Oracle-RLAIF Sample Efficiency vs. Supervised Fine-Tuning on RxR-CE nDTW Metrics

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does Oracle-RLAIF's sample efficiency compare to traditional supervised fine-tuning when evaluated on the RxR-CE benchmark's nDTW score across different training compute budgets. 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. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: How does Oracle-RLAIF's sample efficiency compare to traditional supervised fine-tuning when evaluated on the RxR-CE benchmark's nDTW score across different training compute budgets?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

12 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 improves upon leading fine-tuning frameworks for VLMs, specifically the current SOTA VLM-RLAIF.	✓	0.15
Both Oracle-RLAIF and VLM-RLAIF are trained using the same initial SFT policy model (VLM-SFT 7B checkpoint).	×	0.08
VLM-RLAIF is trained with a pre-trained reward model for 4 epochs and a rollout batch size of 64.	×	0.08
Oracle-RLAIF uses the same training configurations as VLM-RLAIF but omits caption data for training the Oracle ranker re	×	0.10
Both models are trained using 4$\times$NVIDIA H100 80GB GPUs with Quantized Low-Rank Adapter (QLoRA).	×	0.04
Oracle-RLAIF outperforms all baselines in video-question answering performance across MSVD, MSRVTT, and ActivityNet data	×	0.09
Oracle-RLAIF achieves a 4.4% increase in accuracy and a 0.3 increase in score compared to VLM-RLAIF on the MSVD-QA datas	×	0.05
Oracle-RLAIF achieves a 5.0% increase in accuracy and a 0.6 increase in score compared to VLM-RLAIF on the MSRVTT-QA dat	×	0.05
Oracle-RLAIF achieves a 2.0% increase in accuracy and a 0.1 increase in score compared to VLM-RLAIF on the ActivityNet-Q	×	0.05

References

- http://arxiv.org/abs/2208.07204v2
- http://arxiv.org/abs/2110.06500v2
- http://arxiv.org/abs/2510.02561v1