HAS: Hierarchical Adaptive Frame Sampling for Video Understanding

Yuyao Sun[1] **Daiqi Shi**[2] **Shuang Li**[1] **Liao Zhang**[2] **Yang Cai**[2] **Deqing Wang**[1] *
[1]Beihang University [2]Independent Researcher
{sunyuyao,shuangliai,dqwang}@buaa.edu.cn
daiqishi563@gmail.com zliao1992@163.com yangcai1988@gmail.com

Abstract

Due to context-length constraints, most MLLMs cannot process full-length videos and therefore rely on sampling a subset of frames as input. However, existing sampling methods, ranging from uniform sampling to relevance-based selection, are often driven by a single sampling principle and thus struggle to accommodate the heterogeneous evidential requirements of different queries: some demand a holistic understanding of how the video unfolds over time, while others hinge on fine-grained events within a short temporal window. To address this limitation, we propose **Hierarchical Adaptive Frame Sampling (HAS)**, a two-stage frame sampling framework. In the first stage, *Backbone Frame Construction*, we apply a Determinantal Point Process (DPP) to sample frames that are both query-relevant and non-redundant. These selected frames capture key moments across the video and form the backbone of the sampling set, providing a foundation for subsequent enrichment. In the second stage, *Adaptive Contextual Enrichment*, we analyze the temporal distribution of the backbone frames to infer the query's requirements and adaptively allocate the remaining frame budget between Local and Global Context. The Local Context enriches the backbone with fine-grained temporal dynamics and short-range causal relations, whereas the Global Context connects temporally isolated evidence and provides a holistic view of the entire video to support global understanding. Through such a hierarchical two-stage process, HAS effectively addresses diverse query requirements. Incorporated into three leading MLLMs, it demonstrates consistently superior performance across the Video-MME, LongVideoBench, and MLVU benchmarks.

1 Introduction

Multi-modal Large Language Models (MLLMs) have demonstrated impressive performance across a wide range of image perception and reasoning tasks [4, 7, 8, 13, 14, 17]. Extending these capabilities to videos is a natural next step, yet the sheer volume of visual information in videos makes this transition far from trivial. Most current video-based MLLMs [15, 23, 24, 35] follow an image-based paradigm: encoding frames into visual tokens using a vision encoder, and jointly processing these tokens with text tokens in a Large Language Model. However, the constrained context window of LLMs precludes processing all video frames simultaneously; typically, a fixed subset (*e.g.*, 32 or 64) is chosen to represent the entire video. Thus, the frame sampling strategy becomes a pivotal factor influencing the model's understanding.

A common strategy is uniform sampling, which selects frames at fixed intervals. While this approach preserves coarse global coverage of the storyline, it can easily miss brief yet critical moments. To

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

alleviate this issue, relevance-based sampling, as adopted in SeViLA [34] and MVU [22], selects query-relevant frames more effectively. However, this focus on relevance often produces visually redundant selections and weakens the narrative context needed for coherent video understanding. Taken together, these methods highlight a shared limitation: existing approaches are typically governed by a *single* sampling principle—either emphasizing global semantic (uniform sampling) or prioritizing key evidence (relevance-based sampling). Such a one-sided strategy is ill-suited to the heterogeneous nature of video queries. Some questions demand holistic understanding of the overall narrative (e.g., "Why does the mother bird bring a fish to the fox?"), whereas others require fine-grained inspection of a few decisive events (e.g., "How did the cat manage to open the door?"). Since a single principle cannot simultaneously satisfy these competing demands, current methods struggle to adapt across query types, motivating the need for a more adaptive sampling approach.

To address this limitation, we propose **Hierarchical Adaptive frame Sampling** (HAS), a hierarchical framework that tailors frame selection to the specific requirements of each query. HAS proceeds in two stages. In **Frame Backbone Construction**, we cast frame sampling as a Determinantal Point Process (DPP). We build the DPP kernel by injecting frame relevance into the inter-frame similarity matrix, and perform greedy MAP inference to select a set of backbone frames that are both highly relevant and non-redundant. This relevance–diversity trade-off encourages the backbone to cover key moments throughout the video, providing a reliable basis for subsequent enrichment. In **Adaptive Contextual Enrichment**, we allocate the remaining budget to complement the backbone with two types of context. **Local Context** samples temporally neighboring frames around each backbone frame to recover fine-grained dynamics and short-term causal cues, whereas **Global Context** uniformly samples along the full timeline to maintain a holistic narrative view.

The adaptivity of HAS comes from how it distributes the remaining budget between Local and Global Context according to the temporal layout of the backbone frames. We believe that the temporal layout of the backbone frames provides a useful cue for identifying the query's evidential needs. When backbone frames form a dense temporal cluster, the queried content is likely concentrated within a short window; HAS therefore emphasizes Local Context to enrich these key moments with nearby evidence. Conversely, when backbone frames are sparse and widely spread, the queried subject is more likely to recur across the video and require long-range temporal reasoning; HAS thus assigns more budget to Global Context to connect distant events and maintain narrative coherence. In this way, HAS can flexibly shift between zooming in on decisive moments and constructing a holistic overview, addressing the rigidity of one-sided sampling strategies.

To validate HAS, we integrate it into three representative MLLMs: Qwen2.5-VL [4], InternVL3 [41], and LLaVA-Video [38], and evaluate on three widely-used benchmarks: Video-MME [11], LongVideoBench [31], and MLVU [39]. Together, these benchmarks span diverse video lengths and query types, enabling a rigorous evaluation of effectiveness and generalization. Experimental results show that HAS not only significantly outperforms the default uniform sampling strategy used by these models but also surpasses other advanced sampling methods, such as AKS [26] and TSTAR [33].

2 Related Work

2.1 Multimodal Large Language models

The remarkable success of Large Language Models (LLMs) has motivated extending them to incorporate visual information, giving rise to Multimodal LLMs (MLLMs). Early models relied on heavyweight vision–language alignment designs. For instance, BLIP-2 [14] and Qwen-VL [3] adopt a Q-Former to bridge visual embeddings into the language embedding space, while Flamingo [1] injects cross-attention layers into the LLM to fuse visual and textual representations. LLaVA [18] later shows that effective vision–language alignment can be achieved with a much simpler design, using only a lightweight MLP projector and avoiding substantial architectural changes. This paradigm has since been widely adopted by other MLLMs, enabling strong visual understanding performance in models such as MiniGPT-4V [40], ShareGPT4V [6], Qwen2.5-VL [27] and InternVL3 [41].

2.2 Video-based MLLMs

Extending MLLMs from images to videos introduces a key challenge: videos contain far more information, and encoding all frames into tokens would quickly exceed the LLM's context window. Most video-based MLLMs therefore build upon image-based architectures and sample a subset of

frames as input, as in Video-ChatGPT [19], Video-LLaMA [36], Deepseek-VL2 [32], and Video-LLaVA [16]. Inevitably, this discards information from unsampled frames, and the issue becomes more severe as video length increases. To mitigate this, methods such as ADACM [20] and HierarQ [2] employ transformer-based designs to process frames sequentially and compress the entire video into a compact representation. While this reduces the information loss induced by sampling, such architectures generally underperform mainstream sampling-based MLLMs.

