Multimodal Latent Action Models vs. CLAM in LIBERO Success Rate Efficiency

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: Can multimodal latent action models outperform CLAM in success rates on LIBERO by incorporating additional sensory modalities (e.g., audio, depth) while maintaining computational efficiency. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CLAM: Continuous Latent Action Models for Robot Learning from Unlabeled Demonstrations. Research question: Can multimodal latent action models outperform CLAM in success rates on LIBERO by incorporating additional sensory modalities (e.g., audio, depth) while maintaining computational efficiency?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

9 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
CLAM outperforms all baselines and nearly matches the performance of BC with expert data in both state- and image-based	×	0.05
CLAM improves upon the best baseline VPT by more than 2$\times$ average normalized return on the DMControl (locomotion) tasks.	×	0.07
CLAM improves around 2-3$\times$ success rate on the MetaWorld (manipulation) tasks compared to the best baseline VPT.	×	0.12
Transformer-CLAM achieves performance close to or even better than that of BC-Expert which uses the same amount of privi	×	0.08
All variants of CLAM outperform the best baseline VPT.	×	0.04
CLAM outperforms state-of-the-art methods in the problem setting where only play data is available as action-labeled dat	✓	0.17
CLAM can leverage vast, unstructured observation data to learn latent actions in an unsupervised manner.	×	0.10
CLAM enables scalable learning from easy-to-collect, cheap play data avoiding the need for expensive task-specific data	×	0.05
The Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, a feedforward dimension of 2048, 8 attention heads,	×	0.02
The CALVIN Transformer-CLAM model uses 6 encoder layers, 6 decoder layers, a feedforward dimension of 2048, 8 attention	×	0.02
The MetaWorld environment has a maximum episode step of 100, state dimension of 39, action dimension of 4, image shape o	×	0.03
The CALVIN environment has a maximum episode step of 200, state dimension of 39, action dimension of 7, image shape of [	×	0.03
The evaluation environments in simulation include locomotion tasks from the DMControl benchmark (Hopper and HalfCheetah)	×	0.04

References

- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/1705.00470v2
- http://arxiv.org/abs/2507.19375v1