Contrastive and Generative Pre-Training Effects on Agent Confidence Calibration in CALVIN

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the comparative effect of contrastive versus generative pre-training objectives on confidence calibration scores for agents evaluated on CALVIN. 8 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: Distribution Estimation to Automate Transformation Policies for Self-Supervision. Research question: What is the comparative effect of contrastive versus generative pre-training objectives on confidence calibration scores for agents evaluated on CALVIN?.

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 4.4/10.

3 Results

15 papers retrieved. 8 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
The framework can estimate the distribution of visual transformations present in the dataset.	✓	0.23
The obtained histogram for Transformed MNIST is similar to the ground truth distribution.	×	0.06
Complementary pretext tasks lead to learning useful representations for downstream tasks.	×	0.07
The performance gap becomes bigger in the case of translation and scaling.	×	0.03
The pretext task based on scale-based transformations leads to poor performance in the SVHN dataset.	×	0.12
The proposed framework achieves a supervised accuracy of 93.22 on FMNIST, 94.75 on SVHN, 89.58 on CIFAR-10, and 64.00 on	×	0.03
The framework uses a reference subset that includes the most representative and frequent data for distribution estimatio	×	0.06
The transformation parameter distributions of Itrans and Ic are similar.	×	0.04

References

- http://arxiv.org/abs/2111.12265v1
- http://arxiv.org/abs/2412.14737v2
- http://arxiv.org/abs/1901.09960v5