Attention-Based Multimodal Fusion vs. Encoder-Decoder Frameworks in Robotic Vision Under Occlusion

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do attention-based multimodal fusion architectures compare to encoder-decoder frameworks in semantic scene understanding accuracy on robot vision benchmarks under varying occlusion levels. 15 claims were extracted from source literature; 2 were 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 do attention-based multimodal fusion architectures compare to encoder-decoder frameworks in semantic scene understanding accuracy on robot vision benchmarks under varying occlusion levels?.

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 4.5/10.

3 Results

13 papers retrieved. 15 claims extracted; 2 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
Early fusion directly fuses data from different modalities before feature extraction.	×	0.05
Mid-term fusion combines modal features through mechanisms such as feature concatenation or weighting after extracting t	×	0.03
Late stage fusion is achieved by integrating the decision results of each modality after independent decision-making is	×	0.04
Transformer structures have been proposed to improve the applicability of different modal data and capture local feature	×	0.04
Adversarial representation learning is used to create modality invariant embedding spaces and reduce modal gaps.	×	0.03
Post fusion combines the results of decision level independent processing of modalities.	×	0.04
Common techniques in post fusion include weighted averaging, voting mechanisms, and logical rules.	×	0.03
Post fusion offers advantages such as strong modal independence, ease of individual optimization, and scalability of mul	×	0.03
Roitberg et al. compared and analyzed seven decision-level fusion strategies for driver behavior understanding.	×	0.04
Traditional multimodal fusion methods struggle with complex data compared to deep neural networks.	×	0.13
Deep neural networks have driven a shift from explicit to implicit fusion where network design inherently captures modal	×	0.04
The survey categorizes multimodal fusion approaches in semantic scene understanding into encoder-decoder frameworks, att	✓	0.20
The encoder-decoder method represents scene semantics through encoding, interaction, and decoding.	×	0.09
Various sensory inputs including RGB, Depth, LiDAR, GPS, and IMU are processed through multimodal fusion strategies to e	×	0.11
Fused features support core robotic vision tasks such as 3D semantic scene understanding, SLAM, 3D object detection, nav	✓	0.15

References

- http://arxiv.org/abs/2501.16273v2
- http://arxiv.org/abs/2504.02477v3
- http://arxiv.org/abs/2512.14020v1