2.3 Frame sampling for video MLLMs

Given the limited context window of MLLMs, frame sampling has emerged as the dominant solution for adapting long videos to VideoQA tasks, making it a critical factor in model performance. Early influential methods such as SeViLA [34] and MVU [22] focused on query relevance by using MLLMs to assign importance scores to each frame individually. To address the efficiency bottleneck of scoring each frame individually, Hu et al. [12] predict all frame-level scores in a single forward pass. Other concurrent approaches deviate from explicit query-relevance scoring at inference: Koala [25] employs fixed sparse sampling to condition subsequent visual processing, while ViLA [28] trains an end-to-end module to learn a sampling strategy. More recent methods refine the selection criteria: AKS [26] simultaneously considers both the relevance of a frame and the temporal coverage of the already selected frames, and TSTAR [33] converts temporal search into spatial search by arranging frames into a large grid image. However, existing methods often suffer from a key limitation: they are typically governed by a single sampling principle (e.g., uniform coverage or relevance-only selection) and thus struggle to adapt to the heterogeneous evidential requirements of different query types. To the best of our knowledge, we are the first to explicitly highlight this query-dependent demand and propose **Hierarchical Adaptive frame Sampling** (HAS) to address this challenge.

3 Method

3.1 Preliminaries: Problem Formulation

To motivate our method, we first formalize the video question answering pipeline. A video $V = \{v_1, \ldots, v_T\}$ is a sequence of T frames. Given the limited context window of an MLLM $\mathcal{M}$, it is typically infeasible to encode all T frames; instead, the model operates on a subset of K frames with $K \ll T$. We denote the sampling strategy as $\mathcal{S}$, which takes the video V and optionally the query Q as input and returns a set of K frame indices, $\mathcal{I} = \mathcal{S}(V, Q)$. The sampled frames $\{v_i\}_{i\in\mathcal{I}}$, together with Q, are then fed into $\mathcal{M}$ to generate an answer A. Consequently, the answer quality depends critically on the effectiveness of $\mathcal{S}$. Ideally, an optimal strategy $\mathcal{S}^*$ would select indices $\mathcal{I}^*$ that maximize the probability of generating the ground-truth answer A_{gt}:

$$\mathcal{I}^* = \arg \max_{\mathcal{I}\subset\{1..T\},|\mathcal{I}|=K} P(A = A_{\text{gt}} \mid \{v_i\}_{i\in\mathcal{I}}, Q; \mathcal{M}). \tag{1}$$

However, learning such a strategy is intractable, as even finding $\mathcal{I}^*$ for a single instance would require exhaustively evaluating all possible subsets with the MLLM. The goal of this work is therefore to design a practical and effective strategy, $\mathcal{S}_{\text{ours}}$, that serves as a proxy to $\mathcal{S}^*$.

3.2 Hierarchical Adaptive frame Sampling

Motivated by the need to tailor evidence to different VideoQA queries, we introduce HAS, a two-stage frame sampling framework. In **Backbone Frame Construction**, HAS selects query-relevant yet diverse frames to form an informative backbone. In **Adaptive Contextual Enrichment**, it adaptively allocates the remaining budget between Local and Global Context to complement the backbone based on the needs of the query. The following sections detail each stage.

3.2.1 Backbone Frame Construction

HAS begins by selecting a set of backbone frames that are simultaneously highly relevant to the query and internally diverse. We model this as a subset selection problem using a Determinantal Point Process, a natural choice for balancing quality and diversity. The effectiveness of DPP has been well established in tasks like recommendation systems [5, 30] and, more recently, token pruning [37].

In our formulation, we construct a kernel matrix L that jointly incorporates frame-query relevance (quality) and inter-frame redundancy (diversity) to guide backbone frame selection. Specifically,

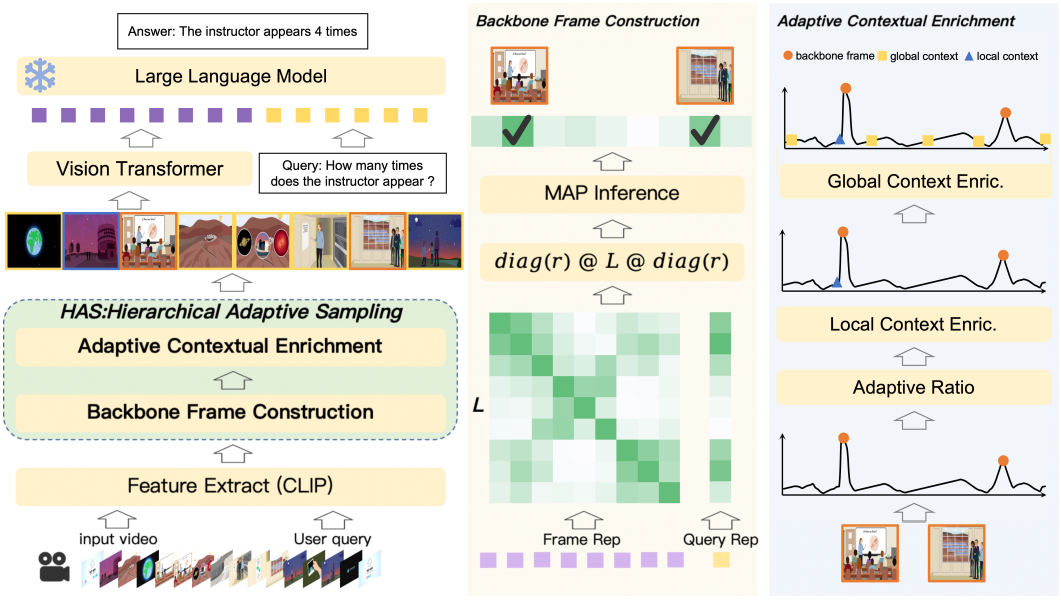


Figure 1: The overall flowchart of HAS. HAS is a training-free approach that enhances model performance solely by modifying the frames sampled from the video. In the Adaptive Contextual Enrichment module, different colors denote different frame types: orange for Backbone Frames, green for Global Context, and blue for Local Context.

we use a pre-trained CLIP model to extract embeddings for each video frame $\mathbf{f}_i$ and the query, and compute a relevance score r_i via cosine similarity. The kernel entries are then defined as:

$$L_{ij} = r_i \cdot \text{sim}(\mathbf{f}_i, \mathbf{f}_j) \cdot r_j, \tag{2}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity between frame embeddings.

Given L, DPP assigns each candidate subset $\mathcal{I}$ a probability proportional to the determinant of its corresponding submatrix, $\det(L_{\mathcal{I}})$, which simultaneously reflects the collective relevance of the selected frames and their diversity. Selecting the backbone frames can thus be formulated as a maximum a posteriori (MAP) inference problem under the DPP model, i.e., finding the subset of size k_b that maximizes:

$$\mathcal{I}_{\text{backbone}} = \underset{\mathcal{I} \subseteq \{1..T\}, |\mathcal{I}| = k_b}{\arg\max} \det(L_{\mathcal{I}}), \tag{3}$$

where k_b denotes the number of backbone frames to be selected.

Since this optimization is NP-hard, we employ an efficient greedy MAP inference algorithm from Chen et al. [5]. This algorithm iteratively selects the sample that provides the largest marginal gain to the log-determinant until the subset size is reached. To prevent a fixed budget from forcing the selection of low-relevance frames, we introduce an adaptive stopping criterion: during this process, we terminate when the marginal gains of adding any remaining candidate frame all fall below a predefined threshold η. We denote the resulting backbone frame indices as $\mathcal{I}_{\text{backbone}}$, which are then used in the subsequent stage.

