# From Label Priors to Task Evidence: Long-Video Frame Selection via Bayesian GRPO

**Shuochen Chang**[1*] **Bingjie Gao**[1*] **Qingyang Liu**[1] **Qianli Ma**[1] **Yibo Miao**[1]
**Haonan Zhao**[1] **Zhaohe Liao**[1] **Xiaofeng Zhang**[1] **Jiangtong Li**[2†] **Li Niu**[1†]
[1]Shanghai Jiao Tong University [2]Tongji University
[*]Equal contribution. [†]Corresponding authors.
{csc1332741686, whynothaha, ustcnewly}@sjtu.edu.cn
jiangtongli@tongji.edu.cn

## Abstract

Multimodal Large Language Models (MLLMs) face significant scalability challenges in long-video understanding due to quadratic attention complexity and context window constraints. Existing retrieval-based approaches typically rely on supervised learning from static keyframe annotations. However, these label priors often fail to align with the actual information required by downstream MLLMs for complex reasoning, creating a misalignment between the training target and inference needs. To bridge this gap, we propose a probabilistic framework that reformulates frame selection as latent-variable inference. We first initialize a lightweight policy using supervised learning to capture general visual saliency. Then, we treat a frozen MLLM as a task environment and use its answer likelihood as dense task evidence to refine the policy via Bayesian Group Relative Policy Optimization (GRPO). Within this formulation, the negative log-likelihood (NLL) over ground-truth answer tokens provides a principled dense reward, while a prior-matching KL term preserves semantic consistency with the elicited label prior. This process transitions the selection principle from static label priors to task evidence. Extensive experiments on Video-MME, MLVU, and LongVideoBench demonstrate that our selector achieves competitive accuracy with MLLM-based scorers while retaining the inference efficiency of lightweight methods. The selector is trained once with a frozen MLLM environment and transfers directly to multiple downstream MLLMs without retraining. Code is available at `github.com/Miraclecsc/Frame_Selection_GRPO`.

## 1 Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) have revolutionized video understanding [1–3, 8, 73, 11, 70, 19], enabling sophisticated reasoning over interwoven visual-textual data. However, despite their impressive zero-shot capabilities and instructional alignment, MLLMs face a fundamental struggle when processing long-form videos [10, 47, 14, 20]. The quadratic computational complexity inherent to self-attention mechanisms [56], coupled with strict context window limitations, typically prohibits the direct ingestion of lengthy, uncompressed frame sequences. Consequently, extracting relevant spatiotemporal information from temporally extended content remains a critical bottleneck, hindering the deployment of MLLMs in real-world applications such as movie analysis and long-horizon agent navigation.

To mitigate this scalability issue, existing approaches [60, 41, 9, 79, 72, 40, 24] generally fall into two dominant paradigms: architectural modification and external frame retrieval. Architectural approaches, such as streaming video LLMs [7, 28, 42, 27, 25] or dynamic memory compression,

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

attempt to fundamentally modify the MLLM's internal token-processing mechanism to accommodate infinite or highly compressed contexts. While theoretically promising, these methods often incur prohibitive training costs and suffer from catastrophic forgetting, degrading the model's general-purpose capabilities. Alternatively, retrieval-based methods [4, 50, 71, 81] bypass architectural changes by employing an external module to select a concise subset of keyframes. However, this paradigm suffers from a severe efficiency-accuracy trade-off. Lightweight selectors (*e.g.*, CLIP-based similarity matchers) [54, 36, 32, 52, 30, 66, 23] rely on shallow semantic alignment and lack the deep causal reasoning required for complex, multi-hop temporal queries. Conversely, utilizing a powerful MLLM as a dense frame scorer [83, 65, 64, 61] bridges the reasoning gap but introduces unacceptable computational overhead and latency during inference.

In this work, we propose a lightweight, plug-and-play selection module that achieves high reasoning accuracy without imposing heavy inference latency. We identify a fundamental theoretical misalignment in current retrieval-based methods: *a static label prior is not strictly identical to task evidence*. Standard supervised selectors trained on fixed human annotations learn the former (typically visually salient or summary frames), but downstream MLLMs require the latter (specific visual evidence necessary to complete a logical reasoning chain). To bridge this semantic gap, we depart from a ranking paradigm and reformulate frame selection as a probabilistic latent-variable modeling problem.

Specifically, we introduce a rigorous two-stage probabilistic framework that transitions from human-derived label priors to MLLM-driven task evidence. First, we construct a lightweight selector that learns a *Label Prior* from keyframe annotations via pairwise ranking. This initializes the policy within a semantically plausible region, drastically reducing the combinatorial search space. Second, we perform *Posterior Inference* to dynamically align this prior with the target MLLM. By treating the frozen MLLM as the environment and maximizing the Evidence Lower Bound (ELBO), we naturally utilize the MLLM's Negative Log-Likelihood (NLL) as dense task evidence, while simultaneously leveraging the elicited prior for principled KL-regularization. We solve this optimization using a novel Bayesian Group Relative Policy Optimization (GRPO) algorithm, which effectively refines the policy into a task-adaptive posterior that autonomously identifies reasoning-critical frames without reward hacking.

Our main contributions are summarized as follows:

- We propose a plug-and-play, lightweight frame selector that decouples selection from inference, enabling highly efficient long-video understanding without modifying the target MLLM's architecture.
- We introduce a probabilistic framework that reformulates frame selection as Bayesian inference, theoretically justifying the sequential transition from supervised Prior Elicitation to MLLM-driven Posterior Inference.
- We derive a novel Bayesian GRPO algorithm. By leveraging the target MLLM's answer likelihood as dense task evidence and the label prior as KL-regularization, our method robustly refines the selector while mitigating reward hacking.
- We empirically demonstrate that our selector acts as a universal plugin across multiple downstream MLLMs, achieving a superior accuracy-efficiency trade-off compared to both training-free lightweight methods and computationally heavy MLLM-based scorers.

## 2 Related Works

### 2.1 MLLM for Long Video Understanding

Recent advancements in Multimodal Large Language Models (MLLMs) have significantly improved video understanding [6, 48, 57, 35, 78, 31], yet processing long-form videos remains a computational bottleneck. Existing approaches largely focus on architectural modifications to handle extended context windows. For instance, StreamingVLM [60] introduces a streaming-aware mechanism to enable infinite-length video processing without retraining, while Video-XL [51, 34, 67, 43] utilizes latent summarization tokens to compress hour-scale visual inputs efficiently. Other works like TimeSuite [69, 58] propose grounded tuning strategies to adapt short-form MLLMs for long contexts. These architecture-centric methods often require substantial computational resources or sacrifice fine-grained temporal reasoning capabilities, highlighting the need for more efficient, selection modules.

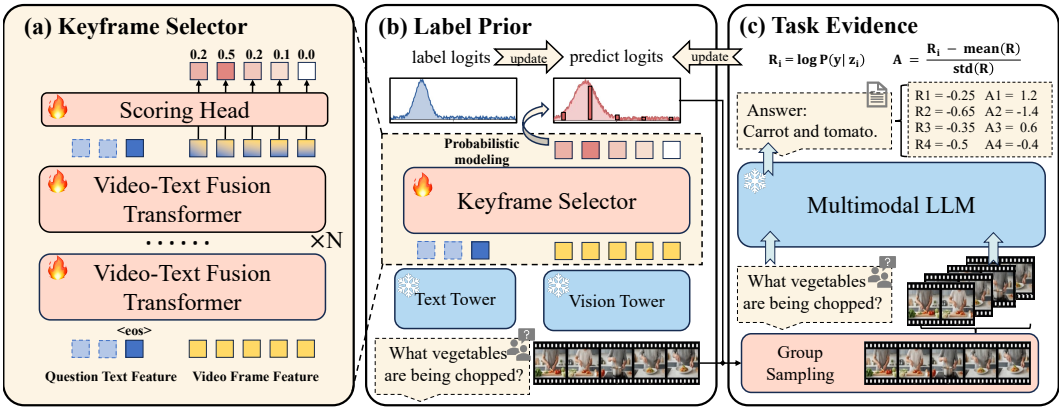


Figure 1: **The overall framework of our methods. (a) Keyframe Selector.** The lightweight selector ($f_\theta$) comprises stacked Video-Text Fusion Transformers. It modulates frame features with query semantics to output importance logits, then parameterizing a Plackett-Luce distribution. **(b) Stage I: Label Prior Elicitation.** We utilize keyframe label annotations to construct a *Label Prior*. The selector is trained via pairwise ranking loss to align its predicted logits with the label logits, establishing a robust initialization. **(c) Stage II:Task-Evidence Posterior Inference.** We freeze the MLLM and treat it as a frozen task environment. The selector performs Group Sampling to propose discrete frame subsets ($\mathbf{z}$). The MLLM evaluates these subsets to provide dense NLL rewards ($R$) and advantages ($A$). The policy is refined via **Bayesian GRPO** to maximize task evidence while maintaining KL-regularization towards the label prior.

