope}} \in \{\text{local}, \text{global}, \text{hybrid}\}$ designates the temporal scope, $\mathcal{C}_{\text{need}} \in \{\text{motion/causal}, \text{static/entity}\}$ records the target visual attributes, τ_{time} specifies temporal boundaries, and $\delta_{\text{detail}} \in [0, 1]$ controls sampling granularity. The scope parameter α_{scope} directly sets the trade-off weights $\boldsymbol{\lambda} = (\lambda_{\text{rel}}, \lambda_{\text{cov}}, \lambda_{\text{smooth}})$ in the submodular objective:

$$\boldsymbol{\lambda}(\alpha_{\text{scope}}) = \begin{cases} (1.0,\ 0.2,\ 0.2), & \text{if } \alpha_{\text{scope}} = \text{local} \quad \text{(relevance-dominant)}, \\ (0.2,\ 1.0,\ 0.8), & \text{if } \alpha_{\text{scope}} = \text{global} \quad \text{(coverage-dominant)}, \\ (0.7,\ 0.6,\ 0.5), & \text{if } \alpha_{\text{scope}} = \text{hybrid} \quad \text{(balanced default)}. \end{cases} \tag{1}$$

Here, τ_{time} clips candidate window boundaries, while δ_{detail} scales partition resolution. The mapping $\Pi(Q)$ is implemented via deterministic keyword and regex rules, adding negligible computational overhead.

3.3 Content Structure Graph (CSG) Formulation

To model video topology without parameter training, COMPASS constructs a Content Structure Graph over the visual embeddings $\{\mathbf{f}_t\}_{t=1}^{N}$.

Adaptive scene partitioning. Adjacent frame cosine distances $d_t = 1 - \mathbf{f}_t^\top \mathbf{f}_{t+1}$ yield empirical mean μ_d and standard deviation σ_d. Scene boundaries are established where $d_t > \mu_d + k_{\text{scene}} \cdot \sigma_d$ ($k_{\text{scene}} = 2.0$). Segments shorter than $L_{\min} = 2$ frames are merged with adjacent intervals, producing M disjoint scenes $\mathcal{S} = \{\mathcal{S}_m\}_{m=1}^{M}$.

Scene descriptors. For each scene $\mathcal{S}_m = [t_{\text{start}}^{(m)}, t_{\text{end}}^{(m)})$, we compute four descriptive metrics. First, the normalized centroid $\mathbf{c}_m = \bar{\mathbf{f}}_m / \|\bar{\mathbf{f}}_m\|_2$ if $\|\bar{\mathbf{f}}_m\|_2 > 10^{-12}$, and $\mathbf{c}_m = \bar{\mathbf{f}}_m$ otherwise, where $\bar{\mathbf{f}}_m = |\mathcal{S}_m|^{-1} \sum_{t \in \mathcal{S}_m} \mathbf{f}_t$. This descriptor characterizes average visual content; the medoid anchor is $r_m = \arg\max_{t \in \mathcal{S}_m} \langle \mathbf{f}_t, \mathbf{c}_m \rangle$. Scene ranking uses the cumulative raw similarity $\mathcal{M}_m = \sum_{t \in \mathcal{S}_m} s_t$, while the set objective uses normalized relevance w_t. In addition, intra-scene visual diversity is defined as

$\mathcal{H}_m = \mathrm{Var}_{t\in\mathcal{S}_m}(1-\mathbf{f}_t^\top \mathbf{c}_m)+\epsilon$, and kinematic motion energy is given by $\mathcal{E}_m = \frac{1}{|\mathcal{S}_m|-1}\sum_{t=t_{\text{start}}^{(m)}}^{t_{\text{end}}^{(m)}-2}(1-\mathbf{f}_t^\top \mathbf{f}_{t+1})$ for $|\mathcal{S}_m| \geq 2$, and zero otherwise.

3.4 Set-Level Submodular Keyframe Optimization

Given a fixed candidate pool $\mathcal{U}$ and frame budget $1 \leq B \leq K$, frame selection maximizes a monotone submodular set function $F : 2^{\mathcal{U}} \to \mathbb{R}_{\geq 0}$:

$$\max_{S\subseteq\mathcal{U},\,|S|\leq B} F(S), \qquad F(S) \triangleq \lambda_{\text{rel}}\widetilde{\text{Rel}}(S) + \lambda_{\text{cov}}\widetilde{\text{Cov}}(S) + \lambda_{\text{smooth}}\widetilde{\text{Smooth}}(S), \tag{2}$$

where $\widetilde{\text{Rel}}(S) = \sum_{i\in S} w_i$. Let $\mathcal{A} = \{\mathbf{a}_j\}$ contain the scene representative embeddings in the window. The temporal reference set $\mathcal{T} = \{\tau_k\}$ contains up to 256 uniformly subsampled candidate timestamps (all candidate timestamps when their number is at most 256), and $\sigma_t = \max\{1, (\max\mathcal{T} - \min\mathcal{T})/B\}$. The coverage and temporal terms are:

$$\widetilde{\text{Cov}}(S) = \sum_{\mathbf{a}_j\in\mathcal{A}} \max_{i\in S} \max\left(0,\, \langle \mathbf{f}_i, \mathbf{a}_j\rangle\right), \quad \widetilde{\text{Smooth}}(S) = \sum_{\tau_k\in\mathcal{T}} \max_{i\in S} \exp\left(-\frac{|t_i - \tau_k|}{\sigma_t}\right). \tag{3}$$

We define each maximum over an empty selected set to be zero. The inner max operators induce diminishing marginal returns over already covered visual clusters and temporal regions, penalizing redundancy. These terms measure temporal coverage, rather than imposing a formal adjacency constraint.

Theoretical guarantees. The relevance term $\widetilde{\text{Rel}}(S)$ is modular, while $\widetilde{\text{Cov}}(S)$ and $\widetilde{\text{Smooth}}(S)$ are non-negative sums of facility-location functions of the form $g(S) = \sum_j \max_{i\in S}\phi_j(i)$. With $w_i \in [0,1]$, non-negative scope weights, unit-normalized feature vectors, fixed anchors and grid, and the empty-set convention above, $F(\emptyset) = 0$ and F is monotone submodular. For $|S| \leq B$, $F(S) \leq \lambda_{\text{rel}}B + \lambda_{\text{cov}}|\mathcal{A}| + \lambda_{\text{smooth}}|\mathcal{T}|$.

Theorem 1. *For a fixed window, candidate pool, anchors, temporal grid, and budget $1 \leq B \leq |\mathcal{U}|$, let S^* maximize Eq. equation 2 subject to $|S| \leq B$. Greedy forward selection for B steps, with any deterministic tie rule, obeys $F(S_{greedy}) \geq [1-(1-1/B)^B]F(S^*)$ and hence:*

$$F(S_{\text{greedy}}) \geq \left(1 - \frac{1}{e}\right) F(S^*) \approx 0.632 \cdot F(S^*). \tag{4}$$

This bound applies to each fixed-window optimization; it does not directly cover the final set after cross-window pooling, temporal NMS, or relevance backfilling. Appendix A gives the proof and its scope. The marginal gain $\hat{\gamma}_{f^*} \triangleq F(S\cup\{f^*\}) - F(S)$ is min-max normalized across selected frames to $\gamma_{f^*} \in [0.05, 1.0]$ to serve as downstream importance weights.

3.5 Hierarchical Zoom (HZ) and Reconciliation

Applying global submodular optimization directly across long uncurated videos dilutes frame density over dynamic action segments. Hierarchical Zoom addresses this issue through a coarse-to-fine schedule.

Semantic window proposals. Guided by planning tuple $\mathcal{P}$ and scene relevance values $\{\mathcal{M}_m\}$, candidate windows $\mathcal{W} = \{W_j\}_{j=1}^{|\mathcal{W}|}$ are constructed according to query scope. Under local scope, candidate windows are formed by expanding the top one or two peak scenes with a $0.10 \times N$ buffer. Under global scope, the video is partitioned into $J \in \{4, 6\}$ uniform segments based on δ_{detail}. Under hybrid scope, four uniform segments are combined with a focused window centered at $\arg\max_m \mathcal{M}_m$ with half-width $N(0.08 + 0.12\delta_{\text{detail}})$. All proposed windows are clipped to the temporal constraint τ_{time}.

Content-aware budget allocation. For each window W_j, visual richness $\mathcal{R}(W_j) = \sum_{m:\mathcal{S}_m\cap W_j\neq\emptyset} \frac{|\mathcal{S}_m\cap W_j|}{|\mathcal{S}_m|}(\beta_{\text{div}}\widetilde{\mathcal{H}}_m + \beta_{\text{mot}}\widetilde{\mathcal{E}}_m + \beta_{\text{dur}}\widetilde{\mathcal{D}}_m)$ aggregates normalized feature diversity $\widetilde{\mathcal{H}}_m$, kinematic motion energy $\widetilde{\mathcal{E}}_m$, and duration $\widetilde{\mathcal{D}}_m \in [0,1]$. The implementation uses

$(\beta_{\text{div}}, \beta_{\text{mot}}, \beta_{\text{dur}}) = (0.4, 0.3, 0.3)$ for all query types. To prevent frame starvation in less dynamic intervals, allocation probabilities are smoothed with a uniform prior:

$$p_j = 0.5q_j + \frac{0.5}{|\mathcal{W}|}, \qquad q_j = \begin{cases} \mathcal{R}(W_j)/\sum_k \mathcal{R}(W_k), & \sum_k \mathcal{R}(W_k) > 0, \\ 1/|\mathcal{W}|, & \text{otherwise.} \end{cases} \tag{5}$$

The total frame budget K is then allocated into discrete window quotas $\{B_j\}_{j=1}^{|\mathcal{W}|}$ satisfying $\sum_j B_j = K$.