3.2.2 Adaptive Contextual Enrichment

The backbone frames provide a foundation basis that captures the key moments across the entire video, but remains incomplete on its own due to limited surrounding context. To enrich this foundation for more complex reasoning, we utilize the remaining frame budget, $k_{\text{context}} = K - |\mathcal{I}_{\text{backbone}}|$, to sample additional context frames. This budget is adaptively allocated between two types: Global Context and Local Context.

Local Context. While the backbone frames pinpoint critical moments, the DPP-enforced diversity often yields temporally sparse snapshots that lack the immediate context needed to recover local dynamics and short-term causal cues. To remedy this, we apply a ripple expansion process to sample

Local Context frames, denoted as $\mathcal{I}_{\text{LC}}$. We first sort the backbone frames by relevance (descending) and push them into a FIFO queue, treating each queued frame as an expansion point. We then repeatedly pop an expansion point and scan neighboring frames outward along the temporal axis in both directions . The search stops upon encountering either an already selected frame or an unselected frame whose similarity to the expansion point falls below τ. All newly qualified frames are added to $\mathcal{I}_{\text{LC}}$ and also enqueued as new expansion points, allowing the expansion to propagate to nearby regions. We stop this stage once $|\mathcal{I}_{\text{LC}}|$ reaches the Local Context budget k_{LC}. Overall, this procedure augments the backbone with temporally adjacent evidence, capturing local dynamics and short-range causal relations around key moments.

Global Context. Some queries require a coherent understanding of how the video unfolds over time, which backbone frames alone may not fully capture. To address this, HAS allocates part of the frame budget to *Global Context* and samples an additional set of frames $\mathcal{I}_{\text{GC}}$ to provide broad coverage of the entire video. We explored multiple Global Context sampling strategies and found that simple uniform sampling performs best. Accordingly, we adopt uniform sampling as the default Global Context strategy in this work, and provide detailed comparisons and analysis in Section 4.3.3.

Adaptive Budget Allocation. A fixed allocation between Global and Local Context is suboptimal, as the appropriate balance is inherently query-dependent. We leverage the temporal layout of the backbone frames as a cue for this allocation. Intuitively, when backbone frames form a dense temporal cluster, the query is likely centered on a short, localized segment, and thus benefits more from Local Context to recover fine-grained dynamics and short-term causal cues. In contrast, when backbone frames are sparse and widely spread, the queried content is more likely to recur throughout the video, calling for broader Global Context to connect distant events and maintain narrative coherence.

To quantify this temporal layout, we compute an entropy-based globalness score $\mathcal{S}_{\text{glob}}$. We partition the video timeline into B equal temporal bins, where we set $B = |\mathcal{I}_{\text{backbone}}|$. Let s_t denote the relevance score of a backbone frame $t \in \mathcal{I}_{\text{backbone}}$. For each bin b, we accumulate the weighted mass

$$a_b = \sum_{t \in \mathcal{I}_{\text{backbone}}} s_t \, \mathbf{1}(t \in b)\,, \tag{4}$$

where $\mathbf{1}(\cdot)$ equals 1 if frame t falls into bin b and 0 otherwise. We normalize $\{a_b\}$ into a discrete distribution $\{p_b\}_{b=1}^{B}$ by

$$p_b = \frac{a_b}{\sum_{j=1}^{B} a_j}, \tag{5}$$

and compute the normalized entropy

$$\mathcal{S}_{\text{glob}} = \frac{-\sum_{b=1}^{B} p_b \log\left(p_b + \epsilon\right)}{\log B}. \tag{6}$$

Finally, we allocate the context budget k_{context} by assigning $\lfloor \mathcal{S}_{\text{glob}} \cdot k_{\text{context}} \rfloor$ frames to Global Context and using the remaining frames for Local Context. The final set of frame indices passed to the MLLM is $\mathcal{I}_{\text{final}} = \mathcal{I}_{\text{backbone}} \cup \mathcal{I}_{\text{GC}} \cup \mathcal{I}_{\text{LC}}$.

4 Experiments

4.1 Experimental Setup

Datasets and Models. To empirically validate the effectiveness of HAS, we conduct comprehensive experiments on three widely used video understanding benchmarks: VideoMME [11], LongVideoBench [31], and MLVU [39]. These benchmarks are characterized by their diverse range of question types and video lengths, with some videos extending beyond one hour. This makes them a rigorous testbed for evaluating the efficacy of frame sampling strategies. For models, we select three powerful MLLMs: Qwen2.5-VL-7B [4], InternVL3-8B [41], and LLaVA-Video-7B [38].

Compared Methods. We compare our proposed method, HAS, against four frame sampling methods. These are: (1) UNI, which samples frames at fixed temporal intervals across the video's timeline and serves as the default method for the models evaluated in this paper; (2) TOP, which selects frames in descending order of their query-relevance scores; (3) AKS [26], an enhancement of TOP that

Table 1: Main results on long video understanding benchmarks. The LVB-60s subset is excluded from our evaluation since our approach re-samples from an initial 1fps stream. **Bold** indicates the best performance and underline indicates the second-best within each group.

Model	Size	Frames	LVB		VideoMME			MLVU		
			600	3600	Short	Med.	Long	Hol.	S-D	M-D
Proprietary										
GPT4o	–	–	66.7	61.6	77.7	66.2	59.8	79.4	53.7	35.7
Gemini-1.5-Flash	–	–	63.1	57.3	71.3	58.6	52.1	73.9	52.5	27.9
Open-source										
VideoChat2	7B	16	39.0	37.5	54.2	42.8	39.9	71.7	50.0	24.9
Long-VITA	14B	64	52.7	46.8	74.6	61.2	55.0	81.0	69.8	50.1
LLAVA-OV	72B	64	61.6	56.5	77.2	63.2	58.2	83.1	74.9	48.6
LLAVA-Video	72B	64	63.9	59.3	80.8	68.6	62.1	81.2	78.7	55.7
Qwen2.5VL	7B	64	59.0	51.1	73.9	62.3	52.2	80.6	66.0	43.9
+ TOP	7B	64	62.9	56.0	74.2	62.3	53.8	76.5	72.2	54.4
+ TStar	7B	64	62.6	55.8	74.2	63.4	50.3	78.8	72.4	55.5
+ AKS	7B	64	62.1	56.2	74.4	65.1	54.9	**82.3**	71.4	51.0
+ HAS	7B	64	**63.8**	**56.4**	**74.8**	**65.3**	**55.4**	80.8	**73.4**	**57.2**
LLAVA-Video	7B	64	59.2	50.9	75.9	63.2	52.1	79.7	75.0	53.7
+ TOP	7B	64	62.1	53.7	76.9	59.8	51.9	75.6	74.4	57.7
+ TStar	7B	64	61.9	53.9	76.6	63.7	51.4	76.7	75.1	56.8
+ AKS	7B	64	61.4	54.3	77.0	**65.1**	**55.0**	79.3	74.9	52.7
+ HAS	7B	64	**63.1**	**56.2**	**77.6**	64.3	**55.0**	**80.6**	**76.1**	**60.6**
InternVL3	8B	64	61.7	53.0	76.2	66.4	55.4	**84.7**	73.9	52.5
+ TOP	8B	64	63.6	56.2	77.8	63.6	54.6	80.8	75.4	57.0
+ TStar	8B	64	64.1	55.7	77.6	66.9	54.1	81.7	76.3	57.5
+ AKS	8B	64	63.3	54.8	77.2	68.9	55.3	82.3	75.0	52.9
+ HAS	8B	64	**64.6**	**56.4**	**78.1**	**70.0**	**56.3**	82.3	**77.6**	**58.1**

simultaneously considers both the relevance of a frame and the temporal coverage of the already selected set of frames and (4) TSTAR [33], which select question-relevant frames via object-centric visual search rather than semantic similarity, iteratively narrowing down the temporal region of interest through coarse-to-fine resampling.

