# Enhancing CLIP Cross-Domain Robustness via Adversarial Pretraining Strategies

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: Can cross-domain robustness of CLIP models (measured by accuracy differentials between ImageNet and ImageNet-Sketch) be improved by integrating adversarial pretraining strategies from tabular. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Improving Adversarial Transferability of Vision-Language Pre-training Models through Collaborative Multimodal Interaction. Research question: Can cross-domain robustness of CLIP models (measured by accuracy differentials between ImageNet and ImageNet-Sketch) be improved by integrating adversarial pretraining strategies from tabular foundation models into their training pipeline?.

## 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 4.5/10.

## 3 Results

14 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.5/10.

## 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 |
|---|---|---|
| VLP models play a crucial role in offering a universal solution for multiple tasks, including image-text retrieval (ITR) | × | 0.12 |
| Recent studies have elucidated the vulnerability and sensitivity of VLP models to adversarial examples. | × | 0.07 |
| Single-modal attacks such as PGD and BERT-Attack exhibit good adversarial performance in the visual and text domains. | × | 0.06 |
| Applying single-modal attacks directly to VLP models still poses challenges because VLP models integrate multimodal info | × | 0.08 |
| Sep-Attack directly combines both BERT-Attack and PGD. | × | 0.05 |
| Co-Attack considers image-text collaborative information and is specifically designed for customized attack forms for di | × | 0.09 |
| The proposed Collaborative Multimodal Interaction Attack demonstrates effectiveness in experimental results. | ✓ | 0.17 |
| VLP models can be classified into two categories: Fused VLP models and Aligned VLP models. | × | 0.07 |
| Fused VLP models (e.g., ALBEF, TCL) use a single encoder to extract feature representations from both images and text, f | × | 0.08 |
| Aligned VLP models (e.g., CLIP) use a single encoder to independently learn feature representations. | × | 0.05 |

## References


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

- http://arxiv.org/abs/2512.03307v1
- http://arxiv.org/abs/2403.10883v2