MELTR Integration with Multimodal Models for Zero-Shot Action Recognition

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the integration of MELTR with multimodal foundation models (e.g., CLIP or Flamingo) affect cross-modal zero-shot action recognition performance on combined video-text benchmarks like. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Language as the Medium: Multimodal Video Classification through text only. Research question: How does the integration of MELTR with multimodal foundation models (e.g., CLIP or Flamingo) affect cross-modal zero-shot action recognition performance on combined video-text benchmarks like Something-Something V2?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

12 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The language model is able to benefit from additional audio information in classifying videos on the UCF-101 test and th	×	0.07
Claude-instant-1 obtains on average the highest accuracy among the three large language models tested.	×	0.07
Both GPT3.5-turbo and Claude-instant-1 outperform Llama2 in interpreting visual and auditory information.	×	0.06
GPT3.5 and Claude-1 benefit from 'seeing' more frame captions, while Llama2's performance is negatively affected by more	×	0.03
The top-1 accuracy for BLIP2(FlanT5-XXL)+Llama2-13B is 49.56%.	×	0.01
The top-1 accuracy for BLIP2(FlanT5-XXL)+GPT3.5 is 66.37%.	×	0.01
The top-1 accuracy for BLIP2(FlanT5-XXL)+Claude-1 is 63.01%.	×	0.01
The top-3 accuracy for BLIP2(FlanT5-XXL)+Llama2-13B is 56.70%.	×	0.01
The top-3 accuracy for BLIP2(FlanT5-XXL)+GPT3.5 is 79.27%.	×	0.01
The top-3 accuracy for BLIP2(FlanT5-XXL)+Claude-1 is 81.49%.	×	0.01
The top-5 accuracy for BLIP2(FlanT5-XXL)+Llama2-13B is 58.51%.	×	0.01
The top-5 accuracy for BLIP2(FlanT5-XXL)+GPT3.5 is 82.04%.	×	0.01
The top-5 accuracy for BLIP2(FlanT5-XXL)+Claude-1 is 85.35%.	×	0.01
The method combines a 'perception' module and a 'reasoning' module to process multimodal textual descriptors and identif	×	0.09

References

- http://arxiv.org/abs/2408.14964v1
- http://arxiv.org/abs/2309.10783v1
- http://arxiv.org/abs/2605.01165v1