Robo-MUTUAL Enhances RoboSuite Task Accuracy with Multimodal Proprioceptive Integration

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of adding proprioceptive robot state information to CLAM's multimodal input (video + audio + proprioception) on RoboSuite task accuracy compared to video-only inputs. 10 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Robo-MUTUAL: Robotic Multimodal Task Specification via Unimodal Learning. Research question: What is the impact of adding proprioceptive robot state information to CLAM's multimodal input (video + audio + proprioception) on RoboSuite task accuracy compared to video-only inputs?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

15 papers retrieved. 10 claims extracted; 4 independently verified. Quality review score: 5.8/10.

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
Robo-MUTUAL can successfully understand both visual and textual goals trained exclusively on visual goals.	×	0.05
Robo-MUTUAL-EPICK fails to transfer from visual to textual goals.	×	0.03
Robo-MUTUAL demonstrates a moderate performance drop when transferring from visual to textual goals.	×	0.03
Robo-MUTUAL is evaluated across more than 130 tasks and 4000 evaluations on both simulated LIBERO benchmark and real rob	✓	0.24
Robo-MUTUAL shows superior capabilities in overcoming data constraints in robotic learning.	✓	0.21
Robo-MUTUAL uses two Collapse and Corrupt operations to bridge the modality gap in the learned multimodal representation	✓	0.24
Robo-MUTUAL is pretrained using extensive out-of-domain data to endow the robot with strong Cross-modality Alignment cap	✓	0.27
GPT4-Synthetic is the main baseline in the common scenario that textual goals are missing.	×	0.01
Robo-MUTUAL-EPICK uses the DecisionNCE pretrained solely on the narrow EPICK-KITCHEN dataset as the robotic multimodal e	×	0.09
Robo-MUTUAL-EPICK may suffer from limited Cross-modality Alignment capability.	×	0.12

References

- http://arxiv.org/abs/2410.01529v1
- http://arxiv.org/abs/2507.02864v2

- http://arxiv.org/abs/1909.11639v3