# Scaling Unlabeled Multimodal Demonstrations for CLAM in Robotic Policy Learning

Assignee Research

June 7, 2026

**Abstract**

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does scaling the size of unlabeled multimodal demonstration datasets (video + audio) affect CLAM's downstream policy accuracy for robotic tasks, compared to scaling labeled action datasets, using. 12 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: CLAM: Continuous Latent Action Models for Robot Learning from Unlabeled Demonstrations. Research question: How does scaling the size of unlabeled multimodal demonstration datasets (video + audio) affect CLAM's downstream policy accuracy for robotic tasks, compared to scaling labeled action datasets, using the success rate on the RoboBench benchmark as a metric?.

## 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. 12 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 |
|---|---|---|
| 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. | × | 0.08 |
| CLAM improves upon the best baseline VPT by around 2-3$\times$ success rate on the Meta-World (manipulation) tasks. | × | 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. | × | 0.05 |
| CLAM outperforms state-of-the-art methods in the problem setting where only play data is available as action-labeled dat | ✓ | 0.17 |
| CLAM scales with \|Dunlabeled\| while supervised IDMs only scale with \|Dlabeled\|. | × | 0.02 |
| VPT learns a suboptimal IDM, underscoring the benefit of latent action models which can leverage vast, unstructured obse | ✓ | 0.15 |
| CLAM enables scalable learning from easy-to-collect, cheap play data avoiding the need for expensive task-specific data | × | 0.05 |
| CLAM is evaluated on DMControl, Meta-World, and CALVIN without modification. | × | 0.03 |
| All domains are continuous control environments and we use a fixed episode length and no termination conditions. | × | 0.07 |
| For DMControl tasks, we report normalized return following [22]. | × | 0.02 |

## References


- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/2102.06386v3
- http://arxiv.org/abs/2010.14374v3