Selection and reconciliation. Submodular optimization (Eq. equation 2) is executed independently within each window using the allocated budget B_j. When integer quotas satisfy $\sum_j B_j = K$, the pooled unique candidate set obeys $|S_{\text{raw}}| \le \sum_j |S_j| \le K$; overlapping windows and deduplication can only reduce its size. Frames selected across windows are pooled, sorted chronologically, and filtered by temporal NMS with threshold $\tau_{\text{nms}} = 0.5\,\text{s}$. Because candidate frames are sampled at 1 FPS (with inter-frame distances $\ge 1.0\,\text{s}$), $\tau_{\text{nms}} = 0.5\,\text{s}$ effectively functions as a deduplication step for overlapping window boundaries, while $\widetilde{\text{Smooth}}(S)$ encourages temporal coverage without guaranteeing continuity. Finally, the budget is regulated to satisfy capacity bounds: if $|S_{\text{nms}}| < K_{\min}$, the set is backfilled with the highest-relevance frames; if $|S_{\text{nms}}| > K$, frames with the lowest marginal gains $\hat{\gamma}_f$ are pruned.

4 Experiments

4.1 Experimental Settings

Benchmarks and metrics. We evaluate COMPASS on four standard long-video benchmarks: LongVideoBench (LVB) (Wu et al.), Video-MME (Fu et al., 2025) under the subtitle-free video-only setting, LVBench (Wang et al., 2025), and MLVU (Zhou et al., 2025). Following standard practice, we report accuracy (Acc, %) across all benchmarks.

Compared baselines. We compare COMPASS with 11 representative baseline methods covering four selection strategies. For query-agnostic selection, we include uniform sampling as a baseline. For pointwise retrieval, we evaluate Top-k selection (Radford et al., 2021), FRAG (Huang et al., 2025), and OneClip-RAG (Chen et al., 2025). For diversity-oriented approaches, we compare against BOLT (Liu et al., 2025), AKS (Tang et al., 2025), MDP^3 (Sun et al., 2025), and Q-Frame (Zhang et al., 2025c). For active and signal-based modeling, we evaluate FOCUS (Zhu et al., 2026), AdaQ (Zhang et al., 2026), and RIDGE (Xu et al., 2026). All baselines are evaluated using official codebases or published results under matched experimental settings.

Implementation details. We conduct evaluations on three MLLMs: LLaVA-OneVision-7B (K=32) (Li et al., 2025), LLaVA-Video-7B (K=64) (Zhang et al., 2025d), and Qwen3-VL-8B (K=64) (Team, 2025), with candidate frames sampled at 1 FPS. Frame feature matrices $\mathbf{F}$ and relevance scores $\mathbf{w}$ are extracted using CLIP (Radford et al., 2021) and LongCLIP (Zhang et al., 2025b), alongside cross-scorer validation on BLIP (Li et al., 2022), SigLIP (Zhai et al., 2023), and BLIP-ITM (Li et al., 2022). Default hyperparameters are set to $k_{\text{scene}} = 2.0$, $L_{\min} = 2$, $K_{\min} = 16$, and a temporal NMS window of $0.5\,\text{s}$. Submodular weights $\boldsymbol{\lambda}$ are determined by rule-based IQP via Eq. equation 1. COMPASS is deterministic and training-free, selecting keyframes in a single forward pass without iterative MLLM calls. Additional implementation details are provided in Appendix C.

4.2 Experimental Results

4.2.1 Main Results and Comparison with SOTA Methods

Table 1 compares COMPASS against existing keyframe selection methods across three MLLMs. Pointwise and retrieval-based baselines, such as Top-k and OneClip-RAG, perform well on localized retrieval tasks (e.g., LongVideoBench), but often struggle on multi-event benchmarks due to redundant frames and missing temporal context. Methods designed for temporal diversity, such as AKS, AdaQ, and RIDGE, yield more balanced performance; for example, RIDGE exhibits strong sensitivity on hour-scale videos in LVBench. COMPASS achieves the best overall results across most settings, yielding average relative improvements of 9.8%, 9.9%, and 10.4% on LLaVA-OneVision-7B, LLaVA-Video-7B, and Qwen3-VL-8B using LongCLIP. It consistently outperforms AdaQ and

Table 1: **Comprehensive performance comparison across four long-video benchmarks on three representative MLLMs.** Baselines are organized chronologically. *Frames* denotes the sampled visual frame budget fed to the downstream MLLM. *Avg.* denotes the mean relative improvement over the uniform baseline across the four disjoint overall benchmark benchmarks ($\frac{1}{4}\sum_D \frac{\mathrm{Acc}(D)-\mathrm{Acc}_{\mathrm{uni}}(D)}{\mathrm{Acc}_{\mathrm{uni}}(D)}$). The best overall results are in **bold** and the best competing baseline results are underlined.

