Non-Autoregressive vs. Autoregressive Multimodal Models: Inference Speed on MSR-VTT and MSVD

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 6 peer-reviewed papers addressing the following research question: How does the inference speed of non-autoregressive multimodal models compare to autoregressive models on the MSR-VTT and MSVD benchmarks when evaluated using frames-per-second (FPS) or latency metrics. It is encouraged to see that progress has been made to bridge videos and natural language. However, mainstream video captioning methods suffer from slow inference speed due to the sequential manner of autoregressive decoding, and prefer generating generic descriptions due to the. 0 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: Non-Autoregressive Coarse-to-Fine Video Captioning. Research question: How does the inference speed of non-autoregressive multimodal models compare to autoregressive models on the MSR-VTT and MSVD benchmarks when evaluated using frames-per-second (FPS) or latency metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 6 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

6 papers retrieved. 0 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.

References

- `http://arxiv.org/abs/2108.02359v2`
- `http://arxiv.org/abs/1911.12018v6`
- `http://arxiv.org/abs/2604.08084v1`