### 2.2 Bayesian Methods in Large Models

Bayesian frameworks [33, 46, 63, 17, 44, 45] have seen a resurgence for quantifying uncertainty and integrating priors in large models. Recent studies such as Bayesian Evaluation of LLM Behavior [39] and Bayesian Mixture of Experts [12] demonstrate that probabilistic modeling can effectively calibrate model outputs and mitigate hallucinations. Recent works [5, 62] formulates LLMs as prior action distributions, showing that Bayesian inference can significantly accelerate exploration in reinforcement learning. This line of research provides strong theoretical grounding for treating supervised signals as priors and downstream task feedback as evidence, a core tenet of our proposed framework.

### 2.3 Policy Optimization and Reinforcement Learning for MLLMs

Reinforcement learning has become a standard paradigm for aligning MLLMs [22, 37, 38, 15, 77, 26, 49, 75] with complex reasoning tasks. General-purpose reasoning has been enhanced by algorithms like Hint-GRPO [21, 29, 76], which applies Group Relative Policy Optimization to boost mathematical and multimodal reasoning capabilities. In the specific domain of frame selection, recent works have explored various decision-making formulations. Focus [82] models keyframe selection as a combinatorial multi-armed bandit problem to balance exploration and exploitation. Similarly, FrameMind [18, 68] employs trajectory-level RL to dynamically request visual information during reasoning. Unlike these methods which often rely on sparse rewards or heuristic bandits, our approach leverages a dense NLL-based reward signal within a rigorous Bayesian inference framework to achieve task-adaptive frame selection.

## 3 Methodology

We formulate long-video frame selection as a probabilistic latent-variable modeling problem, as shown in Figure 1. We use $\mathcal{V} \in \mathbb{R}^{N \times D}$ to denote a sequence of $N$ candidate frames uniformly sampled from the raw video, $\mathcal{Q}$ to denote a textual query, and $\mathbf{y}$ to denote the ground-truth answer. Here, $N$ is reserved for the number of candidate frames, while $L$ denotes the number of Transformer blocks in the selector.

Our objective is to learn a parameterized policy $\pi_\theta(\mathbf{z} \mid \mathcal{V}, \mathcal{Q})$ that selects a discrete subset $\mathbf{z} \subset \{1, \ldots, N\}$ with cardinality $|\mathbf{z}| = K$, maximizing the marginal likelihood of the correct answer under a frozen Multimodal Large Language Model.

### 3.1 Variational Policy Generator

The true posterior distribution $P(\mathbf{z} \mid \mathbf{y}, \mathcal{V}, \mathcal{Q})$ is intractable due to the combinatorial nature of the discrete space $\binom{N}{K}$. To address this, we employ variational inference by introducing a variational family $q_\theta(\mathbf{z} \mid \mathcal{V}, \mathcal{Q})$ to approximate the posterior.

We implement the variational distribution using a lightweight selector network $f_\theta$. The network projects the visual and textual inputs into a sequence of unnormalized importance scores (logits) $\mathbf{s} \in \mathbb{R}^N$:

$$\mathbf{s} = f_\theta(\mathcal{V}, \mathcal{Q}). \tag{1}$$

To define a valid probability measure over ordered discrete subsets without replacement, we parameterize $q_\theta$ using the Plackett-Luce model. The probability of sampling an ordered sequence $\mathbf{z} = (z_1, \ldots, z_K)$ is factorized as:

$$q_\theta(\mathbf{z} \mid \mathbf{s}) = \prod_{k=1}^{K} \frac{\exp(s_{z_k})}{\sum_{j \in \mathcal{U}_k} \exp(s_j)}, \tag{2}$$

where $\mathcal{U}_k = \{1, \ldots, N\} \setminus \{z_1, \ldots, z_{k-1}\}$ represents the set of available frames at step $k$. This formulation ensures a normalized probability distribution over the combinatorial action space, enabling consistent gradient estimation.

### 3.2 Stage I: Prior Elicitation via Ranking

In Bayesian inference, the quality of the prior distribution $p(\mathbf{z})$ critically dictates the efficiency of posterior convergence. A non-informative uniform prior over the vast search space $\binom{N}{K}$ results in severe sparsity issues. We therefore interpret the first training stage as **Prior Elicitation**, leveraging human-annotated keyframes to establish an informative prior $\pi_{\text{prior}}(\mathbf{z})$.

We construct a pairwise preference dataset $\mathcal{D} = \{(i, j) \mid y_i > y_j\}$ derived from human annotations. We model the local relative saliency using the Bradley-Terry-Luce model:

$$P(i \succ j \mid \mathbf{s}) = \frac{\exp(s_i)}{\exp(s_i) + \exp(s_j)} = \sigma(s_i - s_j). \tag{3}$$

The parameters are estimated by minimizing the negative log-likelihood of the observed preferences:

$$\mathcal{L}_{\text{prior}}(\theta) = -\mathbb{E}_{(i,j)\sim\mathcal{D}} \left[ \log \sigma(s_i - s_j) \right]. \tag{4}$$

The resulting policy $\pi_{\text{prior}}$ restricts the support of the latent distribution to a semantically plausible region, effectively initializing the variational approximation near the modes of human visual perception.

### 3.3 Stage II: Task-Evidence Posterior Inference

While $\pi_{\text{prior}}$ captures general saliency, it remains agnostic to the specific reasoning dynamics of the MLLM. To align the selection with task requirements, we perform posterior inference using the frozen MLLM as the likelihood function $P_{\text{MLLM}}(\mathbf{y} \mid \mathbf{z}, \mathcal{Q})$.

#### 3.3.1 Derivation of the Evidence Lower Bound

We aim to maximize the log-marginal likelihood of the ground-truth answer $\mathbf{y}$. By marginalizing over the latent variable $\mathbf{z}$, we have:

$$\log P(\mathbf{y} \mid \mathcal{V}, \mathcal{Q}) = \log \sum_{\mathbf{z}} P_{\text{MLLM}}(\mathbf{y} \mid \mathbf{z}) \pi_{\text{prior}}(\mathbf{z}). \tag{5}$$

Direct maximization of Eq. 5 is intractable. We resort to maximizing its lower-bound by introducing the variational approximation $q_\theta(\mathbf{z})$ and applying Jensen's Inequality:

$$\begin{aligned}\log P(\mathbf{y}) &= \log \mathbb{E}_{\mathbf{z}\sim q_\theta}\left[\frac{P_{\mathrm{MLLM}}(\mathbf{y}\mid\mathbf{z})\pi_{\mathrm{prior}}(\mathbf{z})}{q_\theta(\mathbf{z})}\right] \\ &\geq \mathbb{E}_{\mathbf{z}\sim q_\theta}\left[\log\left(\frac{P_{\mathrm{MLLM}}(\mathbf{y}\mid\mathbf{z})\pi_{\mathrm{prior}}(\mathbf{z})}{q_\theta(\mathbf{z})}\right)\right] \\ &= \mathcal{L}_{\mathrm{ELBO}}(\theta).\end{aligned} \tag{6}$$

Expanding the logarithmic term reveals the structural components of our objective:

$$\mathcal{L}_{\mathrm{ELBO}}(\theta) = \mathbb{E}_{\mathbf{z}\sim q_\theta}\left[\log P_{\mathrm{MLLM}}(\mathbf{y}\mid\mathbf{z})\right] - D_{\mathrm{KL}}(q_\theta\|\pi_{\mathrm{prior}}). \tag{7}$$

The first term represents the *Task Evidence* (Likelihood), ensuring the selected frames support the correct answer. The second term serves as *Semantic Regularization*, explicitly constraining the policy towards the human-derived prior $\pi_{\mathrm{prior}}$ to prevent reward hacking.

The above derivation provides a principled foundation for our reward design within the latent-variable formulation in Eq. 5. Under this formulation, the answer likelihood term naturally induces the token-level negative log-likelihood (NLL) over the ground-truth answer as a dense task-evidence signal. Furthermore, the KL term explicitly regularizes the policy towards the *Prior* (human semantics) rather than an arbitrary previous policy, ensuring semantic consistency.

#### 3.3.2 The Tempered Posterior Objective

To control the strength of the regularization and improve optimization stability, we introduce a temperature coefficient $\beta$. The final maximization objective is:

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{z}\sim q_\theta}[R(\mathbf{z})] - \beta D_{\mathrm{KL}}(q_\theta\|\pi_{\mathrm{prior}}), \tag{8}$$

where $R(\mathbf{z}) \triangleq \log P_{\mathrm{MLLM}}(\mathbf{y}\mid\mathbf{z})$. We show that the optimal policy under this objective takes the form of a tempered posterior.

