Language-Video Attention Ablation Impact on ActivityNet Text-Video Retrieval Performance

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the effect of removing the language-video attention module on inference throughput and recall@1 metrics for text-video retrieval on the ActivityNet Captions benchmark. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Text Is MASS: Modeling as Stochastic Embedding for Text-Video Retrieval. Research question: What is the effect of removing the language-video attention module on inference throughput and recall@1 metrics for text-video retrieval on the ActivityNet Captions benchmark?.

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 3.7/10.

3 Results

10 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 enables better text embedding ts corresponding to the semantics of the raw text t.	×	0.10
The performance tends to be stable from M = 10 to M = 20.	×	0.03
The proposed stochastic text embedding allows a lower similarity for irrelevant pairs and enables lower cross entropy lo	×	0.14
Using stochastic embedding ts gives a better result than t (red curve is lower).	×	0.06
The proposed ts enables lower entropy.	×	0.01
T-MASS achieves R@1 of 50.2, R@5 of 75.3, R@10 of 85.1, MdR of 1.0, and MnR of 11.9 on MSRVTT Retrieval.	×	0.07
T-MASS achieves R@1 of 28.9, R@5 of 48.2, R@10 of 57.6, MdR of 6.0, and MnR of 43.3 on LSMDC Retrieval.	×	0.07

References

- http://arxiv.org/abs/2511.15201v2
- http://arxiv.org/abs/2403.17998v1
- http://arxiv.org/abs/2203.15086v1