Scaling Image-Text Pairs and Adversarial Transferability in CLIP-Based Models

Assignee Research

June 11, 2026

Abstract

As a general-purpose vision-language pretraining model, CLIP demonstrates strong generalization ability in image-text alignment tasks and has been widely adopted in downstream applications such as image classification and image-text retrieval. However, it struggles with fine-grained tasks such as object detection and semantic segmentation. While many variants aim to improve CLIP on these tasks, its robustness to adversarial perturbations remains underexplored. Understanding how adversarial examples transfer across tasks is key to assessing CLIP's generalization limits and security risks. In th

1 Introduction

This paper examines: Bridging the Task Gap: Multi-Task Adversarial Transferability in CLIP and Its Derivatives. Research question: What is the effect of scaling the number of one-to-many image-text pairs during training on the adversarial transferability (measured via attack success rate) of CLIP-based models across different visual-language benchmarks (e.g., COCO, VCR)?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

3 Results

16 papers retrieved. 9 claims extracted; 8 independently verified. Quality review score: 8.1/10.

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
MT-AdvCLIP is a multi-task adversarial attack framework designed for CLIP-based models.	✓	0.20
MT-AdvCLIP employs a task-driven staged perturbation aggregation strategy.	✓	0.22
MT-AdvCLIP contains two core modules: Gradient-Guided Task Decoupling and Transferable Perturbation Weighting.	✓	0.19
In MT-AdvCLIP, strong perturbations are optimized on fine-grained tasks to exploit richer local gradients and avoid grad	✓	0.26
In MT-AdvCLIP, perturbations are refined under CLIP supervision to improve generalization in the global feature space.	✓	0.18
In MT-AdvCLIP, perturbations from fine-grained models are weighted more heavily than those directly on CLIP.	✓	0.20
In MT-AdvCLIP, perturbations generated directly on CLIP are scaled down to protect its feature space.	✓	0.17
Experiments for MT-AdvCLIP were conducted across tasks including image-text retrieval and object detection on MSCOCO.	✓	0.20
MT-AdvCLIP improves attack success by 39% compared to baselines.	×	0.12

References

- http://arxiv.org/abs/1907.02664v2
- http://arxiv.org/abs/2105.04834v3
- http://arxiv.org/abs/2509.23917v1