**Proposition 1.** *The variational distribution $q^*(\mathbf{z})$ that maximizes $\mathcal{J}(\theta)$ is given by*

$$q^*(\mathbf{z}) = \frac{1}{Z}\pi_{\mathrm{prior}}(\mathbf{z})\exp\left(\frac{R(\mathbf{z})}{\beta}\right), \tag{9}$$

*where $Z$ is the normalizing constant.*

*Proof.* Consider the KL divergence from $q_\theta$ to the tempered posterior $q^*$:

$$\begin{aligned}D_{\mathrm{KL}}(q_\theta\|q^*) &= \mathbb{E}_{q_\theta}\left[\log q_\theta(\mathbf{z}) - \log q^*(\mathbf{z})\right] \\ &= \mathbb{E}_{q_\theta}\left[\log q_\theta(\mathbf{z}) - \log \pi_{\mathrm{prior}}(\mathbf{z}) - \frac{1}{\beta}R(\mathbf{z}) + \log Z\right] \\ &= D_{\mathrm{KL}}(q_\theta\|\pi_{\mathrm{prior}}) - \frac{1}{\beta}\mathbb{E}_{q_\theta}[R(\mathbf{z})] + \log Z \\ &= -\frac{1}{\beta}\left(\mathbb{E}_{q_\theta}[R(\mathbf{z})] - \beta D_{\mathrm{KL}}(q_\theta\|\pi_{\mathrm{prior}})\right) + \log Z \\ &= -\frac{1}{\beta}\mathcal{J}(\theta) + \log Z.\end{aligned} \tag{10}$$

Since $Z$ is independent of $q_\theta$, maximizing $\mathcal{J}(\theta)$ is equivalent to minimizing $D_{\mathrm{KL}}(q_\theta\|q^*)$. Therefore, the optimal variational distribution is the tempered posterior in Eq. 9. □

### 3.4 Bayesian Group Relative Policy Optimization

Optimization of Eq. 8 requires estimating gradients over discrete expectations. We derive a Bayesian variant of Group Relative Policy Optimization (GRPO). Unlike standard RL which constrains updates against a previous policy $\pi_{\mathrm{old}}$, our derivation requires specific regularization towards $\pi_{\mathrm{prior}}$.

We utilize the log-derivative trick to estimate the gradient of the expected reward:

$$\nabla_\theta \mathbb{E}_{q_\theta}[R(\mathbf{z})] = \mathbb{E}_{q_\theta}[R(\mathbf{z}) \nabla_\theta \log q_\theta(\mathbf{z})]. \tag{11}$$

To reduce variance, we employ group-based importance sampling. For each query, we sample a group of $G$ trajectories $\{\mathbf{z}_i\}_{i=1}^G$ from the behavior policy $\pi_{\text{old}}$. The normalized advantage $\hat{A}_i$ is computed relative to the group mean:

$$\hat{A}_i = \frac{R(\mathbf{z}_i) - \frac{1}{G}\sum_j R(\mathbf{z}_j)}{\sigma_R + \epsilon}. \tag{12}$$

In practice, $R(\mathbf{z})$ is computed only over the ground-truth answer tokens under a unified short-form answer format. The group-relative normalization in Eq. 13 makes the policy update depend on the relative ranking of sampled frame subsets for the same query. Consequently, the optimization is robust to moderate variations in answer length and output format. The final gradient estimator is given by:

$$\nabla_\theta \mathcal{J} \approx \frac{1}{G}\sum_{i=1}^{G} \Big( \hat{A}_i \nabla_\theta \log q_\theta(\mathbf{z}_i) - \beta \nabla_\theta D_{\text{KL}}(q_\theta(\mathbf{z}_i) \| \pi_{\text{prior}}(\mathbf{z}_i)) \Big). \tag{13}$$

In practice, we implement this using the clipped surrogate objective to ensure monotonic improvement:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G} \min\Big(\rho_i \hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big) + \beta D_{\text{KL}}(q_\theta \| \pi_{\text{prior}}), \tag{14}$$

where $\rho_i = q_\theta(\mathbf{z}_i)/\pi_{\text{old}}(\mathbf{z}_i)$ is the importance sampling ratio. This objective completes the Bayesian loop: the advantage term $\hat{A}_i$ drives the policy towards the likelihood (Task Evidence), while the KL term anchors it to the learned prior, strictly adhering to the ELBO formulation. The complete training procedure, illustrating the transition from Prior Elicitation to Posterior Inference, is summarized in Algorithm 1.

# 4 Experiment

## 4.1 Experimental Setup

**Models and Baselines.** We compare our approach against two distinct categories of frame selection strategies, lightweight scoring and MLLM-based methods. The first category consists of lightweight scoring methods. This includes methods utilizing efficient vision-language models, such as SigLIP2 [55], to measure semantic alignment. Specific approaches include AKS [54], which segments videos based on adaptive similarity; BOLT [32], which computes importance scores followed by inverse transform sampling; and Q-Frame [74], which employs dynamic resolution based on cross-modal retrieval. F2C [52] utilizes CLIP-based anchor keys to adjust clip lengths during selection. Optimization-based methods like MDP3 [53] and Nar-KFC [13] formulate selection as subset or subgraph problems using mathematical estimation. Additionally, Focus [82] employs multi-armed bandits to balance exploration and exploitation, while Frame-Voyager [66] and FrameOracle [30] use pairwise ranking or joint importance prediction to identify keyframe sets. The second category, MLLM-based methods, leverages heavy multimodal models for direct scoring. Examples include ViaRL [61] and Kframes [65], which apply reinforcement learning for temporal grounding; A.I.R. [83], which uses a two-stage iterative selection process; and GenS [64], which treats frame scoring as a generative sampling task. Results are reported over multiple runs. **Efficiency and Training Cost Analysis** are provided in Appendix C.

**Benchmarks and Implementation Details.** We evaluate our method on three widely adopted long-form video benchmarks, covering diverse domains and reasoning granularities. Details about benchmarks are provided in Appendix C. We utilize SigLIP2-SO400M-Patch16-512 [55] as the shared visual-textual backbone for feature pre-extraction. More details are provided in Appendix C.

## 4.2 Results on Long-form Video Benchmarks

We present the quantitative evaluation results on Video-MME, LongVideoBench, and MLVU under different frame budgets in Tables 1, 2, and 3. Notably, our approach maintains competitive performance with all frame budget, demonstrating the efficacy of our keyframe selection strategy.

Table 1: Comparison of frame selection methods on **Qwen2.5-VL-7B**. Methods are categorized by type and grouped by frame count. **Bold**/underline denote the best/second-best performance.

| Method | # Frames | Video-MME | | | | LVB | MLVU | |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Overall | | Needle-QA | Avg |
| *Avg. Video Duration* | | *1.3min* | *9min* | *41min* | *17min* | *12min* | *12min* | |
| ***Uniform Sampling Baseline*** | | | | | | | | |
| Uniform | 8 | 61.7 | 50.6 | 45.7 | 52.6 | 52.6 | 62.8 | 56.1 |
| Uniform | 32 | 71.4 | 60.8 | 49.2 | 60.5 | 56.3 | 72.4 | 62.2 |
| ***Lightweight Scoring Methods*** | | | | | | | | |
| AKS | 8 | 62.1 | 51.4 | 48.8 | 54.1 | 52.0 | - | 53.7 |
| BOLT | 8 | 66.0 | 54.6 | 50.4 | 57.0 | 55.6 | - | 59.0 |
| Q-Frame | 8 | 71.9 | 58.2 | 50.7 | 60.3 | 56.6 | - | 58.2 |
| F2C | 8 | **72.4** | **60.7** | 51.2 | 61.4 | **58.9** | - | 63.1 |
| **Ours** | 8 | 71.2 | 60.4 | **52.9** | **61.5** | 58.2 | **82.0** | **66.1** |
| MDP3 | 32 | - | - | - | 63.8 | 60.0 | - | 66.2 |
| AKS | 32 | 66.4 | 56.9 | 48.6 | 57.3 | 58.3 | - | 63.1 |
| BOLT | 32 | 74.3 | 64.2 | 53.8 | 64.1 | 58.6 | - | 66.3 |
| Q-Frame | 32 | 73.0 | 61.7 | 53.1 | 62.6 | 58.7 | - | - |
| F2C | 32 | 73.8 | **66.3** | **56.6** | **65.6** | 60.8 | - | 66.8 |
| **Ours** | 32 | **74.1** | 65.8 | 55.7 | 65.2 | 61.1 | **82.5** | **70.3** |
| ***MLLM-based Scoring Methods*** | | | | | | | | |
| ViaRL | 8 | 65.1 | 56.1 | 50.8 | 57.3 | - | 73.5 | 58.2 |
| Kframes | 8 | 68.9 | 55.3 | 47.9 | 57.4 | 57.7 | 77.5 | 60.4 |
| Kframes | 32 | **74.1** | 61.4 | 51.7 | 62.1 | 60.5 | 79.4 | 65.9 |
| A.I.R. | 32 | - | - | - | 65.0 | **61.4** | - | 67.5 |
| GenS | 54 | - | - | - | - | 58.7 | - | 64.8 |

