Synthetic vs. Human Feedback Quality Effects on Oracle-RLAIF CIDEr Gains in MSVD

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the impact of varying the quality of AI feedback (e.g., synthetic vs. human-annotated rewards) on the CIDEr score improvement of Oracle-RLAIF on the MSVD benchmark for models with 7B, 13B,. Recent advances in large video-language models (VLMs) rely on extensive fine-tuning techniques that strengthen alignment between textual and visual comprehension. Leading pipelines typically pair supervised fine-tuning (SFT) with reinforcement learning from preference data to. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Oracle-RLAIF: An Improved Fine-Tuning Framework for Multi-modal Video Models through Reinforcement Learning from Ranking Feedback. Research question: What is the impact of varying the quality of AI feedback (e.g., synthetic vs. human-annotated rewards) on the CIDEr score improvement of Oracle-RLAIF on the MSVD benchmark for models with 7B, 13B, and 30B parameters?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.0/10.

3 Results

16 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 5.0/10.

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
Oracle-RLAIF outperforms VLM-RLAIF on MSVD-QA with an accuracy of 72.9 compared to 68.5.	×	0.05
Oracle-RLAIF achieves a score of 3.9 on MSVD-QA, improving upon VLM-RLAIF's score of 3.6.	×	0.05
On MSRVTT-QA, Oracle-RLAIF shows an accuracy improvement of 5.0% over VLM-RLAIF.	×	0.05
Oracle-RLAIF achieves a score of 3.7 on MSRVTT-QA, which is higher than VLM-RLAIF's score of 3.1.	×	0.05
On ActivityNet-QA, Oracle-RLAIF shows a 2.0% increase in accuracy compared to VLM-RLAIF.	×	0.05
The original VLM-RLAIF publication reports better results on ActivityNet due to the use of caption data in reward model	×	0.05
Oracle-RLAIF uses the same SFT model checkpoint (VLM-SFT 7B) as VLM-RLAIF for initialization.	×	0.06
VLM-RLAIF was trained for 4 epochs with a rollout batch size of 64, differing from the original configuration.	×	0.03
Oracle ranker reward model omits caption data during training unlike VLM-RLAIF's reward model.	×	0.10
Both Oracle-RLAIF and VLM-RLAIF were trained using 4$\times$NVIDIA H100 80GB GPUs with QLoRA.	×	0.06

References

- http://arxiv.org/abs/2604.25872v1
- http://arxiv.org/abs/2510.02561v1
- http://arxiv.org/abs/1911.12018v6