Conformal Prediction for Reliable Zero-Shot Multimodal Model Evaluation Under Domain Shifts

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can conformal prediction methods improve the reliability of zero-shot capability evaluations for multimodal models across unseen domain shifts. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Inferring Latent Class Statistics from Text for Robust Visual Few-Shot Learning. Research question: Can conformal prediction methods improve the reliability of zero-shot capability evaluations for multimodal models across unseen domain shifts?.

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 4.1/10.

3 Results

14 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.1/10.

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 experiments use two base datasets: ImageNet and iNaturalist.	×	0.02
iNaturalist is a hierarchical dataset with fine-grained classes.	×	0.04
The test datasets include Caltech, EuroSAT, Food, Flowers, SUN397, DTD, Pets, Cars, and UCF101.	×	0.01
Visual and text features are extracted using the pre-trained CLIP ResNet50 trained on LAION400M.	×	0.05
The method aims to predict the mean and covariance of a class distribution in the feature space from text.	✓	0.23
The visual backbone is denoted by fv and the text encoder by ft.	×	0.03
Text contexts such as 'a photo of a {class}' or GPT3-generated visual descriptions are used.	×	0.05
The method employs two mapping networks, gμ (s, $\theta$$\mu$) and g$\Sigma$ (s, $\theta$$\Sigma$), for predicting the mean and covariance.	×	0.08
The baseline method achieves an average accuracy of 84.70% with 1 shot, 89.02% with 2 shots, 91.15% with 4 shots, 92.10%	×	0.02
Zero-Shot CLIP achieves an average accuracy of 60.43%.	×	0.05

References

- http://arxiv.org/abs/2508.19294v2
- http://arxiv.org/abs/2311.14544v1

- http://arxiv.org/abs/2307.03135v3