Multimodal Caption Distribution Shifts and Vision-Language Model Calibration Error

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: To what extent does variation in multimodal caption distributions affect the confidence calibration error of vision-language models under domain shift. 11 claims were extracted from source literature; 2 were 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: Confidence-calibrated covariate shift correction for few-shot classification in Vision-Language Models. Research question: To what extent does variation in multimodal caption distributions affect the confidence calibration error of vision-language models under domain shift?.

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 4.4/10.

3 Results

16 papers retrieved. 11 claims extracted; 2 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
CalShift achieves a maximum accuracy improvement of 6.8% over CoOp in the zero-shot setting.	×	0.04
CalShift achieves a maximum accuracy improvement of 7.2% over CoOp in the 1-shot setting.	×	0.04
CalShift shows a 5.9% average accuracy improvement over CoOp.	×	0.03
The accuracy improvement percentage of CalShift over CoOp decreases monotonically from 7.2% in the 1-shot setting to 5.1	×	0.03
Integrating CMP with CoOp achieves up to a 12.5% improvement in calibration performance in the 4-shot setting compared t	×	0.03
Integrating CMP with CoOp achieves a 10.2% improvement in calibration performance in the 16-shot setting compared to van	×	0.03
CalShift achieves up to a 5.82% reduction in Expected Calibration Error (ECE).	✓	0.18
CalShift improves accuracy by 3.5% on challenging datasets impacted by covariate shifts.	✓	0.21
The experiments used pre-trained CLIP with ViT-B/16 as the backbone in the prompt learning method.	×	0.05
The covariate shift experiment used CLIP with ViT-B/16 and ResNet-50 backbones for the last four variants of ImageNet.	×	0.08
For the first four datasets, the model was trained on one domain and evaluated on a different test domain.	×	0.06

References

- http://arxiv.org/abs/2412.10372v1

- http://arxiv.org/abs/2502.07847v2
- http://arxiv.org/abs/2604.09529v1