Implementation Details. We conduct all our experiments using the VLMEvalKit framework [9]. HAS is implemented in a plug-and-play manner: we only change which frames are selected as input while keeping the model weights frozen throughout all evaluations. To improve computational efficiency, we first pre-sample a candidate pool from each video at 1 FPS, and then apply our method to select the final set of frames from this pool. For the DPP backbone selection, we construct the candidate set by taking the top-$6K$ frames ranked by relevance, where K is the frame budget. All relevance scores are computed using CLIP-B [21]. Unless otherwise specified, we set $\eta = 0.03$.

4.2 Main results

4.2.1 Quantitative results

As presented in Table 1, our proposed HAS is comprehensively evaluated across three models and three benchmarks. In the great majority of evaluations, HAS outperforms the other sampling strategies, achieving the best performance.

The results on the LVB and VideoMME benchmarks particularly highlight our method's advantage in long-duration video understanding. For instance, on the challenging LVB-3600 benchmark, HAS boosts the LLAVA-Video-7B model's score to 56.2, a clear +1.9 improvement over the next-best method, AKS. Similarly, on the VideoMME Long benchmark, HAS (on InternVL-3) achieves a score of 56.3, significantly outperforming AKS (55.3). Moreover, this strong performance across both LVB, which emphasizes single-scene analysis, and VideoMME, which requires a holistic understanding of the video timeline to answer, showcases our method's ability to adapt to diverse problem types.

The results on the MLVU benchmark further showcase our method's adaptability. HAS achieves top scores on detail-oriented tasks like Single-Detail (S-D) and Multi-Detail (M-D) (e.g., 77.6 and 58.1 with InternVL3-8B), while also securing highly competitive results on Holistic (Hol.) tasks, including

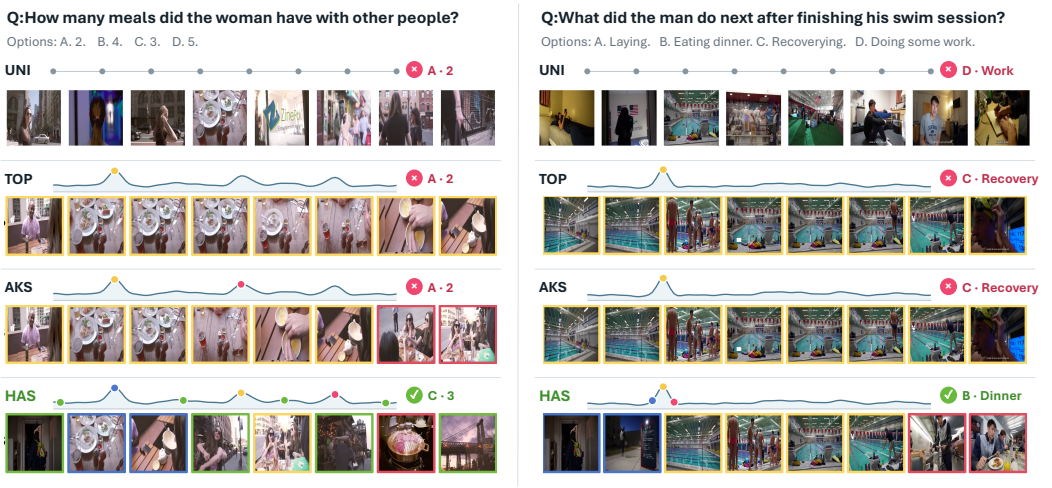


Figure 2: Two examples of how different sampling strategies impact video understanding. The curve illustrates the video's temporal relevance profile. A frame's border color matches the color of its sampling dot on the curve.

scores of 80.6 with LLAVA-Video and 82.3 with InternVL3, the best and second-best results among all sampling strategies. This demonstrates HAS's dual capability: providing a global perspective for holistic queries, while also capturing the local dynamics and short-term context for questions focused on specific moments.

Notably, HAS enables smaller open-source models to achieve performance competitive with, and in some cases even superior to, much larger or proprietary counterparts. For example, HAS-enhanced LLaVA-Video-7B achieves 60.6 on MLVU M-D, outperforming its 10× larger 72B counterpart (55.7), demonstrating immense gains at a fraction of the computational cost. In another compelling case, InternVL3-8B with HAS (70.0) outperforms the proprietary GPT-4o (66.2) on VideoMME Med. These results highlight a key insight: in certain scenarios, employing our economical HAS as a video sampling method is a more effective strategy for performance enhancement than the costly alternative of training ever-larger models.

4.2.2 Qualitative results

Figure 2 illustrates the behavior of different sampling strategies on two representative cases. In the first case, HAS, benefiting from DPP, accurately captures three distinct question-relevant scenes that are all necessary for answering the question, whereas AKS and TOP ignore some of these key scenes to varying degrees. In addition, the global context provide a more complete view of how the video evolves over time, so that the final sampled frame set offers a coherent and sufficient basis for answering the question. In the second case, HAS enriches the selected query-relevant backbone frames with local context which supplies the crucial short term context required to answer the question correctly. In comparison, AKS and TOP rely solely on key frames, which results in incomplete contextual understanding and incorrect predictions. Jointly, these two cases collectively demonstrate the effectiveness of HAS. We provide additional qualitative examples in the appendix, covering diverse scenarios under different frame budgets.

4.3 Ablation study

4.3.1 Ablation Study On Frame Budgets

Table 2 presents a performance comparison under varying input frame budgets (64, 32, and 16). The results demonstrate the strong performance and robustness of our method across different frame budgets. Under ample budgets of 64 and 32 frames, HAS consistently achieves the best performance across all benchmarks. For instance, at 32 frames, HAS secures top scores of 66.7 on VideoMME and 61.5 on LVB, clearly surpassing the next-best scores of 66.2 (AKS) and 60.9 (TOP), respectively. When the frame budget is constrained to 16 frames, splitting the budget into three components (DPP

Table 2: Ablation study on frame budget for HAS on InternVL3.

Metric	LVB			VideoMME			MLVU		
Frames	64	32	16	64	32	16	64	32	16
UNI	59.8	56.4	55.1	66.0	64.4	61.9	73.1	70.0	65.8
TOP	61.8	60.9	60.0	65.3	63.5	61.0	74.6	73.0	69.4
AKS	61.1	59.5	60.1	67.1	66.2	65.8	74.3	73.1	65.0
HAS	**62.2**	**61.5**	**60.7**	**68.1**	**66.7**	**66.2**	**75.8**	**74.1**	**71.4**

Table 3: Ablation study of HAS components on InternVL3.

Model	Frames	LVB		VideoMME			MLVU		
		600	3600	Short	Med.	Long	Hol.	S-D	M-D
TOP	64	63.6	56.2	77.8	63.6	54.6	80.8	75.4	57.0
DPP	64	63.8	55.7	78.0	64.8	55.4	82.2	76.3	56.8
DPP+GC	64	64.3	56.0	77.9	68.3	**57.0**	**82.3**	77.1	57.9
DPP+LC	64	61.7	**56.4**	78.0	67.0	55.2	**82.3**	**77.9**	**59.1**
HAS	64	**64.6**	**56.4**	**78.1**	**70.0**	56.3	**82.3**	77.6	58.1

backbone, Local Context, and Global Context) may allocate too few frames to each part, limiting their effectiveness under such extreme constraints. To address this issue, we simplify HAS by retaining only the first-stage DPP backbone construction at 16 frames, avoiding over-fragmentation of the limited budget. With this adaptation, HAS remains the best-performing method under the 16-frame setting across all benchmarks, achieving 66.2 on VideoMME, 60.7 on LVB, and 71.4 on MLVU.

