Multimodal Video-Language Adapter Robustness Under Noise and Occlusion on MSRVTT-P

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the robustness of multimodal video-language adapters compare to frozen backbones on the MSRVTT-P benchmark under varying levels of Gaussian noise and occlusion. 7 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multimodal Fusion and Vision-Language Models: A Survey for Robot Vision. Research question: How does the robustness of multimodal video-language adapters compare to frozen backbones on the MSRVTT-P benchmark under varying levels of Gaussian noise and occlusion?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 7 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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 transformer structure has been proposed to improve the applicability of different modal data and capture local featu	×	0.04
Adversarial representation learning has been used to create modality invariant embedding spaces, reduce modal gaps, and	×	0.06
Post fusion is a key method in multimodal analysis, which combines the results of decision level independent processing	×	0.09
Common techniques in post fusion include weighted averaging, voting mechanisms, and logical rules.	×	0.03
Roitberg et al. compared and analyzed seven decision-level fusion strategies for driver behavior understanding.	×	0.04
The encoder-decoder method efficiently represents scene semantics through encoding, interaction, and decoding.	×	0.04
Attention-based fusion has been used in multimodal fusion approaches for semantic scene understanding.	✓	0.20

References

- http://arxiv.org/abs/1909.03564v2
- http://arxiv.org/abs/2504.02477v3
- http://arxiv.org/abs/2508.11281v3