Unrestricted Adversarial Attacks on CodeT5 Semantic Consistency in MBPP Benchmarks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of unrestricted adversarial attacks on the semantic consistency scores of CodeT5 when evaluated on the MBPP dataset across different pretraining language distributions. 14 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: SCA: Improve Semantic Consistent in Unrestricted Adversarial Attacks via DDPM Inversion. Research question: What is the impact of unrestricted adversarial attacks on the semantic consistency scores of CodeT5 when evaluated on the MBPP dataset across different pretraining language distributions?.

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. 14 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
The experiments were conducted using the ImageNet-compatible dataset, which includes 1,000 images from the ImageNet vali	×	0.03
The ImageNet-compatible dataset has been frequently employed in related work.	×	0.04
Semantic consistency is evaluated using CLIP Score, SSIM, PSNR, and LPIPS.	×	0.05
SCA is compared with other unrestricted adversarial attacks, including SAE, ADef, cAdv, tAdv, ACE, ColorFool, NCF, AdvST	×	0.06
The evaluation metric employed is attack success rate (ASR), which measures the proportion of images misclassified by th	×	0.04
Target models include MobileNet-V2, Inception-v3, ResNet-50, ResNet-152, DenseNet-161, EfficientNet-b7, MobileViT, Visio	×	0.05
All experiments are performed using PyTorch on an NVIDIA Tesla A100.	×	0.04
The DDPM steps are set to T = 10 $\rightarrow$20, with attack iterations Na = 10, η = 0.04, κ = 0.1, and μ = 1.	×	0.04
The version of Stable Diffusion used is v1.5, and image captions are generated automatically via LLaVA-NeXT.	×	0.03
SCA generates the most natural adversarial examples and maintains a high degree of semantic consistency with the clean i	×	0.10
The core idea of Semantic-Consistent Unrestricted Adversarial Attack is to enhance semantic control throughout the entir	✓	0.19
The pipeline of the proposed method includes Semantic Fixation Inversion and Semantically Guided Perturbation.	×	0.07
The algorithm of SCA is presented in Algorithm 1.	×	0.04
Unrestricted adversarial attacks seek to introduce subtle adversarial perturbations to the input x, resulting in an adve	×	0.07

References

- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/1904.06347v2
- http://arxiv.org/abs/2410.02240v6