Method	Embedding Model	Frames	LongVideoBench		Video-MME		LVBench	MLVU	Avg.
			Long	Overall	Long	Overall			
LLaVA-OneVision-7B	–	32	49.5	56.8	48.1	58.3	38.6	63.7	–
Top-k	CLIP	32	50.7 (+1.2)	58.3 (+1.5)	48.2 (+0.1)	58.9 (+0.6)	43.8 (+5.2)	66.9 (+3.2)	+5.5%
FRAG (arXiv)	MLLM (7B)	32	51.5 (+2.0)	57.6 (+0.8)	48.0 (-0.1)	58.6 (+0.3)	41.5 (+2.9)	65.3 (+1.6)	+1.5%
OneClip-RAG (arXiv)	CLIP	32	54.4 (+4.9)	58.1 (+1.3)	47.7 (-0.4)	57.9 (-0.4)	44.1 (+5.5)	67.6 (+3.9)	+5.5%
BOLT (CVPR 2025)	CLIP	32	53.5 (+4.0)	59.6 (+2.8)	49.6 (+1.5)	59.9 (+1.6)	38.2 (-0.4)	66.8 (+3.1)	+4.2%
AKS (CVPR 2025)	BLIP	32	53.0 (+3.5)	59.3 (+2.6)	49.3 (+1.2)	58.4 (+0.1)	43.5 (+4.9)	66.8 (+3.1)	+5.5%
MDP^3 (ICCV 2025)	SigLIP	32	53.2 (+3.7)	59.0 (+2.2)	48.8 (+0.7)	58.5 (+0.2)	41.6 (+3.0)	68.2 (+4.5)	+5.3%
Q-Frame (ICCV 2025)	LongCLIP	32	54.0 (+4.5)	59.4 (+2.6)	48.3 (+0.2)	58.4 (+0.1)	44.3 (+5.7)	68.5 (+4.8)	+6.8%
FOCUS (ICLR 2026)	BLIP-ITM	32	54.8 (+5.3)	60.7 (+3.9)	48.5 (+0.4)	58.3 (+0.0)	47.5 (+8.9)	67.1 (+3.4)	+7.1%
AdaQ (arXiv)	LongCLIP	32	55.3 (+5.8)	60.0 (+3.2)	49.0 (+0.9)	59.5 (+1.2)	44.8 (+6.2)	69.4 (+5.7)	+8.2%
RIDGE (EMNLP 2026)	BLIP-ITM	32	55.0 (+5.5)	60.1 (+3.3)	**50.2** (+2.1)	59.6 (+1.3)	**48.9** (+10.3)	68.0 (+4.3)	+8.6%
COMPASS (Ours)	CLIP	32	54.8 (+5.3)	60.5 (+3.7)	49.8 (+1.7)	59.8 (+1.5)	48.2 (+9.6)	69.8 (+6.1)	+8.8%
COMPASS (Ours)	LongCLIP	32	**55.9** (+6.4)	**61.2** (+4.4)	50.0 (+1.9)	**60.2** (+1.9)	48.5 (+9.9)	**70.3** (+6.6)	**+9.8%**
LLaVA-Video-7B	–	64	49.6	58.9	53.2	64.2	41.9	69.5	–
Top-k	CLIP	64	54.6 (+5.0)	59.5 (+0.6)	53.2 (+0.0)	64.4 (+0.2)	46.9 (+5.0)	71.2 (+1.7)	+3.9%
FRAG (arXiv 2025)	MLLM (7B)	64	53.0 (+3.4)	60.6 (+1.7)	52.4 (-0.8)	63.7 (-0.5)	43.1 (+1.2)	69.2 (-0.3)	+0.6%
OneClip-RAG (arXiv 2025)	CLIP	64	56.2 (+6.6)	62.5 (+3.6)	54.0 (+0.8)	65.2 (+1.0)	48.2 (+6.3)	71.2 (+1.7)	+6.3%
BOLT (CVPR 2025)	CLIP	64	54.8 (+5.2)	62.2 (+3.3)	53.8 (+0.6)	64.6 (+0.4)	44.5 (+2.6)	70.3 (+0.8)	+2.5%
AKS (CVPR 2025)	BLIP	64	54.7 (+5.1)	62.7 (+3.8)	55.0 (+1.8)	65.3 (+1.1)	47.6 (+5.7)	71.8 (+2.3)	+6.3%
MDP^3 (ICCV 2025)	SigLIP	64	54.2 (+4.6)	61.8 (+2.9)	54.5 (+1.3)	65.1 (+0.9)	46.0 (+4.1)	72.8 (+3.3)	+5.6%
Q-Frame (ICCV 2025)	LongCLIP	64	56.9 (+7.3)	61.5 (+2.6)	53.9 (+0.7)	64.7 (+0.5)	47.1 (+5.2)	72.4 (+2.9)	+5.4%
FOCUS (ICLR 2026)	BLIP-ITM	64	56.5 (+6.9)	62.9 (+4.0)	54.2 (+1.0)	65.0 (+0.8)	48.8 (+6.9)	72.0 (+2.5)	+6.1%
AdaQ (arXiv 2026)	LongCLIP	64	57.4 (+8.0)	63.2 (+4.3)	54.4 (+1.2)	65.3 (+1.1)	49.1 (+7.2)	74.2 (+4.7)	+8.2%
RIDGE (EMNLP 2026)	BLIP-ITM	64	57.6 (+8.0)	63.6 (+4.7)	**55.2** (+2.0)	65.8 (+1.6)	**51.2** (+9.3)	73.6 (+4.1)	+8.6%
COMPASS (Ours)	CLIP	64	57.5 (+7.9)	63.8 (+4.9)	55.1 (+1.9)	66.1 (+1.9)	50.8 (+8.9)	74.4 (+4.9)	+8.7%
COMPASS (Ours)	LongCLIP	64	**58.3** (+8.7)	**64.5** (+5.6)	55.1 (+1.9)	**66.4** (+2.2)	51.0 (+9.1)	**75.1** (+5.6)	**+9.9%**
Qwen3-VL-8B	–	64	50.7	62.2	56.9	67.6	43.6	71.0	–
Top-k	CLIP	64	57.6 (+6.9)	64.1 (+1.9)	57.3 (+0.4)	67.7 (+0.1)	49.2 (+5.6)	73.2 (+2.2)	+4.8%
FRAG (arXiv 2025)	MLLM (7B)	64	55.2 (+4.5)	64.5 (+2.3)	56.2 (-0.7)	67.2 (-0.4)	46.8 (+3.2)	71.8 (+0.8)	+2.3%
OneClip-RAG (arXiv 2025)	CLIP	64	58.3 (+7.6)	65.1 (+2.9)	56.8 (-0.1)	67.3 (-0.3)	50.5 (+6.9)	71.5 (+0.5)	+5.2%
BOLT (CVPR 2025)	CLIP	64	56.8 (+6.1)	64.6 (+2.4)	57.5 (+0.6)	68.0 (+0.4)	46.5 (+2.9)	73.8 (+2.8)	+3.9%
AKS (CVPR 2025)	BLIP	64	56.9 (+8.0)	65.2 (+2.4)	59.0 (+2.1)	68.6 (+1.0)	49.0 (+5.6)	74.2 (+3.2)	+5.8%
MDP^3 (ICCV 2025)	SigLIP	64	57.1 (+6.4)	64.8 (+2.6)	58.2 (+1.3)	68.4 (+0.8)	47.8 (+4.2)	74.5 (+3.5)	+5.0%
Q-Frame (ICCV 2025)	LongCLIP	64	58.3 (+7.6)	65.0 (+2.8)	57.2 (+0.3)	67.9 (+0.3)	50.2 (+6.6)	74.7 (+3.7)	+6.3%
FOCUS (ICLR 2026)	BLIP-ITM	64	58.1 (+7.4)	65.8 (+3.6)	58.5 (+1.6)	68.3 (+0.7)	51.2 (+7.6)	74.1 (+3.1)	+6.8%
AdaQ (arXiv 2026)	LongCLIP	64	59.8 (+9.1)	66.9 (+4.7)	59.4 (+2.5)	69.6 (+2.0)	51.4 (+7.8)	76.4 (+5.4)	+9.0%
RIDGE (EMNLP 2026)	BLIP-ITM	64	59.5 (+8.8)	66.5 (+4.3)	**59.8** (+2.9)	69.2 (+1.6)	**53.2** (+9.6)	75.8 (+4.8)	+9.2%
COMPASS (Ours)	CLIP	64	59.7 (+9.0)	66.8 (+4.6)	59.5 (+2.6)	69.8 (+2.2)	52.8 (+9.2)	76.5 (+5.5)	+9.3%
COMPASS (Ours)	LongCLIP	64	**60.6** (+9.9)	**67.6** (+5.4)	59.7 (+2.8)	**70.2** (+2.6)	53.0 (+9.4)	**77.2** (+6.2)	**+10.4%**

RIDGE on complex reasoning benchmarks like MLVU and LongVideoBench Overall, while remaining competitive on LVBench. These gains persist across different visual backbones (CLIP and LongCLIP) while operating in a fully deterministic, training-free manner. To eliminate potential bias from differing visual backbones, we conduct controlled experiments where the visual encoder is held identical across all methods (Table 5 in Appendix D.1). In addition, Figure 5 in Appendix D.2 illustrates the radar profiles and accuracy gains across all four benchmarks.

4.2.2 Component-Wise Ablation Studies

Table 2 reports ablation results for Qwen3-VL-8B ($K = 64$) across all four benchmarks. Omitting Hierarchical Zoom (Variant g) causes the largest drop (-2.5% on average, -2.7% on LVBench), confirming that single-pass global selection dilutes critical temporal evidence. Replacing CSG with uniform slicing (Variant c) and using a static threshold (Variant d) degrade accuracy by 2.1% and 1.0%, respectively, demonstrating that topology-aware partitioning preserves coherent action segments. Within the submodular objective, removing coverage (Variant e, -1.8%) leads to redundant frame clusters, while omitting smoothness (Variant f, -1.2%) compromises temporal continuity. Adopting a static intent plan (Variant a, -1.6%) or uniform budgeting (Variant h, -1.2%) similarly degrades performance, whereas LLM-based parsing (Variant b, $+0.2\%$, via zero-shot Qwen2.5-7B-

Table 2: **Component-wise ablation study** on Qwen3-VL-8B (K=64) with LongCLIP visual embeddings. Each block isolates one design formulated in Section 3. Δ is relative to the full COMPASS pipeline.

Module	Variant Configuration	LVB	MME	LVBch	MLVU	Avg.	Δ
COMPASS (Full Pipeline)		**67.6**	**70.2**	**53.0**	**77.2**	**67.0**	–
IQP	(a) *w/o* Intent Plan (Fixed λ)	66.0	69.1	51.2	75.3	65.4	−1.6
	(b) IQP via Text-LLM	67.8	70.3	53.1	77.4	67.2	+0.2
CSG	(c) *w/o* CSG (Uniform Slice)	65.5	68.6	50.8	74.8	64.9	−2.1
	(d) *w/o* Adapt. Thresh. (θ_{scene})	66.4	69.4	51.9	76.1	66.0	−1.0
SS	(e) *w/o* Coverage (λ_{cov}=0)	65.7	68.9	51.1	75.1	65.2	−1.8
	(f) *w/o* Smoothness (λ_{smooth}=0)	66.3	69.3	51.6	76.0	65.8	−1.2
HZ	(g) *w/o* Hierarchical Zoom	65.1	68.2	50.3	74.4	64.5	−2.5
	(h) *w/o* Richness Budgeting	66.2	69.2	51.7	75.9	65.8	−1.2

Table 3: **Transferability comparison across four MLLM backbones** under a unified budget of K=32 frames with standard CLIP.

Backbone MLLM	Frame Selector	LVB	MME	LVBch	MLVU	Avg.	Δ_{uni}
LLaVA-OneVision -7B (Li et al., 2025)	Uniform Sampling	56.8	58.3	38.6	63.7	54.4	–
	Top-k (Radford et al., 2021)	58.3	58.9	43.8	66.9	57.0	+2.6
	RIDGE (Xu et al., 2026)	59.8	59.4	**48.6**	67.8	58.9	+4.5
	COMPASS (Ours)	**60.5**	**59.8**	48.2	**69.8**	**59.6**	**+5.2**
Qwen2.5-VL -7B (Bai et al., 2025)	Uniform Sampling	58.9	61.2	38.5	59.7	54.6	–
	Top-k (Radford et al., 2021)	60.0	63.2	39.3	68.4	57.7	+3.1
	AdaQ (Zhang et al., 2026)	62.4	64.0	44.2	69.8	60.1	+5.5
	COMPASS (Ours)	**64.5**	**65.2**	**48.4**	**71.8**	**62.5**	**+7.9**
InternVL-3 -8B (Zhu et al., 2025)	Uniform Sampling	58.5	65.6	43.5	68.4	59.0	–
	Top-k (Radford et al., 2021)	59.2	66.1	44.2	70.8	60.1	+1.1
	RIDGE (Xu et al., 2026)	61.5	66.2	**53.8**	73.5	63.8	+4.8
	COMPASS (Ours)	**62.8**	**66.8**	53.0	**75.2**	**64.5**	**+5.5**
Qwen3-VL -8B (Team, 2025)	Uniform Sampling	59.8	64.5	40.6	67.1	58.0	–
	Top-k (Radford et al., 2021)	62.2	66.1	43.4	70.4	60.5	+2.5
	AdaQ (Zhang et al., 2026)	63.8	67.7	46.8	72.0	62.6	+4.6
	COMPASS (Ours)	**65.4**	**68.6**	**51.2**	**75.5**	**65.2**	**+7.2**

Instruct) yields negligible gains, confirming that rule-based IQP effectively captures intent without latency or parameter overhead. Appendices D.3 and D.4 (Tables 6, 7, and 8) provide hyperparameter sweeps, held-out validation, and ablations under tighter budgets (K=32).

