Causal Structure-Aware Data Augmentation for Zero-Shot Multimodal Model Generalization

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does causal structure-aware data augmentation affect the zero-shot accuracy of multimodal foundation models when evaluated on shifted distribution datasets. 6 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning. Research question: How does causal structure-aware data augmentation affect the zero-shot accuracy of multimodal foundation models when evaluated on shifted distribution datasets?.

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 3.5/10.

3 Results

14 papers retrieved. 6 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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
The benchmark includes object classes such as CellPhone, Desktop, Laptop, Television, Bed, Bowl, Cloth, Cup, Mug, Pan, P	×	0.02
The benchmark categorizes actions into break, clean, class, close, dirty, action, open, turnoff, and turnon.	×	0.03
The benchmark evaluates performance on unseen objects and test data, categorized into compositional shift and systematic	×	0.03
The benchmark includes performance metrics such as accuracy for different seeds (1 through 10).	×	0.02
The average accuracy for dense, token-mean, and token-max methods is reported for seeds 1 through 10.	×	0.02
The benchmark includes results for both ID (in-distribution) and OOD (out-of-distribution) settings.	×	0.03

References

- http://arxiv.org/abs/2008.09301v1
- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2301.05169v2