**Comparison with State-of-the-Art methods.** As evidenced in Tables 1 and 2, our method significantly outperforms existing lightweight selectors while achieving parity with computationally intensive MLLM-based approaches. Our model with $K = 8$ frames surpasses the strong uniform baseline by a substantial margin of 8.9% on Video-MME and 10.0% on MLVU and outperforming lightweight methods such as BOLT and Q-Frame on Qwen2.5-VL(Table 1). Our method even achieves an overall accuracy of 66.8% with $K = 32$ which effectively matches the performance of the MLLM-based scorer A.I.R. on InternVL-3(Table 2). Furthermore, compared to mathematical iterative methods like Nar-KFC and MDP3 that demand extensive test-time computation, our module offers a direct single-pass forward inference.

**Generalization to State-of-the-Art MLLMs.** We extend the evaluation to SOTA MLLMs: Qwen3-VL-8B-Instruct (no-thinking) and InternVL-3.5-8B (thinking) in Table 3 to assess the generalization. On Qwen3-VL ($K = 32$), our method improves the overall Video-MME score from 63.5% to 65.3% and the MLVU average from 65.3% to 72.7%. We observe that the relative margin over uniform sampling is slightly compressed compared to earlier architectures (e.g., Qwen2.5-VL), which can be attributed to the enhanced reasoning robustness of SOTA MLLMs from suboptimal visual contexts or partial evidence.

### 4.3 Ablation Studies

**Cross-environment Stage II Training.** A natural concern is whether the posterior inference stage must be repeated for each downstream MLLM, since the NLL reward is obtained from a frozen MLLM environment. We compare two selectors trained with different Stage II environments and evaluate both on InternVL-3-8B. The first selector is trained with Qwen2.5-VL-7B as the frozen environment, which is our default setting. The second selector is trained by replacing the Stage II environment with InternVL-3-8B.

As shown in Table 4, using InternVL-3-8B as the Stage II environment yields nearly identical performance to the selector trained with Qwen2.5-VL-7B. This result indicates that Stage II does not need to be re-run for each downstream MLLM in our setting. It suggests that the posterior update

Table 2: Comparison of frame selection methods on **InternVL-3-8B** and **LLaVA-OneVision-7B**. Methods are categorized by type, grouped by model, and then by frame count. **Bold**/underline denote the best/second-best performance.

| Method | # Frames | Video-MME | | | | LVB | MLVU | |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Overall | | Needle-QA | Avg |
| *Avg. Video Duration* | | *1.3min* | *9min* | *41min* | *17min* | *12min* | *12min* | |
| ***Uniform Sampling Baseline*** | | | | | | | | |
| **InternVL-3-8B** | | | | | | | | |
| Uniform | 8 | 68.7 | 58.3 | 48.8 | 58.6 | 53.4 | 77.2 | 64.4 |
| Uniform | 32 | 74.2 | 62.8 | 51.7 | 62.9 | 57.9 | 83.7 | 70.3 |
| **LLaVA-OneVision-7B** | | | | | | | | |
| Uniform | 8 | 62.6 | 51.6 | 44.3 | 52.8 | 52.7 | 63.7 | 60.2 |
| Uniform | 32 | 69.3 | 54.7 | 46.2 | 56.7 | 55.2 | 76.6 | 64.7 |
| ***Lightweight Scoring Methods*** | | | | | | | | |
| **InternVL-3-8B** | | | | | | | | |
| **Ours** | 8 | **74.7** | **60.9** | **51.6** | **62.4** | **58.7** | **85.1** | **69.0** |
| MDP3 | 32 | - | - | - | 66.8 | 60.9 | - | 74.0 |
| **Ours** | 32 | **76.2** | **65.9** | **58.4** | 66.8 | 62.1 | **85.4** | 74.3 |
| **LLaVA-OneVision-7B** | | | | | | | | |
| Frame-Voyager | 8 | 67.3 | 56.3 | 48.9 | 57.5 | - | - | 65.6 |
| FrameOracle | 8 | - | - | - | 57.5 | 56.0 | - | 62.9 |
| MDP3 | 8 | **71.8** | **57.6** | **49.6** | **59.6** | **59.0** | - | **69.8** |
| BOLT | 8 | 66.8 | 54.2 | 47.3 | 56.1 | 55.6 | - | 63.4 |
| Nar-KFC | 8 | - | - | - | 57.8 | 56.5 | - | 66.2 |
| **Ours** | 8 | 68.3 | 56.3 | 48.2 | 57.6 | 57.8 | 74.3 | 67.8 |
| Focus | 32 | - | - | - | 58.3 | 60.7 | - | - |
| AKS | 32 | - | - | - | 58.4 | 59.3 | - | - |
| MDP3 | 32 | - | - | - | 60.5 | 59.0 | - | 68.3 |
| BOLT | 32 | 70.1 | 60.0 | 49.6 | 59.9 | 59.6 | - | 66.8 |
| **Ours** | 32 | **70.2** | **61.4** | **50.7** | 60.8 | **60.8** | **76.1** | 69.2 |
| ***MLLM-based Scoring Methods*** | | | | | | | | |
| **InternVL-3-8B** | | | | | | | | |
| A.I.R. | 32 | - | - | - | **68.2** | **62.8** | - | **74.5** |
| **LLaVA-OneVision-7B** | | | | | | | | |
| A.I.R. | 32 | - | - | - | **61.4** | 60.7 | - | **69.3** |

does not merely overfit to the idiosyncratic preference of one frozen environment but helps move the selector more useful across MLLMs.

**Efficacy of Bayesian Posterior Update and Task Evidence.** Table 5 presents results of different training strategies on Qwen2.5-VL. The SigLIP2 baseline selects frames based on cosine similarity. Label Prior (Stage I) improves the selection quality by aligning with label saliency, yet the bias of static annotations limits the performance. Crucially, applying Reinforcement Learning with a label-based sparse reward does not align with the inference dynamics of the downstream model, and the reward signal is overly sparse, failing to produce meaningful gains over the Label Prior. In contrast, our method utilizing the Negative Log-Likelihood (NLL) as a dense reward signal achieves significant performance improvements across all metrics.

**Impact of Frame Budget and Saturation.** We further investigate the efficiency-performance trade-off by varying the number of selected frames $K$ from 8 to 128, as detailed in Table 12. More experiments are provided in Appendix E.

## 5 Conclusion

We formulate frame selection as Bayesian inference, using Bayesian GRPO to refine annotation-derived priors with downstream task evidence. By leveraging the MLLM's negative log-likelihood as a dense reward, our lightweight selector improves video-QA accuracy with modest overhead and transfers across MLLMs. Future work will explore streaming video.

Table 3: Transferability analysis on state-of-the-art MLLMs (**Qwen3-VL-8B** and **InternVL-3.5-8B**). We compare the original model performance (Uniform) with our selector. **Bold** denotes the best performance.

| **Method** | **# Frames** | **Video-MME** | | | | **LVB** | **MLVU** | |
|---|---|---|---|---|---|---|---|---|
| | | Short | Medium | Long | Overall | | Needle-QA | Avg |
| *Avg. Video Duration* | | *1.3min* | *9min* | *41min* | *17min* | *12min* | *12min* | |
| **Qwen3-VL-8B** | | | | | | | | |
| Uniform | 8 | 64.8 | 51.8 | 47.7 | 54.7 | 53.3 | 60.4 | 56.9 |
| **Ours** | 8 | **67.8** | **56.2** | **50.8** | **58.3** | **58.2** | **79.8** | **66.7** |
| Uniform | 32 | 76.1 | 61.6 | 52.8 | 63.5 | 58.9 | 71.6 | 65.3 |
| **Ours** | 32 | **77.6** | **63.6** | **54.8** | **65.3** | **62.7** | **82.3** | **72.7** |
| **InternVL-3.5-8B** | | | | | | | | |
| Uniform | 8 | 64.7 | 55.9 | 46.9 | 55.8 | 56.7 | 75.5 | 62.7 |
| **Ours** | 8 | **70.6** | **56.2** | **48.2** | **58.3** | **59.6** | **86.8** | **67.6** |
| Uniform | 32 | 72.9 | 60.0 | 50.3 | 61.1 | 60.1 | 83.1 | 68.5 |
| **Ours** | 32 | **76.0** | **63.2** | **52.9** | **64.1** | **63.4** | **85.4** | **71.4** |

