Topological Signatures of Adversaries in Multimodal Alignments

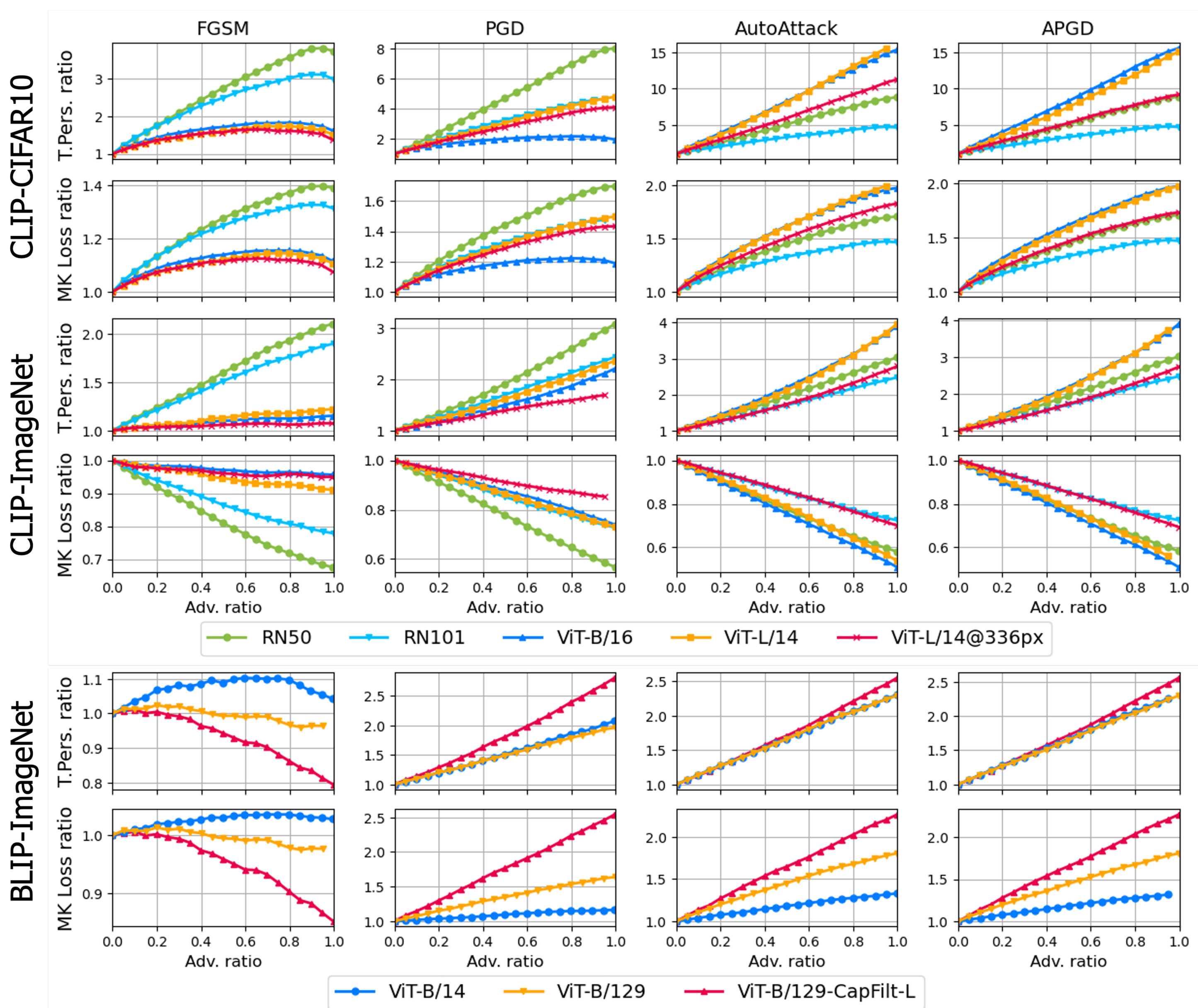
Minh N. Vu, Geigh Zollicoffer, Huy Mai, Ben Nebgen, Boian Alexandrov, Manish Bhattarai
mvu@lanl.gov, gzollicoffer3@gatech.edu, huyqmai1@gmail.com, bnebgen@lanl.gov, boian@lanl.gov, ceodspspectrum@lanl.gov

