Evaluating Segment Anything Model 3 Against Adversarial Perturbations

Matthew Sokoloff, YSP Student, *Cambridge Rindge and Latin School*
Shahriar Rifat, Electrical and Computer Engineering, *Northeastern University*
Mohammad Abdi, Electrical and Computer Engineering, *Northeastern University*
Francesca Meneghello, Principal Research Scientist, Electrical and Computer Engineering, *Northeastern University*
Francesco Restuccia, Associate Professor, Electrical and Computer Engineering, *Northeastern University*

Abstract

Segment Anything Model 3 (**SAM 3**) performs promptable concept segmentation, allowing users to **identify and segment objects using text prompts** rather than a fixed set of classes. Although this flexibility is valuable for safety-critical computer-vision applications, its predictions may remain vulnerable to small input perturbations that are difficult for humans to perceive. This study **evaluated SAM 3 on a balanced subset of the Mapillary Traffic Sign Dataset containing 3,000 annotated traffic-sign targets across 15 common classes**. Using the generic prompt "traffic sign," clean predictions were matched to ground-truth annotations using bounding-box intersection over union (IoU). **SAM 3 failed to localize 24.9% of clean targets**, with substantial variation among sign classes. **We then selected 300 targets detected correctly under clean conditions and compared standard Projected Gradient Descent (PGD), additive PGD, and spatially smooth additive PGD** at perturbation (ε) budgets of 2/255, 4/255, and 8/255. Attack effectiveness was measured using conditional attack success rate (ASR), target IoU, zero-detection rate, and peak signal-to-noise ratio (PSNR). At ε = 8/255, standard, additive, and smooth-additive PGD **caused failure rates (ASR) of 96.0%, 97.3%, and 76.7%, respectively.** These results **demonstrate substantial vulnerability under increasingly physically constrained perturbations** and establish a digital baseline for future evaluation using a calibrated projector-camera system.

Background

SAM 3 uses short noun phrases and/or image exemplars to **locate and segment every matching object instance at the pixel level**. Its Perception Encoder converts the inputs into visual and semantic features. The multimodal detector combines these features to **predict object locations, confidence scores, and separate instance masks**. Compared with SAM 2, **SAM 3 introduces open-vocabulary concept prompting** and uses a **presence head to decouple concept recognition from object localization**, improving discrimination between related prompts. However, SAM 3's robustness to adversarial image perturbations has not yet been studied motivating this evaluation (Fig. 1).

- Adversarial optimization: Gradients identify **small input changes** that **reduce target overlap** and confidence while the model and prompt remain **fixed**.
- Standard PGD: Allows **positive and negative pixel changes**.
- Additive PGD: Restricts the perturbation to **nonnegative** pixel changes.
- Smooth-additive PGD: Applies Gaussian spatial **smoothing** to the **nonnegative** perturbation. (See comparison, Fig. 4)

