Scaling Text Generation Modules in MELTR for Zero-Shot Video Classification Performance

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of scaling the text generation module of MELTR (e.g., using GPT-3.5 vs. GPT-4) on downstream zero-shot video classification performance on ActivityNet, measured by top-1/top-5. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition?. Research question: What is the impact of scaling the text generation module of MELTR (e.g., using GPT-3.5 vs. GPT-4) on downstream zero-shot video classification performance on ActivityNet, measured by top-1/top-5 accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

10 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 3.8/10.

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 study evaluates 16 visual datasets across images, videos, and point clouds.	✓	0.17
The evaluation employs the widely recognized validation sets for the benchmarks.	×	0.02
GPT-4 API version gpt-4-1106-preview was used to generate descriptive sentences.	×	0.04
The default number of descriptive sentences (K) generated per category is 20.	×	0.03
The GPT-4V API version gpt-4-vision-preview was employed in the study.	×	0.04
Kinetics-400 contains 400 classes.	×	0.00
Sth-Sth V1 contains 174 classes and 19,796 validation samples.	×	0.01
CLIP ViT-B/32 has 88M parameters.	×	0.06
CLIP ViT-B/16 has 86M parameters.	×	0.08
CLIP ViT-L/14 has 304M parameters.	×	0.07
EVA ViT-E/14 has 4.4B parameters.	×	0.04
GPT-4V achieved a score of 57.7 / 83.3 on the first reported benchmark metric in Table (p6).	×	0.04
GPT-4V achieved a score of 68.7 / 93.8 on the second reported benchmark metric in Table (p6).	×	0.03
GPT-4V achieved a score of 63.1 / 78.2 on the third reported benchmark metric in Table (p6).	×	0.04
RAF-DB is used as a dataset for 7-class facial expression recognition.	×	0.03
HMDB-51 is used as a dataset for 51-class action recognition.	×	0.08

References

- http://arxiv.org/abs/2311.15732v2

- http://arxiv.org/abs/2303.13375v2
- http://arxiv.org/abs/2305.12477v2