4.2.3 Generalization Across Downstream MLLM Backbones

To evaluate cross-architecture generalization, Table 3 reports the performance of COMPASS across four open-source MLLM families (LLaVA-OneVision-7B (Li et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025), InternVL-3-8B (Zhu et al., 2025), and Qwen3-VL-8B (Team, 2025)) under a unified budget (K=32, 1 FPS, CLIP visual scorer). Across all backbones, COMPASS consistently outperforms heuristic and signal-based baselines, including Top-k, AdaQ, and RIDGE, achieving absolute gains of $+5.2\%$– $+$ 7.9% over uniform sampling. These gains persist on stronger spatiotemporal models such as InternVL-3-8B ($+5.5\%$) and Qwen3-VL-8B ($+7.2\%$), indicating that front-end keyframe selection complements model scaling: stronger reasoning capabilities cannot compensate for missing or redundant visual evidence.

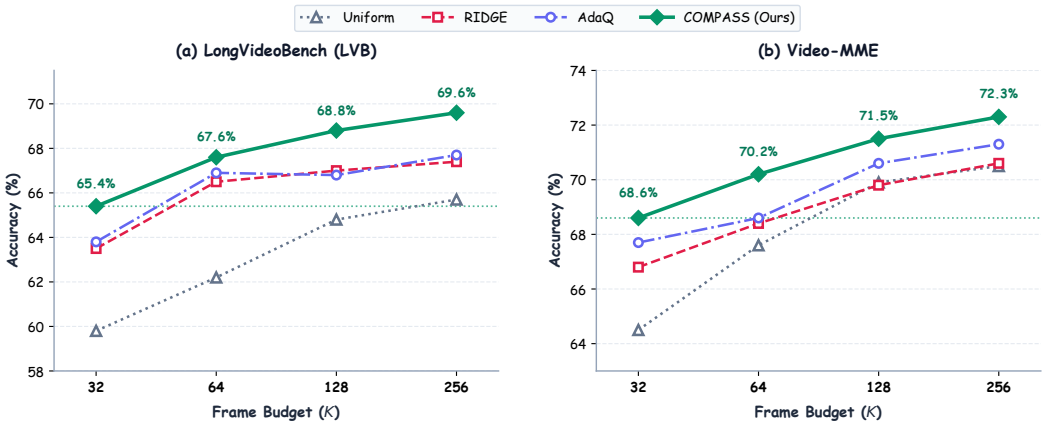

Figure 3: **Reasoning accuracy across frame budgets** ($K \in \{32, 64, 128, 256\}$).

4.2.4 Performance Scaling Across Frame Budgets

Figure 3 plots reasoning accuracy across frame budgets ($K \in \{32, 64, 128, 256\}$) under the settings of Tables 1 and 3. At K=32, COMPASS achieves 65.4% on LVB and 68.6% on Video-MME, outperforming uniform sampling with K=128 on LVB (64.8%) and K=64 on Video-MME (67.6%), thereby reducing visual token overhead by up to 75% ($4\times$ compression at matched accuracy). While baselines such as RIDGE and AdaQ plateau beyond K=64 due to visual noise and cluster redundancy, COMPASS scales steadily to 69.6% on LVB and 72.3% on Video-MME. Across all budgets, jointly optimizing submodular coverage and temporal smoothness maintains a clear advantage, enabling up to $8\times$ compression relative to dense 256-frame sampling. Appendices E.1 and E.2 (Figures 6 and 7) provide stratified analyses across question types and video durations, as well as latency profiling under tight budgets (K=16).

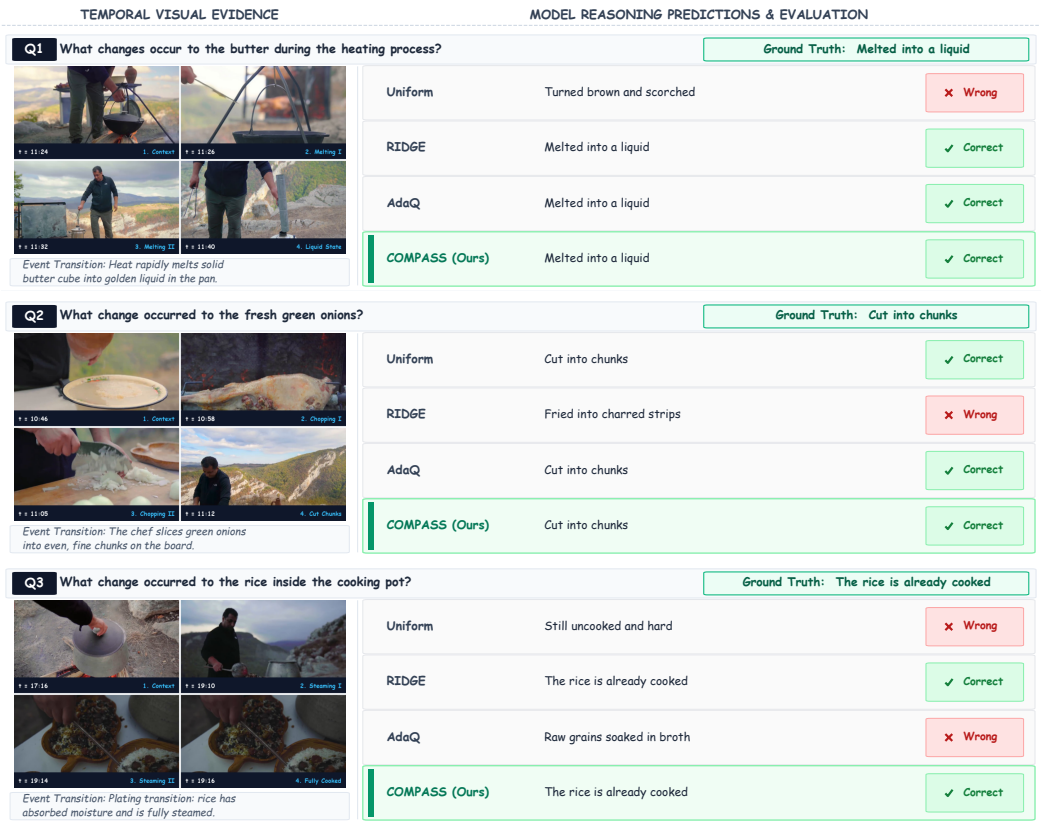

Figure 4: **Qualitative comparison on representative long-video reasoning tasks from LongVideoBench (Wu et al.).**

4.2.5 Qualitative Case Study and Error Analysis

Figure 4 illustrates the impact of keyframe selection on downstream reasoning in Qwen3-VL-8B across three LongVideoBench scenarios. For subtle physical transitions (Q1, butter melting at 11:24–11:40), uniform sampling with coarse strides misses the critical interval, leading to an incorrect prediction of scorched butter; in contrast, COMPASS preserves contiguous state progression through temporal smoothness regularization. For fast-paced actions (Q2, chopping green onions), RIDGE's 1D derivative filter discards rapid transitions as noise, whereas COMPASS structures the sequence on the Content Structure Graph (CSG) to retain dynamic bursts as coherent segments. For long-range causal reasoning (Q3, cooked rice), linking early preparation (17:16) to final plating (19:16) fails under uniform sampling due to omitted steps, and under AdaQ because early visual variance prematurely depletes the frame budget. Maximizing facility-location coverage enables COMPASS to reliably anchor both stages. Appendix E.3 (Figure 8) provides an extended timeline visualization across a 20-minute video.

5 Conclusion

In this paper, we presented COMPASS, a deterministic and training-free framework for question-driven keyframe selection in long-form video understanding. By integrating Intent-guided Query Planning (IQP), Content Structure Graph Submodular Selection (CSG-SS), and Hierarchical Zoom (HZ), COMPASS jointly optimizes query relevance, semantic coverage, and temporal coherence under a monotone submodular objective whose fixed-window greedy step has a theoretical $(1-1/e)$ approximation guarantee. Evaluations across four benchmarks indicate that COMPASS consistently outperforms 11 baseline selectors across diverse MLLMs, achieving relative improvements of up to 10.4% on Qwen3-VL-8B. Furthermore, it also reduces visual token requirements by up to 75%, achieving $4\times$ compression at matched accuracy and up to $8\times$ compression relative to dense 256-frame uniform sampling, while incurring only 8.6 ms of CPU selection latency without auxiliary models.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with language translation and phrasing refinement (tasks with required disclosure). We have not used generative AI tools for other required disclosure tasks, including generating synthetic datasets, formulating theoretical models or mathematical claims, proving theorems, designing experimental methodologies, or interpreting empirical results; all other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure, specifically editing the manuscript to improve readability, formatting LaTeX tables, and assisting with code refactoring for evaluation scripts. We have thoroughly reviewed all AI-assisted outputs. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, Algorithm 1 in the Appendix specifies the COMPASS workflow. Section 3.4 defines the submodular objective, and Appendix A proves the fixed-window greedy approximation guarantee and clarifies its scope. Appendix C provides candidate-frame extraction procedures, the complete hyperparameter configuration (Table 4), and computing environment specifications. The four evaluation benchmarks (LongVideoBench, Video-MME, LVBench, and MLVU) are publicly accessible, and the held-out parameter validation protocol is documented in Appendix D.3. The anonymized code repository linked in the abstract contains the keyframe selection pipeline, reproduction scripts, and evaluation routines.