4.3.2 Ablation study on the components of HAS

To validate the contribution of each component in HAS, we conducted a detailed ablation study, with results presented in Table 3. Our ablation study (Table 3) begins by comparing DPP against TOP. DPP improves over TOP on most metrics, with notable gains on benchmarks like VideoMME-Med (64.8 vs. 63.6) and MLVU HoL (82.2 vs. 80.8). This superiority stems from DPP's inherent ability to balance both relevance and diversity. Therefore, we select DPP for our backbone frame sampling.

Building on this, we observe that different context types cater to different benchmark demands. Specifically, adding Global Context (DPP+GC) yields substantial gains on benchmarks emphasizing global semantics, boosting the VideoMME-Med score from 64.8 to 68.3. Conversely, Local Context (DPP+LC) proves highly effective for tasks requiring fine-grained detail, achieving excellent scores on MLVU S-D 77.9 and M-D 59.1. These results show that GC and LC each offer unique, complementary advantages. This directly motivates our adaptive approach.

By successfully synthesizing these complementary strengths, our full method, HAS, achieves the best or second-best performance across all benchmarks. It secures top scores of 64.6 on LVB-600 and 70.0 on VideoMME-Med, and remains highly competitive even where not strictly the top performer, such as on VideoMME-Long (56.3). In a few cases, a single-context variant (e.g., DPP+LC or DPP+GC) can be slightly better, likely because some benchmarks strongly prefer either local or global context. Overall, HAS provides a strong balance between high performance and generalization, avoiding large swings across benchmarks.

4.3.3 Ablation study of Alternative Strategies for Global Context Sampling.

In HAS, a portion of the budget is dedicated to Global Context to provide a global perspective of the video. To identify the most effective GC sampling strategy, we evaluated four approaches: (i)Uniform (UNI), which samples frames at uniform temporal intervals; (ii)Max-in-Segment (MiS), which divides the video into uniform segments and selects the most relevant frame to the query from each; (iii)Clustering-based (CLUS), which performs clustering within each uniform segment and then selects the centroid of the largest cluster; and (iv)Temporal Bridging (TB), which iteratively samples the midpoint of the largest temporal gap. As shown in Table 4, the simple UNI proves highly effective, with leading performance in benchmarks such as LVB and VideoMME. We attribute this to

Table 4: Ablation study of GC strategies.

Method	LVB	VideoMME	MLVU
UNI	**62.2**	**68.1**	75.8
MiS	61.0	67.4	**76.0**
CLUS	61.1	67.9	75.7
TB	61.9	67.4	75.2

Table 5: Ablation study of different FPS.

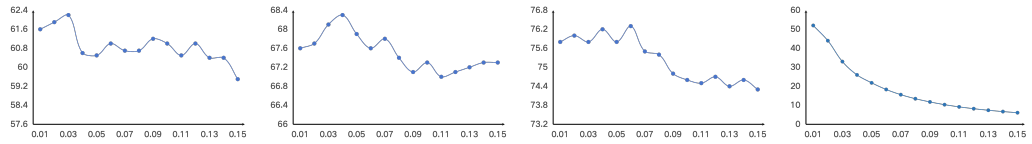
FPS	LVB	VideoMME	MLVU
2	**62.3**	**68.7**	**76.1**
1	62.2	68.1	75.8
0.5	61.3	68.0	75.3
0.25	61.1	67.6	75.4

a mismatch: Global Context aims to offer a broad, unbiased temporal overview, whereas methods like MiS and CLUS introduce relevance or centrality biases. These biases conflict with this goal and compromise the uniform coverage vital for a holistic view. Therefore, we conclude that a simple unbiased strategy like Uniform sampling is the most robust choice for supplementing a query-focused backbone.

4.3.4 Ablation study of different fps

Table 5 reports the performance of HAS under different pre-sampling FPS. Using a higher FPS generally yields better accuracy, as a denser candidate pool provides more fine-grained temporal coverage and is less likely to miss critical moments or key evidence. In our main experiments, we set the pre-sampling rate to 1 FPS to balance performance with efficiency, since computing query-frame relevance scores becomes increasingly expensive as FPS grows.

4.3.5 Ablation study of hyperparameters



(a) Perf. on LVB (b) Perf. on VMME (c) Perf. on MLVU (d) $|\mathcal{I}_{\text{backbone}}|$ on VMME

Figure 3: Analysis of the hyperparameter η. (a)–(c) show the model performance under different values of η on three benchmarks, and (d) shows the number of selected backbone frames under different values of η.

In this section, we analyze η, the marginal-gain threshold in our DPP-based backbone selection (smaller η selects more frames). As shown in Figure 3, the performance exhibits a consistent trend across benchmarks, first increasing and then decreasing. With a low η, the backbone may include low-relevance frames and leave too little budget for the second-stage enrichment. With a high η, too few backbone frames are selected, causing the model to miss some key moments. As a result, performance is best at a moderate η (around $\eta = 0.03$–0.04), which strikes a good balance between capturing sufficient key moments and reserving enough budget for contextual enrichment. This behavior further supports the necessity of our two-stage design. We therefore set $\eta = 0.03$ in all main experiments. The sensitivity to the similarity threshold τ is further analyzed in Appendix 9.

5 Conclusion

In this work, we find that existing video frame sampling methods are often governed by a one-sided principle, making it hard to simultaneously meet the heterogeneous evidential needs of different queries. To address this limitation, we propose HAS, a hierarchical framework that dynamically tailors sampled frames to the specific needs of each query. HAS first constructs a query-aware backbone via DPP to capture diverse and relevant key moments, and then adaptively allocates the remaining sampling budget between Global Context (to cover the broader video timeline for holistic queries) and Local Context (to capture fine-grained temporal dynamics for moment-centric queries). With this adaptive design, HAS consistently delivers strong performance across multiple benchmarks and a range of MLLMs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62276015), the GW2025-09 project, the National Natural Science Foundation of China (Grant No. 62376026), and the Beijing Nova Program (Grant No. 20230484296).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.

[2] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8545–8556, 2025.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in neural information processing systems*, 31, 2018.

[6] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[8] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL `https://arxiv.org/abs/2305.06500`.

[9] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024.

[10] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

[11] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[12] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13702–13712, 2025.

[13] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[15] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[16] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[19] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[20] Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. Adacmˆ 2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8534–8544, 2025.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[22] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. Understanding long videos in one multimodal language model pass. *arXiv preprint arXiv:2403.16998*, 3(4):12, 2024.

[23] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

[24] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169, 2025.

[25] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13581–13591, 2024.

[26] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29118–29128, 2025.

[27] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[28] Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming C Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering. In *European Conference on Computer Vision*, pages 186–204. Springer, 2024.

[29] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025.

[30] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2165–2173, 2018.

[31] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

[32] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[33] Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8579–8591, 2025.

[34] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

[35] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[36] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[37] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025.

[38] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[39] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pages arXiv–2406, 2024.

