Oracle-RLAIF vs. SFT for Video Captioning on MSVD Across Model Scales

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does Oracle-RLAIF training compare to SFT in terms of video captioning accuracy and inference latency on the MSVD benchmark across 1B, 7B, and 13B parameter models. Recent advancements in large language models have influenced the development of video large multimodal models (VLMMs). The previous approaches for VLMMs involved Supervised Fine-Tuning (SFT) with instruction-tuned datasets, integrating LLM with visual encoders, and adding. 6 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: Tuning Large Multimodal Models for Videos using Reinforcement Learning from AI Feedback. Research question: How does Oracle-RLAIF training compare to SFT in terms of video captioning accuracy and inference latency on the MSVD benchmark across 1B, 7B, and 13B parameter models?.

2 Methodology

Systematic literature search across multiple databases yielded 11 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

11 papers retrieved. 6 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
Multimodal SFT models often produce responses that are not temporally and visually grounded to the video input.	×	0.06
The proposed method uses the VLMM to supervise itself by providing self-preference feedback of generated responses using	×	0.12
The proposed method facilitates the alignment of video and text modalities.	×	0.15
In the illustrated example, Output A received a reward score of 0.8 while Output B received a score of 0.3.	×	0.02
The preference data indicates Output A is preferred over Output B (A $>$ B) based on context-aware preferences.	×	0.07
The VLM-SFT training is initiated building on a pre-trained image-text model.	×	0.05

References

- http://arxiv.org/abs/2510.27364v1
- http://arxiv.org/abs/2510.02561v1
- http://arxiv.org/abs/2402.03746v3