Impact of Feature-Level Adversarial Perturbations on CLIP Robustness in Retrieval Tasks

Assignee Research

June 11, 2026

Abstract

Contrastive Language-Image Pre-training (CLIP) models have shown significant potential, particularly in zero-shot classification across diverse distribution shifts. Building on existing evaluations of overall classification robustness, this work aims to provide a more comprehensive assessment of CLIP by introducing several new perspectives. First, we investigate their robustness to variations in specific visual factors. Second, we assess two critical safety objectives–confidence uncertainty and out-of-distribution detection–beyond mere classification accuracy. Third, we evaluate the finesse

1 Introduction

This paper examines: Toward a Holistic Evaluation of Robustness in CLIP Models. Research question: What is the impact of feature-level adversarial perturbations during multimodal pretraining on the robustness of CLIP models against common corruptions in retrieval tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.0/10.

3 Results

10 papers retrieved. 28 claims extracted; 17 independently verified. Quality review score: 7.0/10.

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
CLIPs outperform ImageNet models on six of ten visual factors (e.g., shape, texture, and size).	✓	0.16
Training distribution impacts visual factor robustness of CLIP.	✓	0.21
Zero-shot CLIP is competitive with other models in novelty detection ability (e.g., NINCO).	✓	0.18
Training distribution impacts detection accuracy in OOD detection.	✓	0.18
Fine-tuning raises calibration error in CLIP models.	×	0.14
LP-FT and WiSE-FT recover calibration error with temperature scaling, while FLYP remains over-confident.	✓	0.20
Both data distribution and quantity play key roles in calibration of CLIP models.	✓	0.18
CLIP's Zero-shot retrieval aligns with classification accuracy.	✓	0.17
Data distribution and augmentation shape retrieval quality in CLIP models.	✓	0.18
CNN-based CLIPs are more stable than ViT-based CLIPs in terms of robustness to 3D corruptions and correspondence matchin	✓	0.17
LLaVA outperforms CLIP on ambiguous sets in vision–language interaction.	✓	0.15
Stronger LLMs (Vicuna > Mistral) amplify gains in vision–language interaction.	✓	0.17
No training paradigm dominates in terms of robustness among CLIP, BLIP, SigLIP, ViTamin.	✓	0.18
ViTamin is more robust to 3D corruptions compared to other models.	×	0.12
SigLIP improves robustness but is weak on calibration.	✓	0.15
LLM prompts improve accuracy but may not benefit OOD detection or calibration.	✓	0.19
No prompt method consistently dominates in terms of performance.	×	0.11
LP-FT and WiSE-FT are well-balanced in terms of fine-tuning impact.	✓	0.17
FLYP boosts accuracy but hurts calibration in fine-tuning.	✓	0.19
PromptSRC preserves both accuracy and calibration in fine-tuning.	×	0.13
Data curation boosts classification, OOD, retrieval, and 3D robustness for ViT-CLIP.	✓	0.24
Data curation does not improve calibration for ViT-CLIP.	×	0.10
CLIP models exhibit a stronger shape bias compared to other groups.	×	0.08
CLIP models with CNN-based vision encoders exhibit a significant shape bias	×	0.10

References

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2306.07713v3
- http://arxiv.org/abs/2410.01534v2