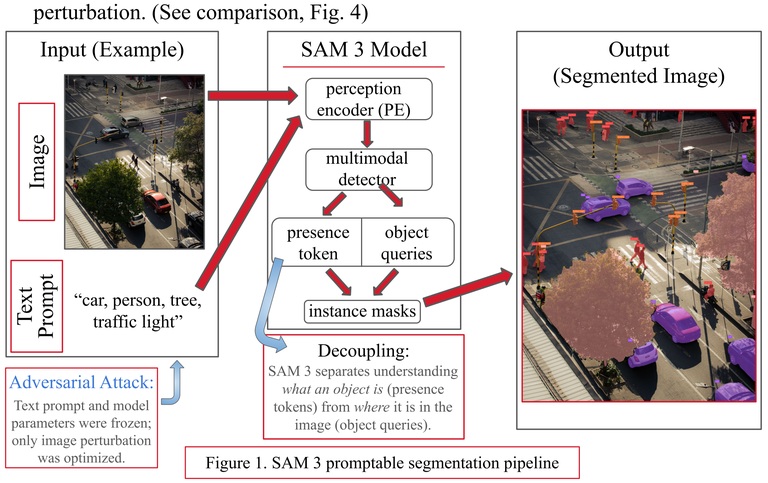


Figure 1. SAM 3 promptable segmentation pipeline

Methods

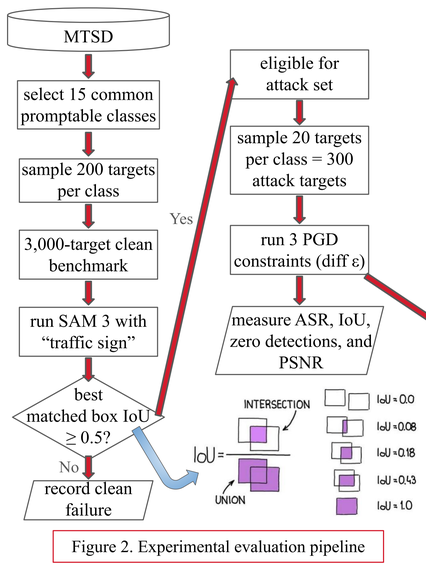


Figure 2. Experimental evaluation pipeline

- **Selected the 15 most frequent promptable traffic-sign classes** in the fully annotated **Mapillary Traffic Sign Dataset (MTSD)**
- Sampled **200 annotated targets per class**, producing a balanced clean benchmark of **3,000 targets**.
- Ran frozen SAM 3 using the **fixed text prompt** "traffic sign."
- Converted predicted masks to bounding boxes; **clean localization succeeded** when the best matched box achieved **IoU ≥ 0.50**.
- Selected **20 clean-correct targets per class, resulting in 300 attack targets.**
- **Optimized** the image **perturbation**; SAM 3 parameters remained fixed (Fig. 3).
- Used 10 PGD steps for standard and additive attacks and 50 steps for smooth-additive attacks following a convergence ablation.
- Defined **attack success** as **reducing** the **target's adversarial IoU below 0.50**.

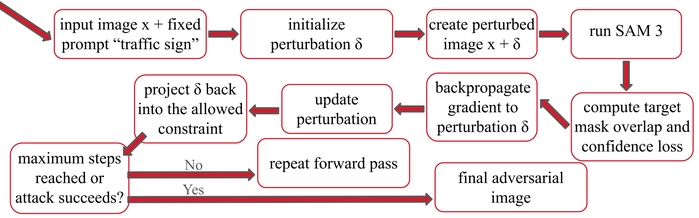
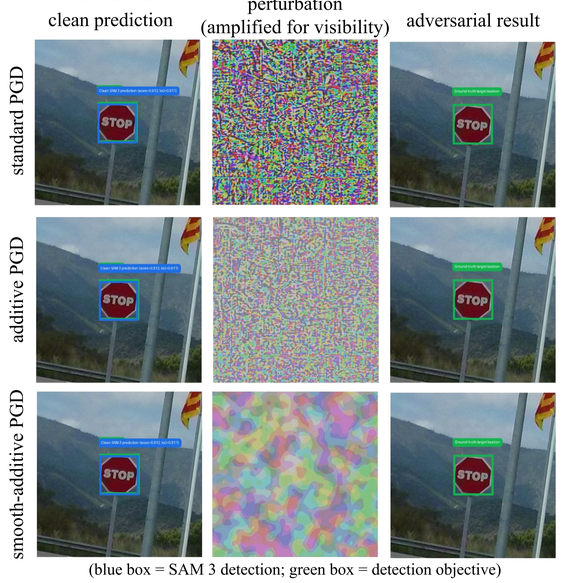


Figure 3. PGD attack optimization process

Results

- SAM 3 failed to localize 747 of 3,000 clean targets, an overall **clean failure rate of 24.9%**.
- Most clean failures were localization or target-matching errors, as **only 2.5% of samples returned zero detections**.
- **Clean reliability varied substantially by class**: parking signs had a 53.5% failure rate, compared with 11.5% for no-stopping signs.



(blue box = SAM 3 detection; green box = detection objective)

Figure 4. Qualitative effects of adversarial perturbations on SAM 3. In the clean image (left), the green and blue boxes are superimposed, meaning correctly predicted. In the corrupted image (right), the absence of the blue box indicates a detection failure.