Table 4: **Cross-environment Stage II training.** Both selectors are evaluated on InternVL-3-8B with $K = 8$ and $K = 32$. The selector trained with Qwen2.5-VL transfers directly to InternVL-3 without additional posterior training and achieves comparable performance to the selector trained with InternVL-3 itself. **Bold** denotes the best performance.

| **Stage II Env.** | **Video-MME** | | **LVB** | | **MLVU** | |
|---|---|---|---|---|---|---|
| | $K = 8$ | $K = 32$ | $K = 8$ | $K = 32$ | $K = 8$ | $K = 32$ |
| **Evaluation: InternVL-3-8B** | | | | | | |
| Qwen2.5-VL-7B | **62.4** | **66.8** | 58.7 | **62.1** | 69.0 | **74.3** |
| InternVL-3-8B | **62.4** | 66.5 | **59.0** | 61.9 | **69.2** | 74.2 |

Table 5: **Ablation study on training strategies.** All results are evaluated using the **Qwen2.5-VL-7B** model. We investigate the impact of the Label Prior (Stage I) and the Bayesian GRPO update (Stage II) across different frame budgets ($K = 8, 32$). Reward Source indicates the reward signal used during GRPO.

| **Method** | **Training Components** | | | **K = 8** | | | **K = 32** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Stage I** | **Stage II** | **Reward** | **V-MME** | **LVB** | **MLVU** | **V-MME** | **LVB** | **MLVU** |
| SigLIP2 | ✗ | ✗ | - | 54.7 | 56.6 | 61.9 | 61.1 | 59.7 | 66.3 |
| Label Prior | ✓ | ✗ | - | 58.5 | 57.4 | 65.2 | 63.6 | 60.4 | 68.2 |
| + Label Reward | ✓ | ✓ | label | 58.3 | 57.6 | 64.9 | 63.8 | 60.2 | 67.6 |
| **+ Task Evidence (Ours)** | ✓ | ✓ | **NLL** | **61.5** | **58.7** | **66.1** | **65.2** | **61.1** | **70.3** |

Table 6: **Comparison of reward types for Stage II.** Results are evaluated on Qwen2.5-VL-7B with $K = 8$ and $K = 32$. Label reward provides medium-granularity but annotation-biased supervision. RLVR-style correctness reward is task-level but sparse. NLL provides dense task evidence and performs best across benchmarks. **Bold** denotes the best performance.

| **Reward Type** | **Reward Properties** | | $K = 8$ | | | $K = 32$ | | |
|---|---|---|---|---|---|---|---|---|
| | **Granularity** | **Bias** | **V-MME** | **LVB** | **MLVU** | **V-MME** | **LVB** | **MLVU** |
| **Stage II reward signal** | | | | | | | | |
| Label | Frame-level / medium | Annotation-biased | 58.3 | 57.6 | 64.9 | 63.8 | 60.2 | 67.6 |
| RLVR-style correctness | Sample-level / sparse | Task-level | 57.3 | 56.5 | 65.3 | 62.1 | 58.9 | 66.7 |
| **NLL** | **Token-level / dense** | **Task-level** | **61.5** | **58.7** | **66.1** | **65.2** | **61.1** | **70.3** |

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62471287).

## References

[1] Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Temporal chain of thought: Long-video understanding by thinking in frames, 2025. URL https://arxiv.org/abs/2507.02001.

[2] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Mingchen Zhuge, Jian Ding, Deyao Zhu, Jürgen Schmidhuber, and Mohamed Elhoseiny. Goldfish: Vision-language understanding of arbitrarily long videos, 2024. URL https://arxiv.org/abs/2407.12679.

[3] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding, 2022. URL https://arxiv.org/abs/2206.01720.

[4] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29071–29082, 2025.

[5] Matteo Cercola, Valeria Capretti, and Simone Formentin. Efficient reinforcement learning from human feedback via bayesian preference inference, 2025. URL https://arxiv.org/abs/2511.04286.

[6] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents, 2025. URL https://arxiv.org/abs/2503.10200.

[7] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video, 2024. URL https://arxiv.org/abs/2406.11816.

[8] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions, 2024. URL https://arxiv.org/abs/2406.04325.

[9] Yuzhe Chen, Jiale Cao, Xuyang Liu, Jin Xie, Aiping Yang, and Yanwei Pang. Stdec: Spatio-temporal stability guided decoding for dllms. *arXiv preprint arXiv:2604.06330*, 2026.

[10] Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. Scaling video-language models to 10k frames via hierarchical differential distillation, 2025. URL https://arxiv.org/abs/2504.02438.

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

[12] Maryam Dialameh, Hossein Rajabzadeh, Weiwei Zhang, Walid Ahmed, and Hyock Ju Kwon. Bayesian mixture of experts for large language models, 2025. URL https://arxiv.org/abs/2511.08968.

[13] Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. Threading keyframe with narratives: Mllms as strong long video comprehenders, 2025. URL https://arxiv.org/abs/2505.24158.

[14] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos, 2024. URL https://arxiv.org/abs/2408.14023.

[15] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. URL https://arxiv.org/abs/2503.21776.

[16] 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, 2025. URL https://arxiv.org/abs/2405.21075.

[17] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

[18] Haonan Ge, Yiwei Wang, Kai-Wei Chang, Hang Wu, and Yujun Cai. Framemind: Frame-interleaved video reasoning via reinforcement learning, 2025. URL https://arxiv.org/abs/2509.24008.

[19] Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning, 2025. URL https://arxiv.org/abs/2506.00318.

[20] Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding, 2025. URL https://arxiv.org/abs/2502.06428.

[21] Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo, 2025. URL https://arxiv.org/abs/2503.23905.

[22] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

[23] Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. Adaptive sample scheduling for direct preference optimization. *arXiv preprint arXiv:2506.17252*, 2025.

[24] Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, et al. Does your reasoning model implicitly know when to stop thinking? *arXiv preprint arXiv:2602.08354*, 2026.

[25] Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuefeng Xiao, Hongyan Xie, Huaqiu Li, Songshi Liang, Zhongxiang Dai, Fuzhen Zhuang, Jianxin Li, Yikun Ban, and Deqing Wang. Real-time aligned reward model beyond semantics. 2026. URL https://api.semanticscholar.org/CorpusID:285240754.

[26] Zixuan Huang, Yang Zhou, Kaixuan Wang, Guli Zhang, Hongyan Xie, Yakun Zhu, Hao Geng, Yikun Ban, and Deqing Wang. Deferred exposure of future trajectories for verifiable reasoning in autonomous driving vlms. *arXiv preprint arXiv:2608.01755*, 2026.

[27] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding, 2024. URL https://arxiv.org/abs/2311.08046.

[28] Donghyuk Kim, Sejeong Yang, Wonjin Shin, and Joo-Young Kim. V-rex: Real-time streaming video llm acceleration via dynamic kv cache retrieval, 2025. URL https://arxiv.org/abs/2512.12284.

[29] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[30] Chaoyu Li, Tianzhi Li, Fei Tao, Zhenyu Zhao, Ziqian Wu, Maozheng Zhao, Juntong Song, Cheng Niu, and Pooyan Fazli. Frameoracle: Learning what to see and how much to see in videos, 2025. URL https://arxiv.org/abs/2510.03584.

[31] Xiaoyuan Li, Yuzhe Wang, Moxin Li, Keqin Bao, Rui Men, Yichang Zhang, Dayiheng Liu, Wenjie Wang, and Fuli Feng. Sage: Scalable automated robustness augmentation for llm knowledge evaluation. *arXiv preprint arXiv:2605.12022*, 2026.

[32] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding, 2025. URL https://arxiv.org/abs/2503.21483.

[33] Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization, 2024. URL https://arxiv.org/abs/2402.03921.

[34] Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-xl-pro: Reconstructive token compression for extremely long video understanding, 2025. URL https://arxiv.org/abs/2503.18478.

[35] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning, 2025. URL https://arxiv.org/abs/2503.13444.

[36] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

[37] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2503.01785.

[38] Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual agentic reinforcement fine-tuning, 2025. URL https://arxiv.org/abs/2505.14246.

[39] Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior, 2025. URL https://arxiv.org/abs/2511.10661.

[40] Wufei Ma, Kai Li, Zhongshi Jiang, Moustafa Meshry, Qihao Liu, Huiyu Wang, Christian Häne, and Alan Yuille. Rethinking video-text understanding: Retrieval from counterfactually augmented data, 2024. URL https://arxiv.org/abs/2407.13094.