[40] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[41] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

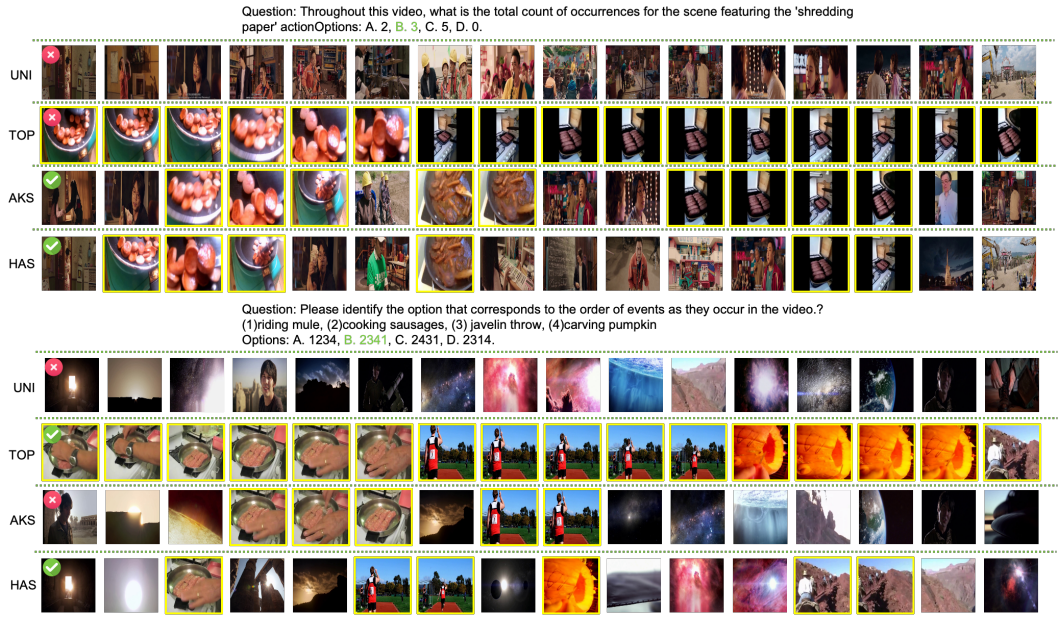


Figure 4: Two qualitative examples on questions from MLVU, comparing four frame sampling strategies: UNI, TOP, AKS, and HAS. For each example, from top to bottom, each block shows the frames sampled by UNI, TOP, AKS, and HAS, respectively.Yellow outlines highlight frames that are directly helpful for answering the question.

A Additional Qualitative Visualization.

In this section, we provide a qualitative comparison of UNI, TOP, AKS, and HAS. As the figure shows, UNI often misses critical evidence frames due to its rigid sampling strategy, which forces the model to reason based on incomplete visual information. TOP, while selecting frames that are mostly directly helpful for answering the question, introduces substantial redundancy and still overlooks some frames that are only weakly related with the question but provide complementary evidence for answering it. AKS generally captures a richer set of informative moments than UNI and TOP, but its sampled frames are still less well aligned with the question than those selected by HAS. In contrast, HAS effectively discovers non-redundant, question-relevant frames distributed and enriches them with both local and global temporal context, thereby providing the model with a more complete and informative visual basis for reliable answer prediction.

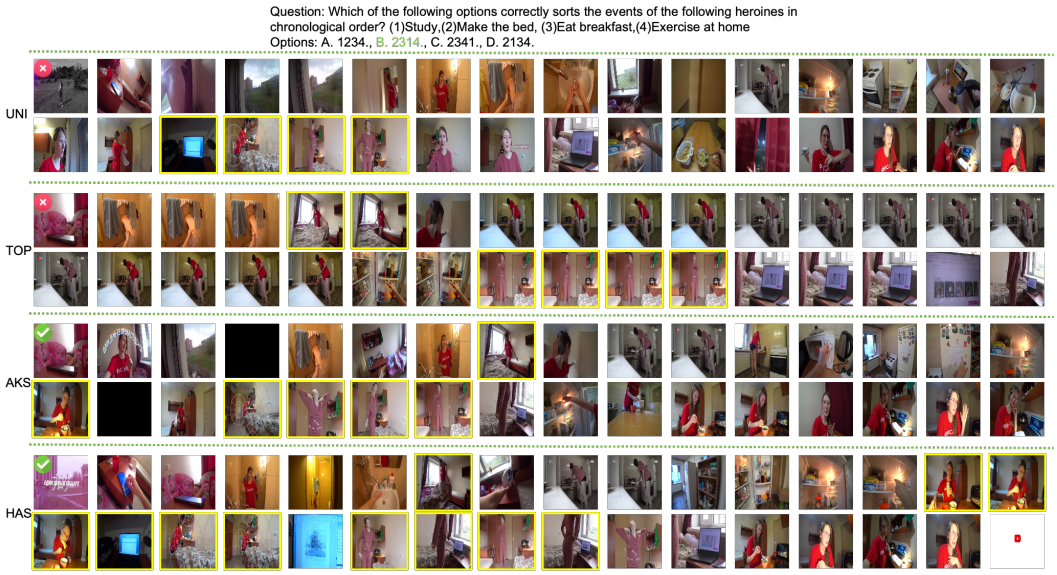


Figure 5: A qualitative result question from VideoMME, comparing four frame sampling strategies: UNI, TOP, AKS, and HAS. For each example, from top to bottom, each block shows the frames sampled by UNI, TOP, AKS, and HAS, respectively.Yellow outlines highlight frames that are directly helpful for answering the question.

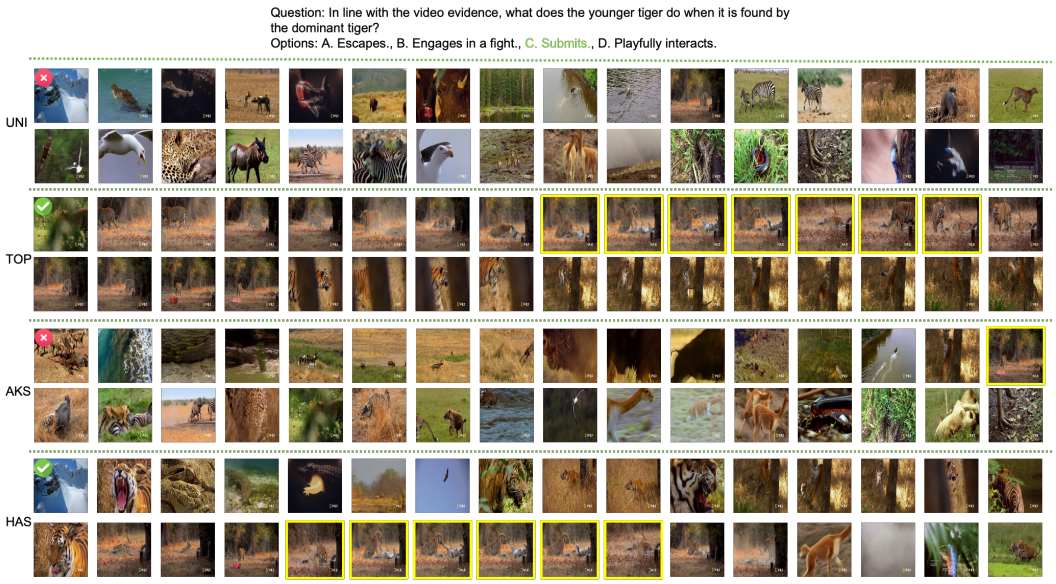


Figure 6: A qualitative result question from VideoMME, comparing four frame sampling strategies: UNI, TOP, AKS, and HAS. For each example, from top to bottom, each block shows the frames sampled by UNI, TOP, AKS, and HAS, respectively.Yellow outlines highlight frames that are directly helpful for answering the question.