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, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Tao Chen, Shaobo Ju, Qiong Wu, Chenxin Fang, Kun Zhang, Jun Peng, Hui Li, Yiyi Zhou, and Rongrong Ji. Towards effective and efficient long video understanding of multimodal large language models via one-shot clip retrieval. *CoRR*, abs/2512.08410, 2025. doi: 10.48550/ARXIV.2512.08410. URL https://doi.org/10.48550/arXiv.2512.08410.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-LLM based video frame selection for efficient video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13702–13712. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01279. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hu_M-LLM_Based_Video_Frame_Selection_for_Efficient_Video_Understanding_CVPR_2025_paper.html.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. FRAG: frame selection augmented generation for long video and long document understanding. *CoRR*, abs/2504.17447,

2025. doi: 10.48550/ARXIV.2504.17447. URL https://doi.org/10.48550/arXiv.2504.17447.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

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

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection. *CoRR*, abs/2407.03104, 2024. doi: 10.48550/ARXIV.2407.03104. URL https://doi.org/10.48550/arXiv.2407.03104.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. BOLT: boost large vision-language model without training for long-form video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 3318–3327. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00315. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_BOLT_Boost_Large_Vision-Language_Model_Without_Training_for_Long-form_Video_CVPR_2025_paper.html.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.679. URL https://doi.org/10.18653/v1/2024.acl-long.679.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. Video-language understanding: A survey from model architecture, model training, and data perspectives. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 3636–3657. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.217. URL https://doi.org/10.18653/v1/2024.findings-acl.217.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/shen25j.html.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 26160–26169. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02436. URL https://openaccess.thecvf.com/content/CVPR2025/html/Shu_Video-XL_Extra-Long_Vision_Language_Model_for_Hour-Scale_Video_Understanding_CVPR_2025_paper.html.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 18221–18232. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01725. URL https://doi.org/10.1109/CVPR52733.2024.01725.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp^3: a training-free approach for list-wise frame selection in video-llms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 24090–24101. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02233. URL https://doi.org/10.1109/ICCV51701.2025.02233.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 29118–29128. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02711. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.html.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. Lvbench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22958–22967. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02131. URL https://doi.org/10.1109/ICCV51701.2025.02131.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Shanqing Xu, Meng Luo, Mengchen Qian, Yuhui Gao, Siyue Peng, Xiaohan Zhong, Xiaojin Zhang, Zhongyu Wei, Wei Chen, and Xiang Bai. Ridge: Region-informed derivative-guided evidence selection for long video understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 11941–11952. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01100. URL https://doi.org/10.1109/ICCV51070.2023.01100.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New*

Mexico, USA, April 29 - May 4, 2025, volume NAACL 2025 of *Findings of ACL*, pp. 881–916. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.FINDINGS-NAACL.51. URL https://doi.org/10.18653/v1/2025.findings-naacl.51.

Kun Zhang, Chenxin Fang, Tao Chen, Baiyang Song, Yunhang Shen, Yiyi Zhou, and Rongrong Ji. Towards fast and effective long video understanding of multimodal large language models via adaptive quasi-gaussian sampling. *CoRR*, abs/2606.24187, 2026. doi: 10.48550/ARXIV.2606.24187. URL https://doi.org/10.48550/arXiv.2606.24187.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2025b. URL https://openreview.net/forum?id=30RAWQVGlx.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 22056–22065. IEEE, 2025c. doi: 10.1109/ICCV51701.2025.02048. URL https://doi.org/10.1109/ICCV51701.2025.02048.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *Trans. Mach. Learn. Res.*, 2025d. URL https://openreview.net/forum?id=EElFGvt39K.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13691–13701. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01278. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

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. *CoRR*, abs/2504.10479, 2025. doi: 10.48550/ARXIV.2504.10479. URL https://doi.org/10.48550/arXiv.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. *CoRR*, abs/2510.04428, 2025. doi: 10.48550/ARXIV.2510.04428. URL https://doi.org/10.48550/arXiv.2510.04428.

A SCOPE AND PROOF OF THE SELECTION GUARANTEE

Fix one window with nonempty candidate domain $\mathcal{U}$, budget $1 \le B \le |\mathcal{U}|$, fixed anchors $\mathcal{A} \neq \emptyset$, and fixed temporal grid $\mathcal{T} \neq \emptyset$. Assume $\|\mathbf{f}_i\|_2 = \|\mathbf{a}_j\|_2 = 1$, $w_i \in [0,1]$, $\sigma_t > 0$, and $\lambda_{\text{rel}}, \lambda_{\text{cov}}, \lambda_{\text{smooth}} \ge 0$. These conditions are satisfied by normalized embeddings and the scope weights in Eq. equation 1; a degenerate empty window is skipped. Write

$$c_j(i) = \max\{0, \langle \mathbf{f}_i, \mathbf{a}_j \rangle\}, \qquad h_k(i) = \exp(-|t_i - \tau_k|/\sigma_t), \qquad \max_{i \in \emptyset}(\cdot) = 0. \tag{6}$$

Both $c_j(i)$ and $h_k(i)$ belong to $[0,1]$. For any nonnegative score $v(i)$, let $m_v(S) = \max_{i \in S} v(i)$. Its marginal gain is

$$\Delta_v(u \mid S) = m_v(S \cup \{u\}) - m_v(S) = \max\{0, v(u) - m_v(S)\}. \tag{7}$$

If $A \subseteq C$ and $u \notin C$, then $m_v(A) \le m_v(C)$ and therefore $0 \le \Delta_v(u \mid C) \le \Delta_v(u \mid A)$. Thus each maximum is normalized, monotone, and submodular. The relevance marginal is $w_u \ge 0$, independent of the current set. Summing the maxima and multiplying all three terms by nonnegative weights proves the same properties for F. Moreover, for $|S| \le B$,

$$\begin{aligned} &0 \le \widetilde{\text{Rel}}(S) \le B, \\ &0 \le \widetilde{\text{Cov}}(S) \le |\mathcal{A}|, \\ &0 \le \widetilde{\text{Smooth}}(S) \le |\mathcal{T}|, \\ &0 \le F(S) \le \lambda_{\text{rel}} B + \lambda_{\text{cov}} |\mathcal{A}| + \lambda_{\text{smooth}} |\mathcal{T}|. \end{aligned} \tag{8}$$

The bound on F is restricted to feasible sets because the relevance term can exceed B when $|S| > B$.

Let O be an optimal feasible set and G_r the greedy set after r steps, $G_0 = \emptyset$. Monotonicity and submodularity give

$$F(O) - F(G_r) \le F(O \cup G_r) - F(G_r) \le \sum_{u \in O \setminus G_r} \Delta_F(u \mid G_r) \le B \max_{u \in \mathcal{U} \setminus G_r} \Delta_F(u \mid G_r). \tag{9}$$

Consequently the greedy choice satisfies $F(G_{r+1}) - F(G_r) \ge [F(O) - F(G_r)]/B$. With $D_r = F(O) - F(G_r)$, induction yields $D_B \le (1 - 1/B)^B F(O) \le e^{-1} F(O)$, proving Theorem 1. This is an approximation to the stated surrogate F under a fixed cardinality budget; it does not establish a corresponding bound on downstream MLLM accuracy.

The window proposal and quota allocation change the feasible domains and objectives. Their independent greedy guarantees cannot be combined into a global $(1 - 1/e)$ guarantee for the final selection. In particular, temporal NMS or pruning can delete a frame with positive marginal utility, and relevance backfilling need not maximize the same objective. If $\sum_j B_j = K$ and $|S_j| \le B_j$, then

$$\left| \bigcup_j S_j \right| \le \sum_j |S_j| \le \sum_j B_j = K, \tag{10}$$

so deduplication and NMS alone cannot trigger an over-budget branch. The guarantee applies before these reconciliation operations, while the final method remains a heuristic composition of locally guaranteed steps.

B PSEUDOCODE OF COMPASS

Algorithm 1 COMPASS Keyframe Selection Pipeline

Require: Candidate frames $\mathcal{V} = \{f_t\}_{t=1}^{N}$, Query Q, Budget K, Lower bound $K_{\min}$.
Ensure: Selected keyframe indices S^*, Calibrated importance weights $\boldsymbol{\gamma}$.