[41] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Bohm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, Mark Todorovich, Ingo Bax, and Roland Memisevic. What to say and when to say it: Live fitness coaching as a testbed for situated interaction, 2024. URL https://arxiv.org/abs/2407.08101.

[42] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models, 2024. URL https://arxiv.org/abs/2405.16009.

[43] Minghao Qin, Xiangrui Liu, Zhengyang Liang, Yan Shu, Huaying Yuan, Juenjie Zhou, Shitao Xiao, Bo Zhao, and Zheng Liu. Video-xl-2: Towards very long-video understanding through task-aware kv sparsification, 2025. URL https://arxiv.org/abs/2506.19225.

[44] Xiaohan Qin, Victor Wang, Ning Liao, Cancheng Zhang, Xiangdong Zhang, Mingquan Feng, Jingzhi Wang, and Junchi Yan. sstoken: Self-modulated and semantic-aware token selection for llm fine-tuning. In *International Conference on Learning Representations*, volume 2026, pages 79347–79368, 2026.

[45] Xiaohan Qin, Xiangdong Zhang, Tuo Dai, Sunan Zou, Huaijin Wu, Yebin Yang, Zhuo Xia, Lin Yao, Yuliang Liu, and Junchi Yan. Multi-stream attention residuals. *arXiv preprint arXiv:2610.13630*, 2026.

[46] Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, January 2026. ISSN 2041-1723. doi: 10.1038/s41467-025-67998-6. URL http://dx.doi.org/10.1038/s41467-025-67998-6.

[47] Saul Santos, António Farinhas, Daniel C. McNamee, and André F. T. Martins. $\infty$-video: A training-free approach to long video understanding via continuous-time memory consolidation, 2025. URL https://arxiv.org/abs/2501.19098.

[48] Yifan Shen, Boyi Li, Meihuan Huang, Yuanzhe Liu, Xu Cao, Jinyang Jin, Zhengyuan Li, Anglin Liu, Junho Kim, Jingyuan Zhu, et al. Decoding children's gait behavior. *arXiv preprint arXiv:2608.00371*, 2026.

[49] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

[50] Yuan Sheng, Yanbin Hao, Chenxu Li, Shuo Wang, and Xiangnan He. Sevices: Unifying semantic-visual evidence consensus for long video understanding, 2025. URL https://arxiv.org/abs/2510.20622.

[51] 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, 2024. URL https://arxiv.org/abs/2409.14485.

[52] Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From frames to clips: Training-free adaptive key clip selection for long-form video understanding, 2025. URL https://arxiv.org/abs/2510.02262.

[53] Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms, 2025. URL https://arxiv.org/abs/2501.02885.

[54] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding, 2025. URL https://arxiv.org/abs/2502.21271.

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

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

[57] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. URL https://arxiv.org/abs/2501.12386.

[58] Zhuohan Wang, Haoran Ma, Tianyu Wu, Yuanlin Duan, Zichun Liao, and Jieming Yu. Solving every step is not enough: Milestone oracles reveal a composition gap in llm math reasoning. *arXiv preprint arXiv:2609.32235*, 2026.

[59] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

[60] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams, 2025. URL https://arxiv.org/abs/2510.09608.

[61] Ziqiang Xu, Qi Dai, Tian Xie, Yifan Yang, Kai Qiu, DongDong Chen, Zuxuan Wu, and Chong Luo. Viarl: Adaptive temporal grounding via visual iterated amplification reinforcement learning, 2025. URL https://arxiv.org/abs/2505.15447.

[62] Adam X. Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. Bayesian reward models for llm alignment, 2024. URL https://arxiv.org/abs/2402.13210.

[63] Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models, 2024. URL https://arxiv.org/abs/2308.13111.

[64] Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding, 2025. URL https://arxiv.org/abs/2503.09146.

[65] Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyan Zhao, Ke Tian, Zhihao Wang, Minghui Qiu, and Tao Wang. K-frames: Scene-driven any-k keyframe selection for long video understanding, 2025. URL https://arxiv.org/abs/2510.13891.

[66] Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, Hao Zhang, and Qianru Sun. Frame-voyager: Learning to query frames for video large language models, 2025. URL https://arxiv.org/abs/2410.03226.

[67] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025.

[68] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[69] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, Yali Wang, Yu Qiao, and Limin Wang. Timesuite: Improving mllms for long video understanding via grounded tuning, 2025. URL https://arxiv.org/abs/2410.19702.

[70] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025. URL https://arxiv.org/abs/2501.13106.

[71] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering, 2024. URL https://arxiv.org/abs/2312.17235.

[72] Gengyuan Zhang, Jisen Ren, Jindong Gu, and Volker Tresp. Multi-event video-text retrieval, 2026. URL https://arxiv.org/abs/2308.11551.

[73] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. URL https://arxiv.org/abs/2306.02858.

[74] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025. URL https://arxiv.org/abs/2506.22139.

[75] Xiangdong Zhang, Xiaohan Qin, Sunan Zou, Tuo Dai, Xiaoming Shi, Huaijin Wu, Yebin Yang, Zhuo Xia, Shaofeng Zhang, Lin Yao, et al. xhc: Expanded hyper-connections. *arXiv preprint arXiv:2607.14530*, 2026.

[76] Xiangdong Zhang, Debing Zhang, Shaofeng Zhang, Xiaohan Qin, Yu Cheng, and Junchi Yan. Nitp: Next implicit token prediction for llm pre-training. *arXiv preprint arXiv:2605.24956*, 2026.

[77] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Tinyllava-video-r1: Towards smaller lmms for video reasoning, 2025. URL https://arxiv.org/abs/2504.09641.

[78] Zhixia Zhang, Zixuan Huang, Xin Xia, Deqing Wang, Fuzhen Zhuang, Shuai Ma, Ning Ding, Yaodong Yang, Jianxin Li, and Yikun Ban. Heterogeneous agent collaborative reinforcement learning. *arXiv preprint arXiv:2603.02604*, 2026.

[79] Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. Streaming video model, 2023. URL https://arxiv.org/abs/2303.17228.

[80] 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, 2025. URL https://arxiv.org/abs/2406.04264.

[81] Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review, 2023. URL https://arxiv.org/abs/2302.12552.

[82] Zirui Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding, 2025. URL https://arxiv.org/abs/2510.27280.

[83] 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, 2025. URL https://arxiv.org/abs/2510.04428.

**Algorithm 1** Bayesian Frame Selection via Prior-Guided GRPO

**Input:** Video dataset $\mathcal{D}$, label dataset $\mathcal{D}_{\text{label}}$, frozen MLLM, budget $K$, group size $G$, KL coefficient $\beta$.
**Output:** Frame-selection policy $\pi_\theta$.
*// Stage I: prior elicitation*
Initialize policy parameters $\theta$.
**for** each batch $(\mathcal{V}, \mathcal{Q}, \mathbf{y}_{\text{pref}}) \in \mathcal{D}_{\text{label}}$ **do**
  Compute frame scores $\mathbf{s} = f_\theta(\mathcal{V}, \mathcal{Q})$.
  Update $\theta$ by minimizing the BTL ranking loss $\mathcal{L}_{\text{prior}}$ in Eq. 4.
**end for**
Freeze the elicited prior policy $\pi_{\text{prior}} \leftarrow \pi_\theta$.
*// Stage II: posterior inference*
**for** each batch $(\mathcal{V}, \mathcal{Q}, \mathbf{y}) \in \mathcal{D}$ **do**
  Sample $G$ frame-selection trajectories $\{\mathbf{z}_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid \mathbf{s})$.
  **for** $i = 1, \ldots, G$ **do**
    Evaluate task evidence $R(\mathbf{z}_i) = \log P_{\text{MLLM}}(\mathbf{y} \mid \mathbf{z}_i, \mathcal{Q})$.
  **end for**
  Normalize group-relative advantages $\{\hat{A}_i\}_{i=1}^{G}$.
  Compute prior regularization $D_{\text{KL}}(\pi_\theta(\mathbf{z}_i) \,\|\, \pi_{\text{prior}}(\mathbf{z}_i))$.
  Update $\theta$ by minimizing the Bayesian GRPO loss in Eq. 14.
**end for**

# A Overview

This appendix provides supplementary details to support the findings presented in the main paper. Algorithm is provided in Section B. Implementation details, including hyperparameters, evaluation prompts, and benchmark descriptions, are provided in Section C. We then describe the training dataset construction in Section D. Finally, we present extended ablation studies and sensitivity analyses on state-of-the-art MLLMs in Section E, followed by additional visualization cases in Section F.

# B Algorithm

The complete training procedure, illustrating the transition from Prior Elicitation to Posterior Inference, is summarized in Algorithm 1.

