Manifold-Aware Defense Mechanisms in Vision-Language Models: Performance and Throughput Trade-offs

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: Does integrating manifold-aware defense mechanisms into vision-language models degrade inference throughput or zero-shot classification performance on standard benchmarks like ImageNet and MSCOCO. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Enhancing CLIP Robustness via Cross-Modality Alignment. Research question: Does integrating manifold-aware defense mechanisms into vision-language models degrade inference throughput or zero-shot classification performance on standard benchmarks like ImageNet and MSCOCO?.

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 3.2/10.

3 Results

4 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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 study evaluates methods assuming full access to model weights and gradients by the attacker.	×	0.01
Table 1 reports classification accuracy on clean and adversarially perturbed images across 9 diverse datasets.	×	0.06
TeCoA, PMG, and FARE are identified as fine-tuning-based methods.	×	0.06
Table 2 presents classification accuracy (%) on ImageNet and its variants datasets.	×	0.03
Table 4 presents classification accuracy (%) on ImageNet and its variants datasets under PGD attacks.	×	0.04
In the proof of Cosine Similarity Distortion Bound, clean feature vectors x1 and x2 are defined with unit norm (xi= 1)	×	0.01
Adversarial features are denoted as the sum of clean features and a perturbation vector with norm bounded by epsilon.	×	0.05
When the perturbation contains a non-zero orthogonal component, the ratio of projected cosine distortion to full cosine	×	0.02
Figure 2 displays accuracy comparisons across 10 image augmentation methods on ImageNet.	×	0.02
Figure 3 displays accuracy comparisons across varying numbers of Principal Components.	×	0.01
The experiment sets the class name augmentation parameter C to 256.	×	0.01

References

- https://www.semanticscholar.org/paper/8d7f1a07420f9a296d8199f1e372c3a068063556

- https://arxiv.org/abs/2503.03613
- https://arxiv.org/abs/2510.24038