MELTR Integration and Zero-Shot Video Captioning Accuracy on ActivityNet Captions

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: What is the impact of MELTR integration on zero-shot video captioning accuracy for Flamingo compared to baseline CLIP on ActivityNet Captions. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Video Is Worth 4096 Tokens: Verbalize Videos To Understand Them In Zero Shot. Research question: What is the impact of MELTR integration on zero-shot video captioning accuracy for Flamingo compared to baseline CLIP on ActivityNet Captions?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

8 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
The proposed method converts long videos into coherent textual stories by verbalizing keyframes, audio, and text-overlai	×	0.06
The proposed method outperforms state-of-the-art story generation methods in experimental tests.	×	0.07
Experiments were conducted on five benchmark datasets covering fifteen video understanding tasks.	×	0.13
The proposed method achieves better results than fine-tuned video understanding baseline models without using any human-	×	0.09
The proposed method achieves better results than zero-shot video understanding baseline models without using any human-a	×	0.13
In the Zero-shot GPT-3.5 Generative benchmark, the proposed framework using Vicuna achieved a score of 17.4 compared to	×	0.03
The Zero-Shot approach using GPT-3.5 generated stories and a GPT-3.5 classifier achieved a score of 64.1 on a specific m	×	0.03
The Zero-Shot approach using GPT-3.5 generated stories and a GPT-3.5 classifier achieved a score of 98.89 on a specific	×	0.03
The system architecture includes modules for OCR, Automatic Speech Recognition (Closed Captions), Blip 2, Emotion Classi	×	0.07
The method utilizes WikiData to retrieve company name and business information related to the video content.	×	0.02

References

- http://arxiv.org/abs/1806.08854v1
- http://arxiv.org/abs/2305.09758v3
- http://arxiv.org/abs/1907.05092v1