# C Implementation Details

In this section, we provide comprehensive implementation details to ensure the reproducibility of our method.

## C.1 Benchmarks.

**Video-MME** [16] serves as a large-scale multimodal evaluation suite, assessing capabilities in video understanding, reasoning, and knowledge-grounded tasks. It incorporates both short and long-form videos to test context management strategies comprehensively. **LongVideoBench** [59] focuses specifically on long-form scenarios, with videos spanning tens of minutes to hours. It challenges models with tasks such as VQA, summarization, and temporal grounding, rigorously testing temporal reasoning efficiency. **MLVU** [80] evaluates hierarchical comprehension. It tests models on multiple levels of reasoning, including frame-level recognition, clip-level temporal understanding, and video-level holistic comprehension, ensuring that the selection method preserves both local details and global context.

## C.2 Implementation Details.

For all frame-selection methods evaluated under our controlled setting, candidate frames are uniformly sampled from the raw video at 1 FPS before selection. Given a video of duration $T$ seconds, this produces $N = T$ candidate frames, and each method selects $K$ frames from the same candidate pool. Unless otherwise specified, lightweight baselines including AKS, BOLT, Q-Frame, F2C, and

our method use the same SigLIP2-SO400M features, ensuring that performance differences are not caused by different feature extractors.The lightweight selector consists of $L = 4$ blocks of our proposed Video-Text Fusion Transformers. The hidden dimension is set to $D = 1152$ to align with the visual and textual features. We train the selector on the GenS-Video-150K dataset with 150k video-question pairs. The training is conducted in two stages. In Stage I, we optimize the pairwise ranking loss with a learning rate of $1e^{-4}$. In Stage II, we fine-tune the selector using Bayesian GRPO with a learning rate of $5e^{-6}$. The Stage II environment is Qwen2.5-VL-7B, which is kept frozen throughout training. After training, the selector is directly deployed to other downstream MLLMs without additional posterior training. All training and evaluation experiments are conducted on 8 NVIDIA H100 (80GB) GPUs.

### C.3 Hyperparameters of Bayesian GRPO

We detail the hyperparameters used in both label prior elicitation (Stage I) and task evidence inference (Stage II) in Table 7. The training was conducted on 8 NVIDIA H100 GPUs.

Table 7: Hyperparameter settings for the two-stage training process.

| **Hyperparameter** | **Stage I: Label Prior** | **Stage II: Task Evidence** |
|---|---|---|
| Batch Size | 32 | 1 |
| Learning Rate | $1e^{-4}$ | $5e^{-6}$ |
| Optimizer | AdamW | AdamW |
| Weight Decay | 0.01 | 0.01 |
| Epochs | 10 | 1 |
| KL Coefficient ($\beta$) | - | 0.2 |
| Group Size ($G$) | - | 8 |
| PPO Epochs | - | 3 |
| Clip Ratio ($\epsilon$) | - | 0.2 |

### C.4 Efficiency and Training Cost Analysis

A central goal of our method is to preserve the inference efficiency of lightweight selectors while avoiding the heavy test-time overhead of MLLM-based scorers. We therefore report both computational cost and measured latency under a common setting. Following the evaluation protocol, candidate frames are sampled at 1 FPS. For a 300-second video with $K = 32$, we estimate the selector-side computational cost with fp16 model weights and report the peak memory usage.

As shown in Table 8, our selector introduces only a small overhead over SigLIP2-based lightweight methods, increasing the estimated cost from 240–241 GFLOPs to 264 GFLOPs and memory from 0.80GB to 0.93GB. In contrast, MLLM-based scorers require repeated large-model inference over candidate frames or windows, leading to orders-of-magnitude higher cost.

We further measure wall-clock latency on Video-MME using a single NVIDIA H100 GPU, with $K = 32$ and candidate frames sampled at 1 FPS. The average per-sample latency is reported in Table 9. Our method remains close to lightweight training-free methods and is substantially faster than generative MLLM-based scoring.

Although our method is not training-free, the training cost is incurred only once. Stage I learns the label prior from keyframe annotations, and Stage II uses a frozen Qwen2.5-VL-7B environment for posterior refinement. After training, the selector is reused as a standalone module and directly transferred to different downstream MLLMs without re-running Stage II. Therefore, the training cost is amortized across benchmarks, frame budgets, and downstream MLLMs, while inference requires only one selector forward pass for each video-query pair.

### C.5 Evaluation Details

We utilize the following structured prompt to evaluate MLLMs across all video benchmarks. The temperature is set to 0.0 to ensure deterministic generation. Inference is performed using `transformers==4.57`, and accuracy is calculated using Exact Match.

Table 8: **Selector-side efficiency comparison.** GFLOPs and peak memory are measured on a 300-second video sampled at 1 FPS with $K = 32$. BOLT, F2C, and ours use the same SigLIP2-SO400M feature extractor.

| Method | Encoder / Scorer | GFLOPs | Mem. (GB) |
|---|---|---|---|
| *Lightweight feature-based scoring* | | | |
| BOLT | SigLIP2-SO400M | 240 | 0.80 |
| F2C | SigLIP2-SO400M | 241 | 0.80 |
| **Ours** | SigLIP2-SO400M + selector | **264** | **0.93** |
| *MLLM-based scoring* | | | |
| A.I.R. | CLIP-L + Qwen2.5-VL-7B | 72K–431K | 15.81 |
| GenS | Aria-25B-A3.9B | 315K | 50.60 |

Table 9: **Average per-sample inference latency on Video-MME.** Latency is measured on a single NVIDIA H100 GPU with $K = 32$ and 1 FPS candidate sampling.

| Method | Latency (s/sample) |
|---|---|
| BOLT | 1.29 |
| Ours | 1.38 |
| Q-Frame | 1.53 |
| GenS | 214.6 |

**Evaluation Prompt Template**

```
<Selected Frames>
You are a helpful assistant.
<Questions>
<Options>
Answer with the option's letter from the given choices directly.
```

# D Dataset Details

We utilize the GenS-Video-150K [64] dataset for training. The dataset annotation pipeline leverages GPT-4o to construct grounded video QAs and score frame relevance from 0 (irrelevant) to 5 (most relevant). To ensure comprehensive coverage of video understanding capabilities, the dataset includes 12 specific question types categorized into three primary domains: (1) **Reasoning Tasks**, which test the model's ability to infer logical relationships involving objects, actions, spatial configurations, and temporal changes; (2) **Perception Tasks**, focusing on the basic visual understanding of scenes, attributes, and spatial layouts; and (3) **Specialized Tasks**, which assess fine-grained capabilities including video detail referring, counting, OCR, and temporal sequence perception.

# E More Experimental Results

In this section, we provide additional ablation studies and sensitivity analyses on **Qwen3-VL-8B** and **InternVL-3.5-8B**. These experiments further validate the effectiveness and generalizability of our proposed Bayesian GRPO framework across different model architectures.

## E.1 Ablation on Training Components on SOTA Models

We investigate the contribution of each training stage on Qwen3-VL and InternVL-3.5. Table 10 and Table 11 demonstrates that while the Label Prior (Stage I) provides a reasonable initialization, the Bayesian GRPO update (Stage II) with NLL-based Task Evidence is essential for maximizing performance.

Table 10: **Extended Ablation Study on Qwen3-VL-8B.** Comparison of training strategies across different frame budgets ($K = 8, 32$).

| Method | Training Components | | | K = 8 | | | K = 32 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Stage I | Stage II | Reward | V-MME | LVB | MLVU | V-MME | LVB | MLVU |
| SigLIP2 | ✗ | ✗ | - | 55.4 | 56.7 | 61.0 | 62.1 | 59.8 | 69.2 |
| Label Prior | ✓ | ✗ | - | 57.3 | 58.1 | 64.3 | 63.9 | 61.4 | 71.0 |
| + Label Reward | ✓ | ✓ | Label | 57.1 | 57.4 | 64.9 | 64.1 | 61.0 | 71.4 |
| **+ Task Evidence** | ✓ | ✓ | **NLL** | **58.3** | **58.2** | **66.7** | **65.3** | **62.7** | **72.7** |

Table 11: **Extended Ablation Study on InternVL-3.5-8B.** Comparison of training strategies across different frame budgets ($K = 8, 32$).

| Method | Training Components | | | K = 8 | | | K = 32 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Stage I | Stage II | Reward | V-MME | LVB | MLVU | V-MME | LVB | MLVU |
| SigLIP2 | ✗ | ✗ | - | 56.7 | 58.1 | 64.8 | 60.9 | 60.6 | 69.7 |
| Label Prior | ✓ | ✗ | - | 57.6 | 58.9 | 66.2 | 62.4 | 62.3 | 70.7 |
| + Label Reward | ✓ | ✓ | Label | 57.3 | 58.1 | 66.5 | 62.8 | 62.7 | 70.3 |
| **+ Task Evidence** | ✓ | ✓ | **NLL** | **58.3** | **59.6** | **67.6** | **64.1** | **63.4** | **71.4** |

### E.2 Sensitivity Analysis: Frame Budget on SOTA Models

We analyze the impact of the frame budget $K$ on Qwen2.5-VL, Qwen3-VL and InternVL-3.5, both models exhibit a consistent trend: performance improves rapidly as $K$ increases from 8 to 32 and saturates around $K = 64$. This confirms that our selector effectively identifies the most critical evidentiary frames within a constrained budget, regardless of the downstream MLLM architecture.

