VLA-Adapter Robustness to Visual Perturbations in RoboBench Distraction Scenarios

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the robustness of VLA-Adapter to visual perturbations compare against other lightweight VLA architectures when evaluated on the RoboBench distraction scenarios. 14 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: Robustness Analysis of Video-Language Models Against Visual and Language Perturbations. Research question: How does the robustness of VLA-Adapter to visual perturbations compare against other lightweight VLA architectures when evaluated on the RoboBench distraction scenarios?.

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. 14 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
Experiments were performed on the YouCook2-P and MSRVTT-P benchmarks.	×	0.03
For the MSRVTT-P benchmark, zero-shot models typically achieve higher absolute and relative robustness scores compared t	×	0.05
For long, complex activities in the YouCook2-P benchmark, fine-tuned models are typically more relatively robust than ot	×	0.05
The FIT model was pre-trained on the HowTo100M dataset.	×	0.05
The FIT model has 180.9M parameters.	×	0.01
The COOT model has 7.6M parameters.	×	0.01
On the MSRVTT-P benchmark under AddText perturbation, the FIT (zs) model achieved a relative robustness score (γr) of 1.	×	0.03
On the YouCook2-P benchmark under ChangeChar perturbation, the COOT (scratch) model achieved a relative robustness score	×	0.04
The VideoClip model uses S3D as its video encoder.	×	0.03
The UniVL model uses BERT as its text encoder.	×	0.02
On the MSRVTT-P benchmark, the VideoClip (ft) model achieved an aggregated absolute robustness score (γa) of 0.94$\pm$0.05 f	×	0.07
On the YouCook2-P benchmark, the VideoClip (zs) model achieved an aggregated absolute robustness score (γa) of 0.97$\pm$0.02	×	0.07
The MIL NCE model uses Word2Vec for text input encoding.	×	0.05
The FIT model uses a ViT (Vision Transformer) as its video encoder.	×	0.03

References

- http://arxiv.org/abs/2207.02159v4
- http://arxiv.org/abs/2504.09480v1
- http://arxiv.org/abs/2506.05429v1