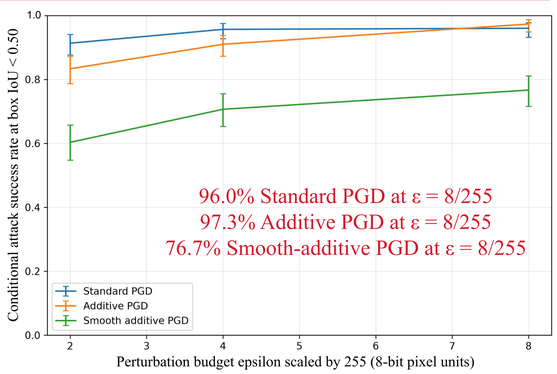


Figure 5. Conditional attack success versus perturbation budget

Across all targets:

- At **ε = 2/255**, conditional ASR was **91.3%** for standard PGD, **83.3%** for additive PGD, and **60.3%** for smooth-additive PGD.
- At **ε = 8/255**, conditional ASR reached **96.0%**, **97.3%**, and **76.7%**, respectively (Fig. 5).
- **Most successful attacks eliminated the matching detection entirely**. At ε = 8/255, **zero-detection rates** were **94.3%**, **95.7%**, and **68.7%**, respectively.
- At ε = 2/255, mean PSNR remained between 43.35 and 46.41 dB, consistent with low-amplitude image distortion.
- Increasing smooth-additive optimization from **10** to **50** steps **improved ASR** by **32–44** percentage points while changing mean PSNR by less than approximately 0.3 dB.
- The **smooth constraint** was therefore **more difficult to optimize**, rather than inherently ineffective.

Conclusion and Future Steps

- SAM 3 exhibited substantial class-dependent clean localization failure on traffic signs.
- All three PGD methods caused **high failure rates** on targets detected correctly under clean conditions.
- At larger budgets, additive PGD approached or exceeded standard signed PGD; **effective attacks did not require negative pixel perturbations**.
- Smooth-additive PGD remained effective despite its spatial constraint, reaching 76.7% conditional ASR.
- These findings establish a projector-inspired **digital baseline**.
- Future work will calibrate a projector-camera system, model optical and environmental effects, project optimized patterns onto physical traffic signs, and compare clean, digital, and physical outcomes (Fig. 6).

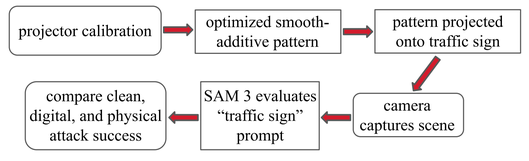


Figure 6. Proposed physical projected-light evaluation

References

Bolya, D., Huang, P.-Y., Sun, P., Cho, J. H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, D., Dollár, P., & Feichtenhofer, C. (2025). Perception encoder: The best visual embeddings are not at the output of the network [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2504.13181

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1712.09665

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., . . . Feichtenhofer, C. (2025). SAM 3: Segment anything with concepts [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.16719

Ertler, C., Mislej, J., Ollmann, T., Porzi, L., Neuhold, G., & Kuang, Y. (2020). The Mapillary traffic sign dataset for detection and classification on a global scale. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), Computer vision—ECCV 2020 (Lecture Notes in Computer Science, Vol. 12368, pp. 68–84). Springer. https://doi.org/10.1007/978-3-030-58592-1_5

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramèr, F., Prakash, A., Kohno, T., & Song, D. (2018). Physical adversarial examples for object detectors. In 12th USENIX Workshop on Offensive Technologies (WOOT 18). USENIX Association. https://www.usenix.org/conference/woot18/presentation/eykholt

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1412.6572

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2304.02643

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1706.06083

Nichols, N., & Jasper, R. (2018). Projecting trouble: Light based adversarial attacks on deep learning classifiers [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1810.10337

Acknowledgements

MENTIS Laboratory
Francesco Restuccia, Associate Professor
Francesca Meneghello, Principal Research Scientist
Center for STEM Education
Claire Duggan, Executive Director
Jennifer Love, Associate Director
Ahmed Othman, Kenneth Santizo & Frances Corcell, YSP Coordinators
Nicolas Fuchs, Program Manager
Mary Howley, Administrative Officer

Young Scholars Program at Northeastern University
Claire Duggan, Program Director

This work was supported by the National Science Foundation