# Multimodal vs. Single-Modality Models in Sim-to-Real Transfer Performance on RoboWatch

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: Can multimodal models (e.g., vision-language-action) trained on high-fidelity synthetic environments achieve comparable sim-to-real transfer performance to single-modality models, as measured by the. 0 claims were extracted from source literature; 0 were 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: A Survey of Deep Learning-Based Object Detection. Research question: Can multimodal models (e.g., vision-language-action) trained on high-fidelity synthetic environments achieve comparable sim-to-real transfer performance to single-modality models, as measured by the RoboWatch benchmark score?.

## 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. 0 claims extracted; 0 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.

## References


- https://doi.org/10.1109/access.2019.2939201
- https://doi.org/10.48550/arxiv.2309.07864
- https://openalex.org/W3163842339