Multimodal-Aligned Large Language Models Outperform Specialized Video Encoders in Cross-Domain Video Question Answering

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Do large language models with multimodal alignment exhibit superior cross-domain generalization on video question answering benchmarks compared to specialized video encoders. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PruneVid: Visual Token Pruning for Efficient Video Large Language Models. Research question: Do large language models with multimodal alignment exhibit superior cross-domain generalization on video question answering benchmarks compared to specialized video encoders?.

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 4.3/10.

3 Results

15 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 4.3/10.

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
PLLaVA with PruneVid achieves 47.6% accuracy with 16.2% visual tokens and 0.23$\times$ speedup.	×	0.05
ST-LLM with PruneVid achieves 54.3% accuracy with 15.1% visual tokens and 0.26$\times$ speedup.	×	0.05
LLaVA-OneVision with PruneVid achieves 57.5% accuracy with 17.0% visual tokens and 0.20$\times$ speedup.	×	0.04
PruneVid is designed to efficiently process video data by minimizing redundancy in visual tokens before inputting them i	×	0.12
In the pre-filling stage, the model processes the input question tokens and visual tokens to construct the initial repre	×	0.04
The combined input sequence X is formed by concatenating the question tokens and the merged visual tokens.	×	0.07
The model employs a Transformer architecture with L layers.	×	0.04
The attention scores are computed using scaled dot-product attention with causal masking to prevent attending to future	×	0.04

References

- http://arxiv.org/abs/2410.09380v1
- http://arxiv.org/abs/2510.17722v2
- http://arxiv.org/abs/2412.16117v1