Compute visual embeddings $\mathbf{F} \in \mathbb{R}^{N\times D}$, query embedding $\mathbf{g}_Q \in \mathbb{S}^{D-1}$, and normalized relevance $\mathbf{w} \in [0,1]^N$.
// Step 1: Intent-guided Query Planning
$\mathcal{P} = (\alpha_{\text{scope}}, \mathcal{C}_{\text{need}}, \tau_{\text{time}}, \delta_{\text{detail}}) \leftarrow \Pi(Q)$.
Retrieve submodular weights $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}(\alpha_{\text{scope}})$ via Eq. equation 1.
// Step 2: Content Structure Graph Construction
Partition $\mathcal{V}$ into scene segments $\mathcal{S} = \{\mathcal{S}_m\}_{m=1}^{M}$ using adaptive threshold $\theta = \mu_d + k_{\text{scene}}\sigma_d$.
Compute structural descriptors $\{\mathbf{c}_m, r_m, \mathcal{M}_m, \mathcal{H}_m, \mathcal{E}_m\}_{m=1}^{M}$ for all scene segments in $\mathcal{S}$.
// Step 3: Hierarchical Window Proposals and Budgeting
$\mathcal{W} = \{W_j\}_{j=1}^{|\mathcal{W}|} \leftarrow \text{ProposeWindows}(\mathcal{P}, \{\mathcal{M}_m\}, N)$ clipped by τ_{time}.
Allocate discrete budgets $\{B_j\}_{j=1}^{|\mathcal{W}|} \leftarrow \text{AllocateBudget}(\mathcal{W}, \{\mathcal{H}_m, \mathcal{E}_m\}, K)$.
// Step 4: Intra-Window Submodular Maximization
Initialize pooled sets $S_{\text{raw}} \leftarrow \emptyset$, $\boldsymbol{\gamma}_{\text{raw}} \leftarrow \emptyset$.
for all $W_j \in \mathcal{W}$ **with** $B_j > 0$ **do**
  Define window domain $\mathcal{U}_j \leftarrow W_j \cap \mathcal{U}$ and anchors $\mathcal{A}_j \leftarrow \{\mathbf{f}_{r_m} \mid \mathcal{S}_m \cap W_j \neq \emptyset\}$.
  $S_j, \boldsymbol{\gamma}_j \leftarrow \text{SubmodularGreedy}(\mathcal{U}_j, \mathbf{w}, \mathbf{F}, \mathcal{A}_j, B_j, \boldsymbol{\lambda})$ via Eq. equation 2.
  $S_{\text{raw}} \leftarrow S_{\text{raw}} \cup S_j, \quad \boldsymbol{\gamma}_{\text{raw}} \leftarrow \boldsymbol{\gamma}_{\text{raw}} \cup \boldsymbol{\gamma}_j$.
end for
// Step 5: Deduplication, Temporal NMS, and Regulation
$S_{\text{uniq}}, \boldsymbol{\gamma}_{\text{uniq}} \leftarrow \text{DeduplicateAndSort}(S_{\text{raw}}, \boldsymbol{\gamma}_{\text{raw}})$.
$S_{\text{nms}}, \boldsymbol{\gamma}_{\text{nms}} \leftarrow \text{TemporalNMS}(S_{\text{uniq}}, \boldsymbol{\gamma}_{\text{uniq}})$.
if $|S_{\text{nms}}| < K_{\min}$ **then**
  $S^*, \boldsymbol{\gamma} \leftarrow \text{SupplementByRelevance}(S_{\text{nms}}, \boldsymbol{\gamma}_{\text{nms}}, \mathbf{w}, K_{\min})$.
else if $|S_{\text{nms}}| > K$ **then**
  $S^*, \boldsymbol{\gamma} \leftarrow \text{TruncateByImportance}(S_{\text{nms}}, \boldsymbol{\gamma}_{\text{nms}}, K)$. ▷ *Defensive branch; unreachable if quotas sum to* K
else
  $S^* \leftarrow S_{\text{nms}}, \quad \boldsymbol{\gamma} \leftarrow \boldsymbol{\gamma}_{\text{nms}}$.
end if
return S^*, $\text{NormalizeImportance}(\boldsymbol{\gamma})$.

C EXTENDED EXPERIMENTAL SETTINGS AND IMPLEMENTATION DETAILS

C.1 BENCHMARK SUITES AND EVALUATION PROTOCOL

We evaluate COMPASS across four long-video benchmark suites covering diverse durations and reasoning tasks. LongVideoBench (LVB) (Wu et al.) comprises 6,678 multiple-choice questions across 3,763 videos on its validation split, with video durations ranging from seconds up to one hour, primarily diagnosing fine-grained referring and multi-frame causal reasoning. Video-MME (Fu et al., 2025) contains 900 curated videos and 2,700 question-answer pairs evaluated under the subtitle-free video-only setting, categorized into short ($<$ 2 min), medium (2–15 min), and long (15–60 min) durations. LVBench (Wang et al., 2025) focuses on hour-scale videos with 1,549 questions over 103 instances, testing long-horizon event comprehension, key information retrieval, entity recognition, and temporal grounding. Finally, MLVU (Zhou et al., 2025) is evaluated on its development split of 1,730 videos with an average duration of approximately 12 minutes, covering nine diagnostic tasks including anomaly detection, summarization, action ordering, and needle-in-a-haystack retrieval. Across all benchmarks, model answers are generated deterministically using greedy decoding (temperature $= 0$) with standardized multiple-choice evaluation protocols via the `lmms-eval` toolkit (Zhang et al., 2025a).

C.2 Baseline Implementation and Reproductions

All eleven baseline methods are evaluated under matched candidate frame pools sampled at 1 FPS and identical frame budgets of $K \in \{32, 64\}$. Uniform sampling distributes K frames at evenly spaced temporal intervals across the full duration $[0, T)$ without query conditioning. For point-wise retrieval, Top-k selection (Radford et al., 2021) scores frames using cosine similarity $s_t = \langle \mathbf{f}_t, \mathbf{g}_Q \rangle$ and selects the top-K scoring frames sorted chronologically. Clip-level retrieval baselines FRAG (Huang et al., 2025) and OneClip-RAG (Chen et al., 2025) are evaluated using their official implementations based on text-visual similarity matching.

For diversity-oriented and temporal partitioning methods, BOLT (Liu et al., 2025) applies query-guided inverse transform sampling reproduced via its official repository with temperature $\tau_{\text{bolt}} = 0.5$. AKS (Tang et al., 2025) balances temporal coverage and relevance using its default recursive binning configuration, while MDP^3 (Sun et al., 2025) enforces list-wise diversity via determinantal point processes with dynamic programming. Q-Frame (Zhang et al., 2025c) perturbs softmax similarity distributions with Gumbel noise; because it involves stochastic sampling, its results are averaged over three independent runs. For active exploration and signal-based modeling, FOCUS (Zhu et al., 2026) executes a two-stage bandit search over short temporal clips using Bernstein confidence bounds. AdaQ (Zhang et al., 2026) performs quasi-Gaussian sampling scaled by score variance $\mathrm{Var}(s)$ with default scaling factor $\gamma = 0.5$ over a 3-σ interval. Finally, RIDGE (Xu et al., 2026) segments smoothed similarity curves via numerical derivatives using default parameters $\sigma = 2.0$, prominence $p_{\min} = 0.15$, and window half-width $w_0 = 3$.

C.3 COMPASS Parameter Specifications and Execution Flow

Table 4 lists the hyperparameter configurations for COMPASS across each module.

Table 4: Hyperparameter configurations of the COMPASS framework.

Module	Hyperparameter	Default Value
Candidate Extraction	Sampling frame rate	1.0 FPS
	Embedding normalization	ℓ_2 unit sphere $\mathbb{S}^{D-1}$
IQP Planning	Decomposer mode	Deterministic rule-based (Regex/keyword)
	Scope trade-offs $(\lambda_{\text{rel}}, \lambda_{\text{cov}}, \lambda_{\text{smooth}})$	Local: $(1.0, 0.2, 0.2)$; Global: $(0.2, 1.0, 0.8)$; Hybrid: $(0.7, 0.6, 0.5)$
CSG Construction	Scene boundary threshold k_{scene}	2.0 ($\theta_{\text{scene}} = \mu_d + 2.0 \cdot \sigma_d$)
	Minimum scene segment length $L_{\min}$	2 frames
	Centroid anchor selection	Medoid frame $r_m = \arg\max_{t \in \mathcal{S}_m} \langle \mathbf{f}_t, \mathbf{c}_m \rangle$
Hierarchical Zoom (HZ)	Global uniform windows	4 ($J = 4$ if $\delta_{\text{detail}} < 0.5$, else $J = 6$)
	Hybrid zoom window half-width	$N \cdot (0.08 + 0.12 \cdot \delta_{\text{detail}})$
	Richness mixture prior	$0.5 \times$ Richness $+ 0.5 \times$ Uniform
Selection & Regulation	Fixed-window greedy approximation	$(1 - 1/e) \approx 0.632$ lower bound
	Temporal NMS suppression window	0.5 s
	Capacity bounds $[K_{\min}, K]$	$[16, 32]$ for $K = 32$; $[16, 64]$ for $K = 64$

C.4 Hardware and Software Environment

Candidate frame extraction, graph construction, and submodular optimization are executed on standard CPUs using NumPy and PyTorch without GPU requirements or training overhead. Downstream MLLM inference is conducted on NVIDIA RTX 4090 (24 GB) and A100 (80 GB) GPUs using PyTorch 2.4.0, CUDA 12.4, and Transformers 4.45+ on Ubuntu 22.04 LTS.

D Extended Experimental Results and Ablations

D.1 Matched-scorer comparison

In Table 1, baseline selectors use the visual embedding models specified by their default setups (e.g., BOLT uses CLIP, AdaQ uses LongCLIP, RIDGE uses BLIP-ITM, and MDP^3 uses SigLIP). To isolate the impact of the keyframe selection policy from the representational capacity of the visual scorer, we conduct controlled experiments where the vision-language backbone is held identical across methods.

Table 5: **Matched-scorer comparison** on Qwen3-VL-8B (K=64). Within each block, the visual encoder is held identical to isolate the frame selection policy. Bold denotes best.