B Ablation study of different visual encoders

To quantify the impact of the underlying vision encoder on HAS, we conduct an ablation study comparing CLIP and SigLIP2 on VideoMME and LongVideoBench. As shown in Table 6, although SigLIP2 is substantially stronger than CLIP at the image–text level, the resulting gains on these VideoQA benchmarks are modest. We conjecture that this is because HAS performance is not determined by keyframe quality alone: the final evidence set is jointly composed of backbone, Global

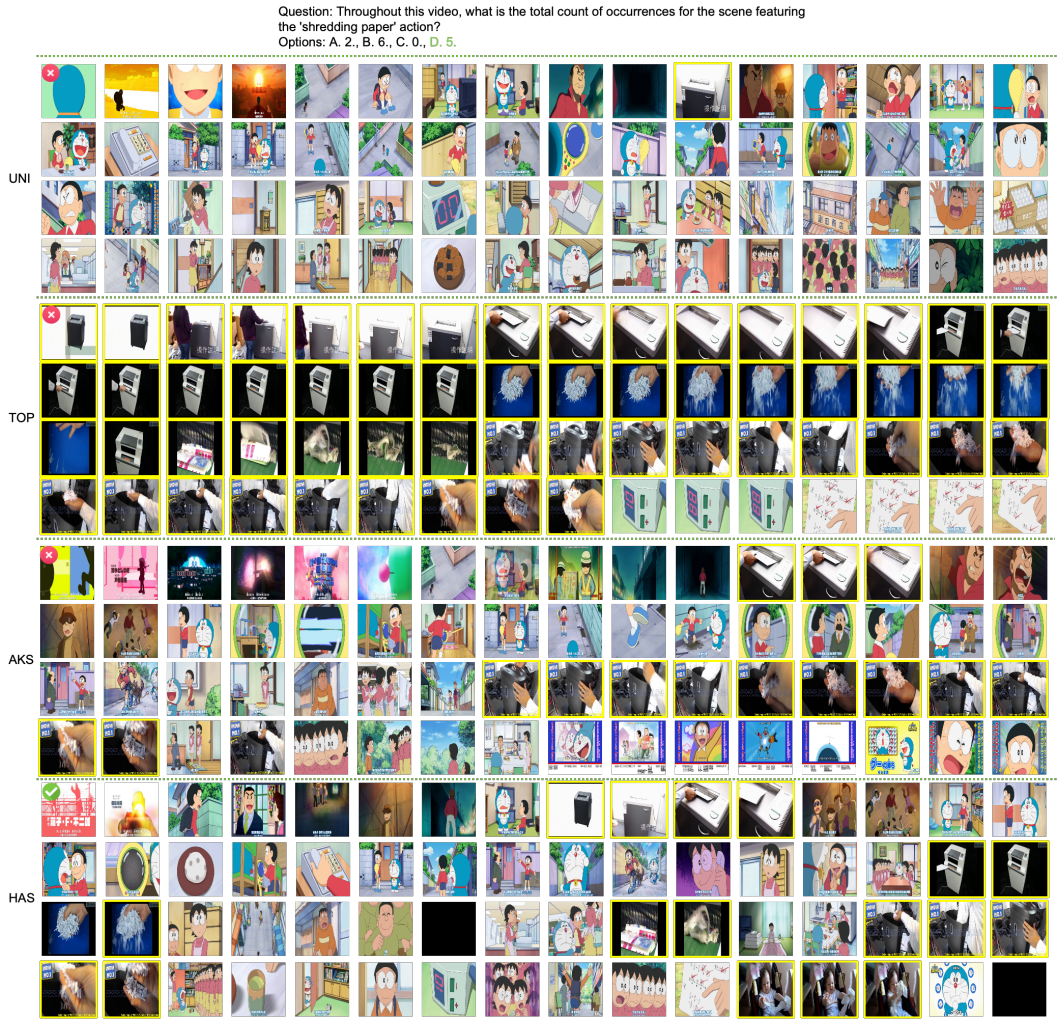


Figure 7: A qualitative result question from MLVU, comparing four frame sampling strategies: UNI, TOP, AKS, and HAS. For each example, from top to bottom, each block shows the frames sampled by UNI, TOP, AKS, and HAS, respectively.Yellow outlines highlight frames that are directly helpful for answering the question.

Table 6: Ablation study of different vision encoders.

Vision encoder	VideoMME	LongVideoBench
CLIP	68.1	62.2
SigLIP2	68.3	62.5
Δ (SigLIP2 $-$ CLIP)	+0.2	+0.3

Context, and Local Context frames. In addition, a portion of questions on long-video benchmarks may hinge on signals beyond static vision (e.g., audio cues) or be limited by the VLM's inherent capacity, which further reduces the marginal benefit of a stronger vision encoder.

C Scaling to Larger Backbones

To further examine whether HAS scales to larger vision–language models, we additionally evaluate HAS with InternVL3-38B and with GPT5 on the VideoMME benchmark. For InternVL3-38B, we follow the same frame budget and evaluation protocol as in the 7B setting and report results on the

