Multimodal Model Performance on Long-Form Video-Language Inputs: Accuracy and Cost Trade-offs

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How do multimodal models like Gemini 1.5 Pro compare to prior models in terms of accuracy and computational cost when processing interleaved video-language inputs of varying lengths, particularly for. 7 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MaxInfo: A Training-Free Key-Frame Selection Method Using Maximum Volume for Enhanced Video Understanding. Research question: How do multimodal models like Gemini 1.5 Pro compare to prior models in terms of accuracy and computational cost when processing interleaved video-language inputs of varying lengths, particularly for videos exceeding 30 minutes?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

1 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.5/10.

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
MaxInfo is the first training-free method based on the maximum volume principle for key-frame selection in video underst	✓	0.35
MaxInfo is available in Fast and Slow versions and a Chunk-based version.	✓	0.21
MaxInfo selects and retains the most representative frames from a video by maximizing the geometric volume formed by sel	✓	0.27
MaxInfo achieves a 3.28% improvement on LongVideoBench and a 6.4% improvement on EgoSchema for LLaVA-Video-7B.	✓	0.27
MaxInfo boosts LongVideoBench performance by 3.47% on LLaVA-Video-72B and 3.44% on MiniCPM4.5.	✓	0.27
MaxInfo is simple to implement and works with existing VLLMs without the need for additional training and has very low l	✓	0.21
The code for MaxInfo is available at https://github.com/FusionBrainLab/MaxInfo.git.	✓	0.20

References

- https://doi.org/10.1109/wacv61042.2026.00695