Table 13: Performance with varying frame budgets ($K$) on **Qwen3-VL-8B**.

| Metric | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|
| Video-MME | 58.3 | 59.7 | 65.3 | 65.6 | 67.2 |
| LVB | 58.2 | 59.2 | 62.7 | 63.7 | 63.9 |
| MLVU | 66.7 | 70.0 | 72.7 | 76.3 | 77.7 |

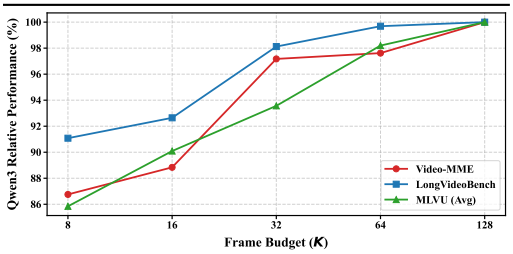


Figure 2: Relative performance trend on **Qwen3-VL-8B**.

Table 14: Performance with varying frame budgets ($K$) on **InternVL3.5-8B**.

| Metric | 8 | 16 | 32 | 64 | 128 |
|---|---|---|---|---|---|
| Video-MME | 58.3 | 59.6 | 64.1 | 63.7 | 64.5 |
| LVB | 59.6 | 60.7 | 63.4 | 62.7 | 63.3 |
| MLVU | 67.6 | 69.8 | 71.4 | 73.1 | 72.3 |

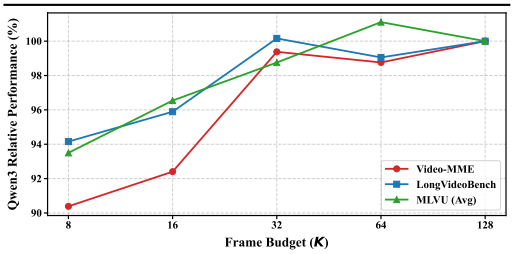


Figure 3: Relative performance trend on **InternVL-3.5-8B**.

## F More Visualization Results

We provide additional visualization cases in Figure 5, Figure 4. These examples further illustrate how our Bayesian GRPO selector effectively identifies evidentiary frames that align with the reasoning requirements of the MLLM, contrasting with the static label priors.

Table 12: Performance with varying frame budgets ($K$) on Video-MME, LongVideoBench, and MLVU. All results are evaluated using **Qwen2.5-VL**.

| **Metric** | $K = 8$ | $K = 16$ | $K = 32$ | $K = 64$ | $K = 128$ |
|---|---|---|---|---|---|
| Video-MME | 61.5 | 63.7 | 65.2 | 66.2 | 66.9 |
| LVB | 58.2 | 59.8 | 61.1 | 62.4 | 63.3 |
| MLVU (Avg) | 66.1 | 68.7 | 70.3 | 73.3 | 73.4 |

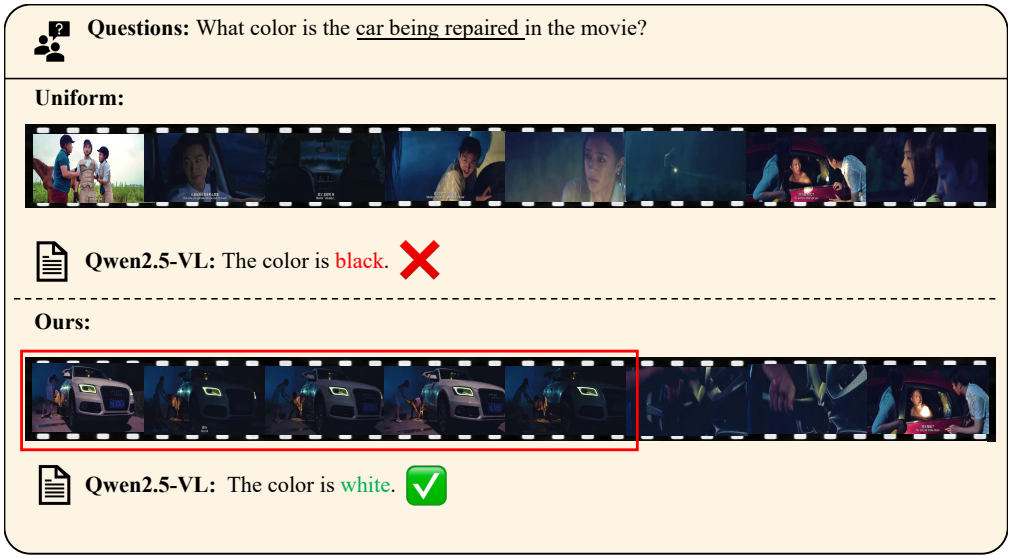


Figure 4: **Qualitative Examples 1 of Frame Selection.** From top to bottom, we visualize the frames selected by our method compared to the baseline.

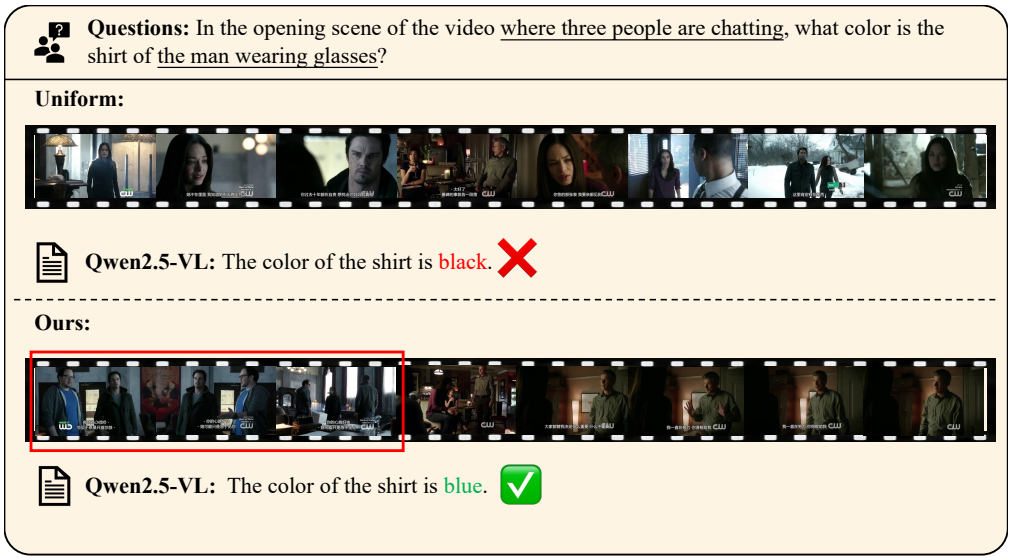


Figure 5: **Qualitative Examples 2 and 3 of Frame Selection.** From top to bottom, we visualize the frames selected by our method compared to the baseline.

## 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 abstract and introduction summarize the main method.

   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: In conclusion we discuss the limitations and future work.

   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: [Yes]

Justification: All assumptions and proofs are stated in Section 3.

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: Model design and training setup are specified in main paper and appendix for reproducibility.

   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: [No]

Justification: The code will be released after acceptance, the reproduction instructions is detailed in appendix, and the experiments use publicly available datasets.

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: Training and test details, hyperparameters are provided in main paper and appendix.

   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: [Yes]

   Justification: Results are reported over multiple runs.

   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: Compute resources are reported in main paper and appendix.

   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: All experiments use public datasets and follow the NeurIPS Code of Ethics.

   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: Potential impacts are discussed in conclusion.

    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: No high-risk generative model, scraped sensitive dataset, or dual-use asset.

    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: Existing datasets, baselines, and software assets are cited with access information where available.

    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: [Yes]

    Justification: Code will be released. No new dataset or benchmark asset is introduced.

    Guidelines:

    - 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: No new crowdsourcing or human-subject experiments are conducted.

    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: Experiments rely only on previously released public datasets.

    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: We use several MLLMs as RL environment to train and evaluate our frame selection method. Implementation details are comprehensively described in main paper and appendix.

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.