Scorer	Selection Method	LVB	MME	LVBch	MLVU	Avg.	Δ_{base}
CLIP	Top-k (Radford et al., 2021)	64.1	67.7	49.2	73.2	63.6	–
	BOLT (Liu et al., 2025)	64.6	68.0	46.5	73.8	63.2	−0.4
	COMPASS (Ours)	**66.8**	**69.8**	**52.8**	**76.5**	**66.5**	**+2.9**
LongCLIP	Q-Frame (Zhang et al., 2025c)	65.0	67.9	50.2	74.7	64.5	–
	AdaQ (Zhang et al., 2026)	66.9	69.6	51.4	76.4	66.1	+1.6
	COMPASS (Ours)	**67.6**	**70.2**	**53.0**	**77.2**	**67.0**	**+2.5**
BLIP-ITM	FOCUS (Zhu et al., 2026)	65.8	68.3	51.2	74.1	64.9	–
	RIDGE (Xu et al., 2026)	66.5	69.2	**53.2**	75.8	66.2	+1.3
	COMPASS (Ours)	**67.2**	**69.9**	53.1	**76.8**	**66.8**	**+1.9**
SigLIP	MDP^3 (Sun et al., 2025)	64.8	68.4	47.8	74.5	63.9	–
	COMPASS (Ours)	**66.9**	**69.7**	**52.5**	**76.9**	**66.5**	**+2.6**

Table 5 compares COMPASS against the baselines natively paired with each of the four vision-language backbones on Qwen3-VL-8B (K=64) across all four benchmarks. Across all configurations, COMPASS consistently achieves higher accuracy than the corresponding baseline selectors. Under standard CLIP embeddings, COMPASS achieves an average accuracy of 66.5%, outperforming Top-k (63.6%) and BOLT (63.2%) by 2.9% and 3.3%, respectively. When paired with LongCLIP, which accommodates longer text prompts, COMPASS attains 67.0% average accuracy, surpassing Q-Frame (64.5%) and AdaQ (66.1%). Under BLIP-ITM cross-modal matching features, COMPASS reaches 66.8%, exceeding FOCUS (64.9%) and RIDGE (66.2%). Finally, using SigLIP embeddings, COMPASS achieves 66.5%, outperforming MDP^3 (63.9%) by 2.6%.

These controlled comparisons highlight two clear trends. COMPASS achieves the highest average accuracy across all four scorer groups, indicating that its performance gains stem from query planning and submodular optimization rather than coupling with a particular visual encoder. In addition, algorithmic design proves more impactful than visual backbone scaling. Even when paired with the standard CLIP model, COMPASS achieves an average accuracy of 66.5%, exceeding baselines that use specialized or extended backbones, such as MDP^3 on SigLIP (63.9%) and Q-Frame on LongCLIP (64.5%). This suggests that modeling video topology and enforcing set-level coverage provides greater benefits than simply upgrading the visual scorer.

D.2 Cross-Dataset Consistency and Robustness

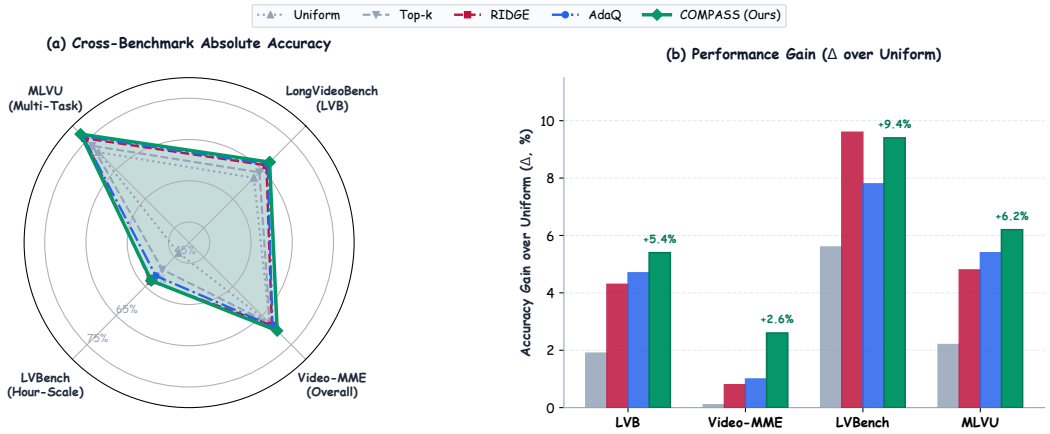

Figure 5: **Cross-benchmark performance consistency** on Qwen3-VL-8B (K=64) across LongVideoBench (LVB), Video-MME, LVBench, and MLVU. Results align with Table 1. (a) Absolute accuracy profiles across benchmarks. (b) Absolute accuracy improvement over uniform sampling (Δ, %).

Figure 5 evaluates performance consistency across four benchmarks covering different video durations and task categories. As shown in Figure 5(a), baseline selectors exhibit clear performance trade-offs across evaluation suites. For example, Top-k accuracy drops on hour-scale videos in LVBench (49.2%), and AdaQ shows lower accuracy on long sequences (51.4%). In contrast, COMPASS achieves the highest accuracy on LVB (67.6%), Video-MME (70.2%), and MLVU (77.2%), while performing comparably to RIDGE on LVBench (53.0% vs. 53.2%). In addition, Figure 5(b) shows that COMPASS provides consistent improvements over uniform sampling, with absolute gains ranging from 2.6% to 9.4% across all four benchmarks. This stability indicates that combining query planning with topological video structuring generalizes reliably without benchmark-specific parameter tuning.

D.3 Hyperparameter Sensitivity and Held-Out Validation

Hyperparameter sensitivity analysis. To assess the robustness of COMPASS to its operational parameters, we conduct sensitivity experiments on Qwen3-VL-8B (K=64) across LongVideoBench and MLVU by varying one hyperparameter at a time while holding all others fixed. First, varying the scene boundary multiplier $k_{\text{scene}} \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ in $\theta_{\text{scene}} = \mu_d + k_{\text{scene}} \cdot \sigma_d$ controls segment granularity. While small values of k_{scene} over-partition minor camera motions into fragmented sub-scenes and large values merge dynamic transitions, performance remains stable across the full range, fluctuating within 0.8% (Table 6). Second, sweeping the temporal NMS window $\tau_{\text{nms}} \in \{0.2\,\text{s}, 0.5\,\text{s}, 1.0\,\text{s}, 1.5\,\text{s}\}$ shows consistent accuracy, with minor degradation observed only at larger intervals (≥ 1.5 s) where rapid actions risk being discarded. Finally, adjusting the capacity floor $K_{\min} \in \{8, 12, 16, 20\}$ preserves stable accuracy between 71.8% and 72.4%, confirming that the lower bound prevents window starvation without distorting global frame allocation. Across all tested configurations, COMPASS consistently maintains a clear advantage over uniform sampling (58.0% on LVB and 67.1% on MLVU), indicating that performance gains stem from the submodular optimization framework rather than sensitive parameter tuning.

Held-out development validation. To verify that COMPASS does not overfit to specific evaluation benchmarks, we evaluate its parameters under a held-out calibration protocol. We first construct a video-disjoint split on LongVideoBench by isolating a 20% development subset (LVB-dev) from the remaining 80% evaluation split (LVB-held-out). We then perform a grid search over 16 parameter configurations exclusively on LVB-dev. The resulting configuration (k_{scene}=1.8, $K_{\min}$=16, τ_{nms}=0.4 s) is frozen and evaluated directly on LVB-held-out, Video-MME, LVBench, and MLVU without further adjustment. As shown in Table 7, this held-out configuration achieves performance comparable to the default configuration (67.0% vs. 67.0% four-benchmark average) while outper-

Table 6: **Hyperparameter sensitivity** of COMPASS on Qwen3-VL-8B (K=64) across LongVideoBench and MLVU. One parameter is varied at a time while others remain at default values (gray).

Varied Parameter	Setting	LVB	MLVU	Avg.	Δ_{default}
Default Configuration	–	**67.6**	**77.2**	**72.4**	–
Boundary Mult. k_{scene}	1.0	66.8	76.5	71.7	−0.7
	1.5	67.3	76.9	72.1	−0.3
	2.5	67.2	77.0	72.1	−0.3
	3.0	66.7	76.4	71.6	−0.8
Temporal NMS τ_{nms}	0.2 s	67.1	76.8	72.0	−0.4
	1.0 s	67.4	76.9	72.2	−0.2
	1.5 s	66.9	76.5	71.7	−0.7
Capacity Floor $K_{\min}$	8	66.9	76.6	71.8	−0.6
	12	67.4	77.0	72.2	−0.2
	20	67.3	77.1	72.2	−0.2

Table 7: **Held-out parameter generalization** on Qwen3-VL-8B (K=64). The configuration is tuned strictly on a 20% disjoint development set (*LVB-dev*), frozen, and evaluated across unseen splits and benchmarks. *Avg.* excludes *LVB-dev*.

Benchmark / Split	Uniform	Default	Held-out
LVB-dev (20% Disjoint Set)	57.5	68.1	**68.4**
LVB-held-out (80% Test)	60.1	**67.5**	67.4
Video-MME	67.6	70.2	**70.4**
LVBench	43.6	53.0	**53.2**
MLVU	71.0	**77.2**	77.1
Avg. (Unseen Benchmarks)	60.6	**67.0**	**67.0**

forming uniform sampling by 6.4%. These results confirm that parameters calibrated on a small disjoint subset generalize across unseen datasets and video distributions.

