Adversarial Perturbation Tolerance in Segment Anything Model vs. Multimodal Code-Generation Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does adversarial perturbation tolerance in Segment Anything Model compare to multimodal code-generation models like CodeT5 when evaluated on structural integrity metrics. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Attack-SAM: Towards Attacking Segment Anything Model With Adversarial Examples. Research question: How does adversarial perturbation tolerance in Segment Anything Model compare to multimodal code-generation models like CodeT5 when evaluated on structural integrity metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.1/10.

3 Results

12 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.1/10.

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
In the vanilla version of SAM, a single point is chosen in the image and a single mask near the given point is predicted	×	0.06
The attack succeeds in the mask removal task if Maskadv is empty or has a much smaller area than Maskclean.	×	0.10
Clean images for visualization were randomly selected from SAM project demo images or the SA-1B dataset.	×	0.06
FGSM and PGD attacks generate adversarial images with imperceptible perturbations.	×	0.07
Both FGSM and PGD attacks are able to remove the area of Maskclean.	×	0.03
The PGD attack performs better than the FGSM attack in the mask removal task.	×	0.09
The IoU metric is used for quantitative evaluation of adversarial attacks on SAM.	×	0.07
mIoU is calculated as the average IoU between Maskadv and Maskclean over N data pairs.	×	0.04
The maximum value of mIoU is 1 when the perturbation vector is zero (no attack).	×	0.04
With the proposed ClipMSE, the mIoU drops from 1 to close to zero after a PGD attack.	×	0.02
The FGSM attack achieves an mIoU much smaller than 1.	×	0.02
SAM solves promptable segmentation where the model generates masks based on image and prompt inputs.	×	0.08
Generated masks in SAM cut out detected objects but do not have semantic labels.	×	0.05
In SAM, a pixel is marked within the mask area if the predicted confidence value is positive.	×	0.04
The final predicted masks (Maskpred) in SAM are binary matrices with shape H*W.	×	0.06

References

- http://arxiv.org/abs/2305.00866v2
- http://arxiv.org/abs/2405.18770v6

- http://arxiv.org/abs/2407.13111v1