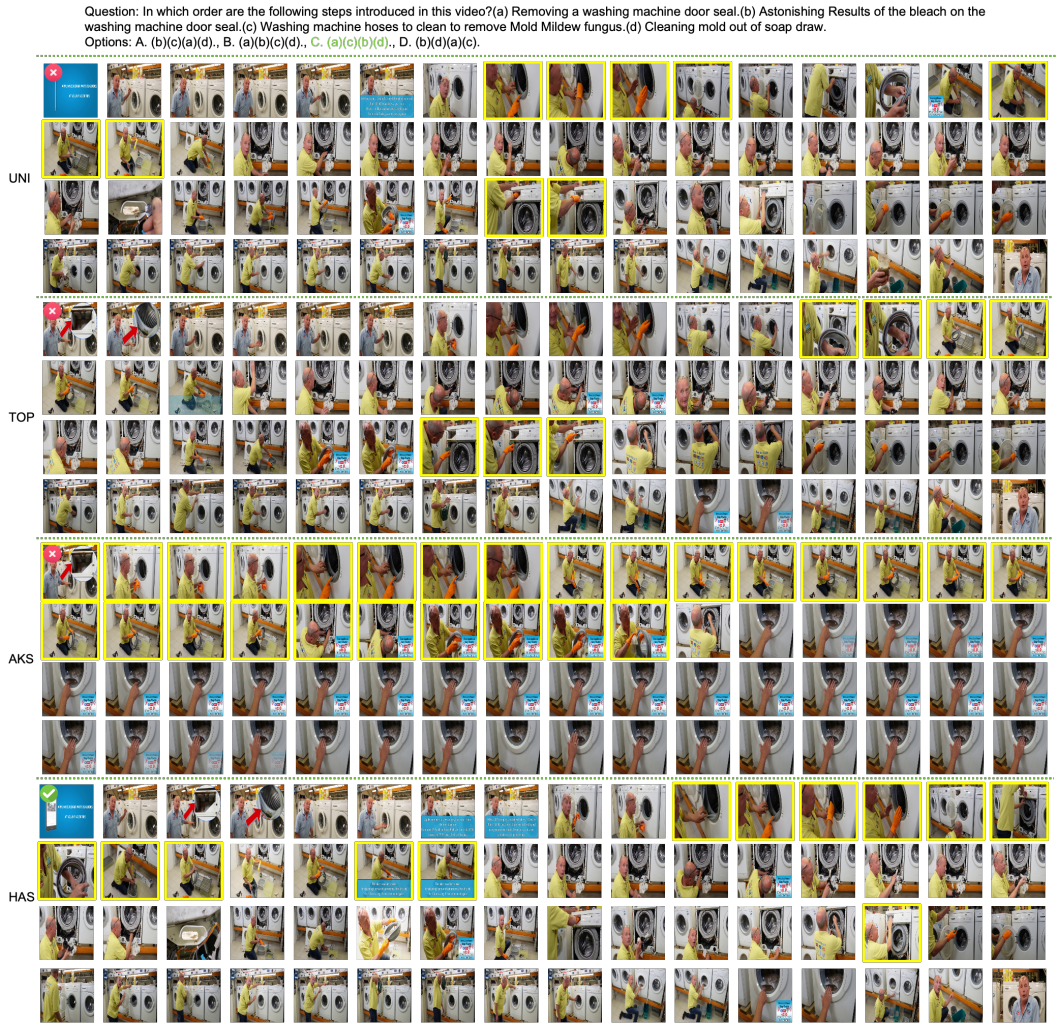


Figure 8: A qualitative result question from VideoMME, comparing four frame sampling strategies: UNI, TOP, AKS, and HAS. For each example, from top to bottom, each block shows the frames sampled by UNI, TOP, AKS, and HAS, respectively.Yellow outlines highlight frames that are directly helpful for answering the question.

Table 7: HAS on larger backbones on VideoMME.

Backbone	Sampling	Videos	Medium	Long
InternVL3-38B	UNI	Full	72.1	62.1
InternVL3-38B	HAS	Full	73.8	64.0
GPT5	UNI	128-subset	77.3	75.9
GPT5	HAS	128-subset	81.3	78.1

medium and long splits. For GPT5, we evaluate only 128-video subsets of the VideoMME-medium and VideoMME-long splits, obtained by random sampling 128 videos from each split. As shown in Table 7, HAS consistently outperforms uniform sampling across both backbones, improving InternVL3-38B from 72.1 to 73.8 on the medium split and from 62.1 to 64.0 on the long split, and improving GPT5 from 77.3 to 81.3 on the medium (subset) split and from 75.9 to 78.1 on the long (subset) split.

D Runtime of Key Frame Selection

Table 8: Average runtime for key frame selection on VideoMME. For TOP/AKS/HAS, we report the decomposition as relevance score computation time + selection time.

Method	TSTAR	VideoTree	TOP	AKS	HAS
Time	101.24 s	12.64 s	5.46 s + 1 ms	5.46 s + 2 ms	5.46 s + 43 ms

We measured the average time required by different methods to perform key frame selection on the VideoMME benchmark. As summarized in Table 8, HAS incurs a runtime comparable to TOP and AKS, while being substantially faster than TSTAR and VideoTree [29]. Notably, since VideoTree evaluates an LLM's capability of understanding videos, it is not directly comparable to our plug-and-play frame selection in terms of task performance; therefore, we only include VideoTree in the runtime comparison. The similar runtime of TOP, AKS, and HAS is expected, as most of the computation is spent in CLIP-based feature extraction for frame–query relevance scoring; the three methods differ only in the lightweight selection rules applied afterward. In our implementation, the total time for HAS is $5.46\,\text{s} + 43\,\text{ms}$, where $5.46\,\text{s}$ is spent on relevance computation and only $43\,\text{ms}$ on selection. Since relevance scoring dominates the overall cost, HAS introduces negligible additional overhead, keeping the total runtime close to that of TOP and AKS.

E Sensitivity of τ

Table 9: Performance of InternVL3-8B with HAS on VideoMME under different τ.

τ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Acc.	68.0	68.2	68.1	67.7	67.9	67.5	67.1	66.5	66.7

We study the effect of the similarity threshold τ used in Adaptive Contextual Enrichment on VideoMME, while keeping all other settings identical to our default configuration. As shown in Table 9, HAS is robust to τ when it is relatively high: for $\tau \in [0.5, 0.9]$, performance remains stable with only minor fluctuations. In contrast, when τ is reduced below 0.5, performance tends to degrade as τ decreases. We attribute this to an overly permissive stopping criterion: with a smaller τ, the expansion does not terminate until it encounters a frame whose similarity to the expansion point is *very* low. As a result, the search tends to propagate farther along the timeline and may include temporally distant frames that are weakly related to the key moment, deviating from our goal of capturing immediate local context and short-term dynamics, thus hurting performance.

F Transfer to short-video benchmarks

Table 10: Performance on MMbench-video. HAS is evaluated with a 4 FPS pre-sampled candidate pool.

Benchmark	UNI	HAS
MMbench-video	1.70	1.96

Although our main evaluation focuses on long-video benchmarks, HAS is not limited to long-form videos and can be directly applied to shorter videos. We primarily benchmark on long-video settings because frame sampling is more critical there: when the video is long, a small frame budget inevitably introduces information loss, and the choice of sampling strategy has a larger impact. In contrast, for short videos, even uniform sampling can often cover most of the relevant content within the limited timeline, making the sampling problem less challenging.

To verify that HAS remains effective beyond long-form benchmarks, we additionally evaluate it on MMbench-video[10], a short-video benchmark. As shown in Table 10, HAS yields a clear improvement over uniform sampling, increasing the score from 1.70 to 1.96. For this experiment, we

pre-sample the candidate pool at 4 FPS to match the shorter temporal scale of MMbench-video. More broadly, this result further indicates that HAS is not tied to a specific FPS; in our main experiments, we use 1 FPS as a practical trade-off between the cost of computing query–frame similarity and overall performance.

Broader Impact

This work aims to improve the performance of video MLLMs through more effective frame sampling under a fixed frame budget. Since HAS only changes the input frame selection strategy and does not introduce new data collection, model training, or deployment scenarios, we do not foresee direct negative societal impacts beyond those already associated with existing video understanding models.

Limitations

The main goal of HAS is to improve the quality of the visual input by selecting more informative frames under a fixed frame budget. However, frame sampling only determines what visual evidence is provided to the model, and does not directly improve the reasoning, perception, or language generation ability of the underlying MLLM. Therefore, the upper bound of our method is still largely constrained by the capability of the base model. If the model fails to understand the selected frames, reason over temporal relations, or generate the correct answer, better frame selection alone may not be sufficient to resolve the error. This also suggests that HAS is complementary to, rather than a replacement for, stronger video MLLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide the limitations in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results such as theorems or lemmas that require formal assumptions and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the main information needed to reproduce the experimental results, including the evaluated backbones and benchmarks, key hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the core code for the proposed frame sampling method, including the main HAS procedure.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides the key experimental settings, including backbones, benchmarks, evaluation framework, baselines, budgets, and main hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars, confidence intervals, or statistical significance tests for the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mainly use RTX 4090 GPUs for our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide broader impacts in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new data or models that pose a high risk for misuse, such as pretrained language models, image generators, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use existing models, datasets, and evaluation tools only for research purposes. All pretrained model backbones, datasets, baseline methods, and evaluation frameworks are credited through citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets.

    Guidelines: The paper does not release new assets.

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. The experiments are conducted using existing MLLM backbones, established multimodal benchmarks, and previously available training data, so participant instructions, screenshots, and compensation details are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, potential risks to study participants, risk disclosure, and IRB or equivalent approvals are not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: MLLMs are evaluated backbones, not autonomous agents used to generate data or labels.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.