D.4 Component-Wise Module Ablation

Table 8 evaluates the contribution of each COMPASS component on LongVideoBench (LVB) and Video-MME under a fixed budget of K=32 frames with Qwen3-VL-8B. Removing Content Structure Graph Submodular Selection (CSG-SS) causes the largest performance drop, reducing average accuracy by 3.6% to 63.4%. Under this configuration, selecting frames via standard top-k similarity fails to capture inter-frame dependencies, yielding redundant selections in static scenes while missing key temporal context. In the absence of CSG, HZ falls back to uniform sliding windows centered around top pointwise similarity peaks. This confirms that graph-based submodular optimization ($\mathcal{F}_{\text{cov}} + \mathcal{F}_{\text{smooth}}$) is essential for maintaining both visual diversity and temporal coverage.

Table 8: **Component-wise ablation study** of COMPASS on Qwen3-VL-8B (K=32 frames). We ablate the three core pillars: Intent-guided Query Planning (**IQP**), Content Structure Graph Submodular Selection (**CSG-SS**), and Hot-Zone Zooming (**HZ**). Latency denotes pure CPU selection time on LongVideoBench.

Configuration	IQP	CSG-SS	HZ	LVB	Video-MME	Avg.	Latency
Uniform Baseline	×	×	×	59.8	64.5	62.2	$<$ 1.0 ms
Top-k Ranking	×	×	×	62.2	66.1	64.2	0.8 ms
w/o IQP	×	✓	✓	64.0	67.2	65.6 (-1.4)	8.4 ms
w/o CSG-SS	✓	×	✓	61.8	64.9	63.4 (-3.6)	2.1 ms
w/o HZ	✓	✓	×	64.2	67.4	65.8 (-1.2)	6.2 ms
COMPASS (Full)	✓	✓	✓	**65.4**	**68.6**	**67.0**	**8.6 ms**

Disabling Intent-guided Query Planning (IQP) replaces structured query decomposition with global question-level matching, resulting in a 1.4% decrease in average accuracy. Without structured query guidance, the visual encoder tends to focus on salient foreground objects rather than the specific actions or causal transitions targeted by the query. Omitting Hot-Zone temporal zooming (HZ) further reduces accuracy by 1.2%. While CSG-SS preserves broad coverage across the full timeline, HZ allocates higher frame density to intervals with rapid motion or state changes, ensuring that transient, critical events are captured under constrained sampling budgets. Together, these three components improve accuracy by 4.8% over uniform sampling while adding only 8.6 ms of CPU runtime.

E Extended Visual and Diagnostic Studies

E.1 Stratified Analysis

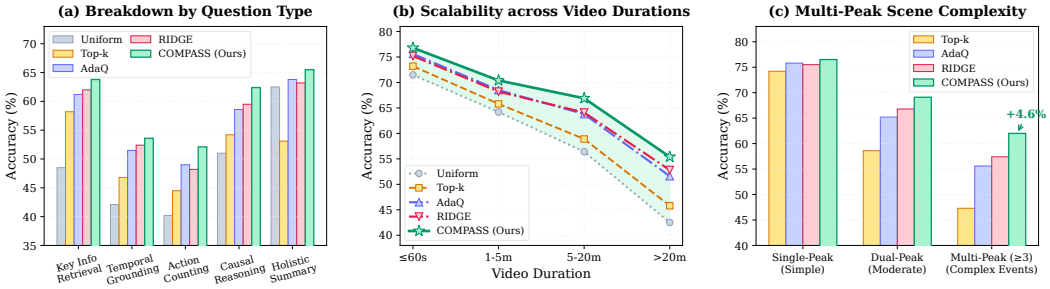

Figure 6: **Stratified evaluation of COMPASS** on Qwen3-VL-8B (K=64). (a) Accuracy breakdown across five representative question categories from MLVU and LVBench. (b) Performance scaling curve across progressive video duration intervals. (c) Accuracy under increasing temporal complexity (single-peak vs. multi-peak event distributions).

Figure 6 illustrates the performance of COMPASS across question categories, video durations, and temporal scene complexity. Across reasoning tasks (Figure 6a), pointwise retrieval (Top-k) performs poorly on Holistic Summarization (53.1%, falling 9.4% below uniform sampling) due to localized frame clustering. In contrast, COMPASS achieves 52.1% on Action Counting and 62.4% on Causal Reasoning, outperforming RIDGE and AdaQ by 3.9% and 2.9%, respectively. This indicates that IQP effectively balances submodular coverage ($\mathcal{F}_{\text{cov}}$) and transition smoothness ($\mathcal{F}_{\text{smooth}}$). Regarding video duration (Figure 6b), uniform sampling degrades to 42.5% on videos longer than 20 min, while COMPASS maintains consistent gains, exceeding uniform sampling by 12.9% and RIDGE by 2.6% through the use of Hierarchical Zoom (HZ) and content-adaptive budgeting. Finally, under increasing temporal complexity (Figure 6c), Top-k drops to 47.3% on multi-peak sequences ($\geq$ 3 peaks) due to over-concentration on the dominant peak. COMPASS reaches 62.0% (4.6% higher than RIDGE), showing that constructing a Content Structure Graph on the visual feature manifold provides balanced, redundancy-free coverage across interleaved sub-events.

E.2 Budget Scalability and Selection Latency

Figure 7 evaluates the sample efficiency and runtime overhead of COMPASS on Qwen3-VL-8B. Across frame budgets (Figure 7a), COMPASS shows the largest gains under constrained budgets, achieving a 7.7% improvement over uniform sampling and 2.3% over AdaQ at K=16, as well as a 6.5% gain over uniform sampling at K=32. As the budget increases to K=128, performance gradually saturates across selectors (67.6% for COMPASS vs. 64.3% for uniform sampling). Regarding selection latency (Figure 7b), we measure CPU post-processing time following the standard LongVideoBench protocol (N=32, excluding shared feature extraction). Pointwise ranking (Top-k) and variance-based sampling (AdaQ) require 0.75–1.0 ms, while recursive binning (AKS) and derivative analysis (RIDGE) take 5.2 ms and 8.1 ms, respectively. In comparison, while RIDGE requires an auxiliary LLM (+0.6 s) for peak performance and iterative frameworks like A.I.R. take over 3.4 s, the default COMPASS pipeline relies on rule-based Intent-guided Query Planning (IQP), requiring no additional LLM queries or iterative model invocations. It completes graph construction and submodular optimization in 8.6 ms on CPU, adding minimal deployment overhead to the downstream inference pipeline.

E.3 Hierarchical Temporal Structuring Analysis

Figure 8 illustrates how COMPASS coordinates frame selection across three temporal granularities: macro scenes, intermediate query-relevant windows, and fine-grained similarity peaks. At the macro level, long videos naturally partition into distinct semantic segments (Scenes 1–3). While uniform sampling dilutes frame budgets across inactive intervals and variance-based methods like AdaQ allocate over 65% of the budget to Scene 2 due to local texture variations, COMPASS regu-

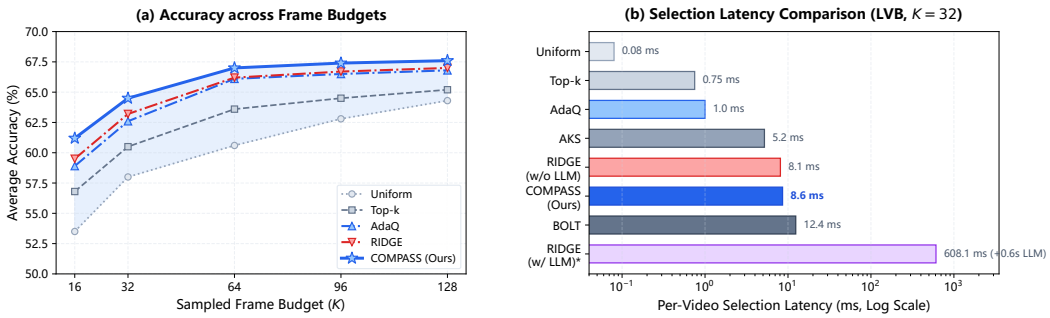

Figure 7: **Budget scalability and selection latency analysis** on Qwen3-VL-8B. (a) Average accuracy across four benchmarks under varying frame budgets $K \in \{16, 32, 64, 96, 128\}$. (b) Per-video selection latency on LongVideoBench (K=32). Methods such as RIDGE (with LLM) and A.I.R. require auxiliary LLM processing (+0.6 s) or iterative model calls (3.4 s), whereas COMPASS runs in 8.6 ms on CPU without external model queries.



Figure 8: **Hierarchical temporal decomposition and frame allocation** across a 20-minute video (T=1200 s). Top: Multi-scale visual structuring across macro scenes (CSG), query-relevant windows (IQP), and local similarity peaks. Middle: Frame selection distributions across methods. Bottom: Selected keyframes within the primary target window (630–750 s), capturing physical state transitions without visual redundancy.

lates allocation through CSG clustering, ensuring each scene retains representative frames. At the intermediate level, Intent-guided Query Planning (IQP) and Hot-Zone zooming (HZ) identify salient temporal windows (630–750 s and 1100–1180 s), concentrating frame density within dynamic intervals that fixed-stride sampling typically misses. At the fine-grained level, where 1D derivative filters such as RIDGE often trigger on local visual noise, COMPASS jointly optimizes facility-location coverage ($\widetilde{\text{Cov}}$) and temporal smoothness ($\widetilde{\text{Smooth}}$). This formulation captures key event anchors ($E_1 \rightarrow E_4$ across the timeline) while retaining precise transition frames ($P_1 \rightarrow P_4$, bottom panel) without introducing redundant frames.