Multimodal Contrastive Learning for Robust Vulnerability Detection in Adversarial Code-Comment Pairs

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does multimodal contrastive learning (as proposed in MultiVul) improve the robustness of vulnerability detection models like Llama3 and Deepseek R1 when tested on adversarially. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Contrastive Video Representation Learning via Adversarial Perturbations. Research question: To what extent does multimodal contrastive learning (as proposed in MultiVul) improve the robustness of vulnerability detection models like Llama3 and Deepseek R1 when tested on adversarially perturbed code-comment pairs from the Big-Vul dataset, measured by F1-score degradation under noise injection?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

13 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The optimization produces useful representations in about 50 iterations and takes about 5 milli-seconds per frame on a s	×	0.03
The slack regularization constant C is set to 1.	×	0.02
HMDB-51 consists of 6766 Internet videos over 51 classes; each video is about 20 – 1000 frames.	×	0.03
The standard evaluation protocol for HMDB-51 reports average classification accuracy on three-folds.	×	0.07
For HMDB-51, features from the pool5 layer of each stream are sequences of 2048D vectors.	×	0.03
NTU-RGBD has 56,880 video sequences across 60 classes, 40 subjects, and 80 views.	×	0.03
Videos in NTU-RGBD have on average 70 frames and consist of people performing various actions; each frame is annotated f	×	0.01
For NTU-RGBD, two evaluation protocols are used, namely cross-view and cross-subject evaluation.	×	0.05
For NTU-RGBD, 256D features from the bottleneck layer (before their global average pooling layer) are used as input to t	×	0.04
YUP++ dataset has 20 scene classes with 60 videos in each class.	×	0.01
Half of the sequences in each class of YUP++ are collected by a static camera and the rest are recorded by a moving came	×	0.02

References

- http://arxiv.org/abs/2504.07887v2

- http://arxiv.org/abs/1807.09380v3
- http://arxiv.org/abs/2502.13141v1