MaxInfo Fast and Slow Variants: Computational Overhead and Latency in Multimodal Video Processing

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the computational overhead of MaxInfo's Fast vs. Slow variants compared to uniform sampling when processing videos of varying lengths (1-60 minutes) on GPUs, and how does this trade-off. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adaptive Keyframe Sampling for Long Video Understanding. Research question: What is the computational overhead of MaxInfo's Fast vs. Slow variants compared to uniform sampling when processing videos of varying lengths (1-60 minutes) on GPUs, and how does this trade-off affect inference latency in multimodal models like Gemini 1.5 Pro?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

15 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 4.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
AKS improves video QA accuracy beyond strong baselines upon selecting informative keyframes.	✓	0.25
The study reveals the importance of information pre-filtering in video-based MLLMs.	✓	0.28
AKS brings consistent accuracy gain over three baselines, e.g., upon Qwen2VL, the improvement is 5.0% on LongVideoBench	×	0.06
With AKS, LLaVA-Video-7B reports 62.7% on LongVideoBench, which is 0.8% higher than the LLaVA-Video-72B model without AK	×	0.10
VideoMME contains many questions that require a high-level comprehension of the video content in which uniform sampling	×	0.08
AKS still finds more informative frames and improves the accuracy on VideoMME, although the gain is smaller than that on	×	0.04
GPT-4V reports 61.3% on LongVideoBench and 59.9% on VideoMME with 256 input frames.	×	0.03
GPT-4o reports 66.7% on LongVideoBench and 71.9% on VideoMME with 256 input frames.	×	0.04
Gemini-1.5-Flash reports 61.6% on LongVideoBench and 70.3% on VideoMME with 256 input frames.	×	0.05
Gemini-1.5-Pro reports 64.0% on LongVideoBench and 75.0% on VideoMME with 256 input frames.	×	0.06
LLaVA-Video-7B with AKS reports 62.7% on LongVideoBench and 65.3% on VideoMME.	×	0.09
UNI sampling strategy reports 58.9% on LongVideoBench and 64.4% on VideoMME.	×	0.02
TOP sampling strategy reports 62.4% on LongVideoBench and 63.7% on VideoMME.	×	0.03
BIN sampling strategy reports 60.2% on LongVideoBench and 65.2% on VideoMME.	×	0.03
ADA sampling strategy reports 62.7% on LongVideoBench and 65.3% on VideoMME.	×	0.03

References

- `http://arxiv.org/abs/2502.21271v1`

- http://arxiv.org/abs/2502.03183v3
- http://arxiv.org/abs/2403.05530v5