Text Data Augmentation Effects on CLIP and ALIGN Zero-Shot ImageNet Classification

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does text data augmentation impact the zero-shot image classification accuracy of CLIP compared to ALIGN on ImageNet variants. Contrastive Language-Image Pretraining (CLIP) has emerged as a powerful paradigm for cross-modal learning, using image-text pairs to achieve remarkable zero-shot classification performance. However, its calibration on noisy data has been less explored, especially in. 10 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Analyzing the Calibration of CLIP Models Under Noisy Data Conditions. Research question: How does text data augmentation impact the zero-shot image classification accuracy of CLIP compared to ALIGN on ImageNet variants?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

4 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.4/10.

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
CLIP uses image-text pairs to achieve zero-shot classification performance.	✓	0.15
The study evaluates CLIP's calibration using 10 noise types from the ImageNet-C dataset.	✓	0.18
The noise types used in the evaluation include natural noise, digital noise, weather noise, and blur noise.	✓	0.16
The study proposes test-time augmentation (TTA) to improve calibration by increasing prediction diversity and reducing o	✓	0.23
Approximately 600 simulations were conducted in this study.	×	0.05
In in-distribution settings on brightness noise, ViT-B/32 achieves 94.66% accuracy compared to ResNet50's 62.37% accurac	✓	0.19
In in-distribution settings on brightness noise, ViT-B/32 achieves 0.59% Expected Calibration Error (ECE) compared to Re	✓	0.24
In out-of-domain (OOD) situations on defocus noise (fine-tuned on glass), ResNet50 achieves 3.14% ECE while ViT-B/32 ach	✓	0.27
Applying Test-Time Augmentation (TTA) reduces ViT-B/32's ECE to 13.07% on defocus noise when fine-tuned on glass.	✓	0.27
The code for this study is available at https://github.com/AIPMLab/CLIPSimulationsGao.	✓	0.16

References

- https://www.semanticscholar.org/paper/c36e20d972a21fe4ac3c8f4dbbbbb2db4e1a0c1e
- https://www.semanticscholar.org/paper/cf14839be008fc5761c61fd818b89f5d42ad82da

- https://arxiv.org/abs/2508.20760