Synthetic Multimodal Caption Diversity and Zero-Shot Vision-Language Model Stability

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does diversity in synthetic multimodal caption distributions affect the zero-shot classification stability of vision-language models across unseen visual domains. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Toward a Holistic Evaluation of Robustness in CLIP Models. Research question: To what extent does diversity in synthetic multimodal caption distributions affect the zero-shot classification stability of vision-language models across unseen visual domains?.

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 3.8/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

References

- http://arxiv.org/abs/2410.01534v2
- http://arxiv.org/abs/2504.09480v1
- http://arxiv.org/abs/2303.07771v3