Causally Augmented Fine-Tuning for Cross-Domain Robustness in Multimodal Foundation Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: To what extent does causally augmented fine-tuning improve the cross-domain alignment robustness of multimodal foundation models as measured by zero-shot classification scores. 8 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: Robust Fine-Tuning of Vision-Language Models for Domain Generalization. Research question: To what extent does causally augmented fine-tuning improve the cross-domain alignment robustness of multimodal foundation models as measured by zero-shot classification scores?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 8 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
Ensembling of weights from zero-shot and fine-tuned CLIP models improved both accuracy under distribution shifts and acc	×	0.12
Weight ensembling strategies in CLIP models fine-tuned using different hyper-parameters improved robustness and inferenc	×	0.05
Domain Prompt Learning automatically generated tailored text prompts for CLIP by estimating domain-specific features fro	×	0.05
Zero-shot CLIP outperforms a trained ResNet50-based logistic regression classifier on 16 lower-complexity datasets.	×	0.07
CLIP uses an InfoNCE loss function with temperature scaling (τ), popularized by van den Oord et al. and adapted for imag	×	0.04
Zero-shot classification performance of CLIP does not match that of a fine-tuned SoTA vision-only model under challengin	✓	0.15
Few-shot CLIP demonstrates superior performance over a few-shot vision-only model in limited data environments containin	✓	0.17
A fine-tuning strategy for CLIP that combines cross-entropy training and stochastic weight averaging improves out-of-dis	×	0.07

References

- http://arxiv.org/abs/2601.04110v2
- http://arxiv.org/abs/2306.06048v3
- http://arxiv.org/abs/2311.02236v1