Robustness of RLAIF vs. Supervised Fine-Tuning in Multimodal Video Captioning

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the robustness of RLAIF-trained multimodal models compare to SFT baselines on out-of-domain video captioning benchmarks like MSR-VTT versus in-domain MSVD. 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. 10 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. 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 robustness of RLAIF-trained multimodal models compare to SFT baselines on out-of-domain video captioning benchmarks like MSR-VTT versus in-domain MSVD?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

4 papers retrieved. 10 claims extracted; 3 independently verified. Quality review score: 5.3/10.

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
Non-autoregressive decoding generates words in parallel to achieve significant inference speedup.	✓	0.18
Fine-grained descriptions are generally long.	×	0.10
Visual words are visually-grounded and highly associated with semantic correctness.	×	0.09
Treating visual and non-visual words equally causes insufficient training of meaningful words.	×	0.12
The proposed NACF model employs a bi-directional self-attention based network as its language model.	✓	0.19
The NACF model is trained with masked language modeling objective.	×	0.03
The NACF model generates visual words first to form a coarse-grained sentence template.	✓	0.19
The NACF model achieves 42.0 BLEU-4 score.	×	0.03
The NACF model achieves 28.7 METEOR score.	×	0.07
The NACF model achieves 51.4 ROUGE-L score.	×	0.03

References

- http://arxiv.org/abs/1806.08854v1
- http://arxiv.org/abs/1907.05092v1
- http://arxiv.org/abs/1911.12018v6