Abstract

Multimodal Machine Learning systems, such as CLIP/BLIP models, have become increasingly prevalent, yet remain susceptible to adversarial attacks. This work investigates the **topological signatures that arise between image and text embeddings** and shows how adversarial attacks disrupt their alignment. We specifically leverage persistent homology and introduce two novel **Topological-Contrastive losses** based on Total Persistence and Multi-scale kernel methods to analyze the topological signatures introduced by adversarial perturbations. We observe **a pattern of monotonic changes in the proposed topological losses** emerging in a wide range of attacks as more adversarial samples are injected in the data. We then integrate these signatures into Maximum Mean Discrepancy tests, creating a novel class of tests that leverage topological signatures for better adversarial detection.

Topological Signatures of Adversaries

Monotonic behavior of Topological Signatures: the topological signatures of the logits exhibit a consistent, **monotonic** change as the proportion of adversarial examples in the data increases.



Topological Contrastive Losses

Total Persistence Loss: For a dimension i, the α-total persistence of dimension i is computed on the persistence diagram $D_i(X)$:

$$\mathrm{Pers}_i^{\alpha}(X) := \sum_{(b,d)\in D_i(X)} (d-b)^{\alpha}$$

The TP loss of order α between two point clouds is the summation of the difference at all homology groups:

$$\mathcal{L}_{TP}^{\alpha}(X,Y) = \sum_i \left|\mathrm{Pers}_i^{\alpha}(X) - \mathrm{Pers}_i^{\alpha}(Y)\right|$$

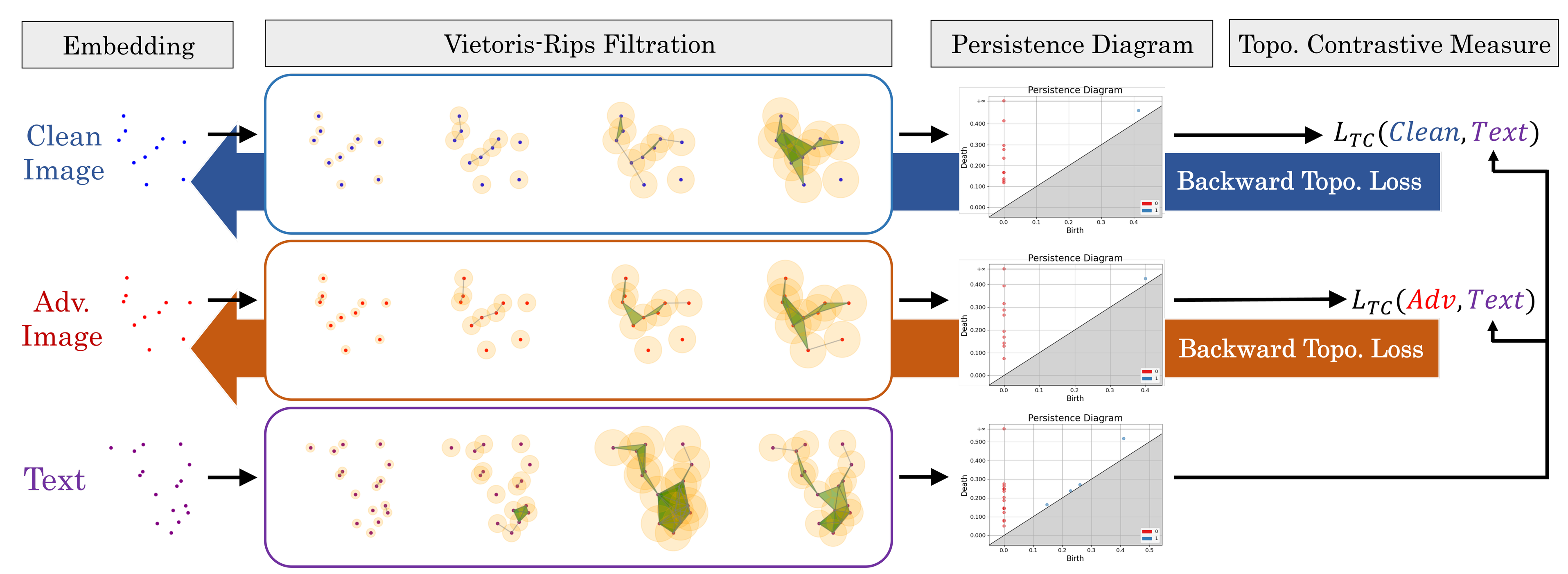
Multi-scale Kernel Loss: The loss is based on the a kernel $k_\sigma : \mathcal{D}\times\mathcal{D}\to\mathbb{R}$ acting on persistence diagrams of point clouds X and Y:

$$k_\sigma(D_i(X), D_i(Y)) := \frac{1}{8\pi\sigma} \sum_{p\in D_i(X), q\in D_i(Y)} e^{-\frac{\|p-q\|_2^2}{8\sigma}} - e^{-\frac{\|p-\bar{q}\|_2^2}{8\sigma}}$$

where p and q are the birth-death pairs from the corresponding persistence diagrams, and $\bar{q} = (d,b)$ denotes the mirror of $q = (b,d)$ through the diagonal. For our purpose, we define the MK loss of scale σ between two point clouds by:

$$\mathcal{L}_{MK}^{\sigma}(X,Y) = \sum_i k_\sigma(D_i(X), D_i(Y))$$

Detection with Topological Features: We utilize $\mathcal{L}_{TC}$ for detection by computing **sample-level** features derived from the topological loss: $\dot{Y} = \nabla_Y \mathcal{L}_{TC}(Y,T)$, where Y represents the image's logits and T denotes the text embedding.

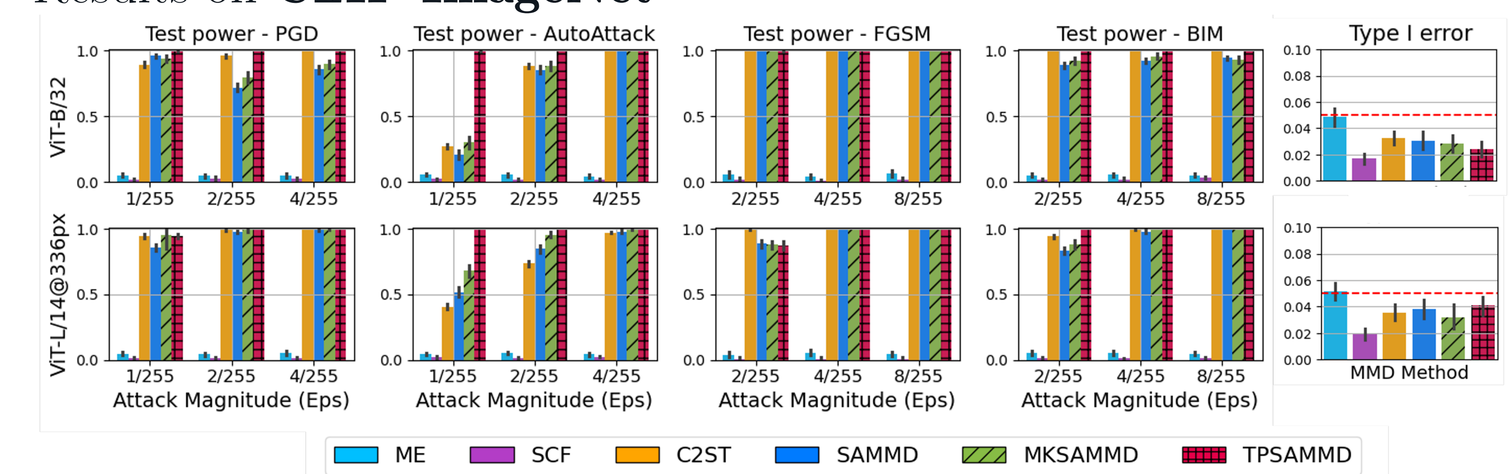


To incorporate topological features for detection, we propose the following topological-contrastive deep kernel k_τ:

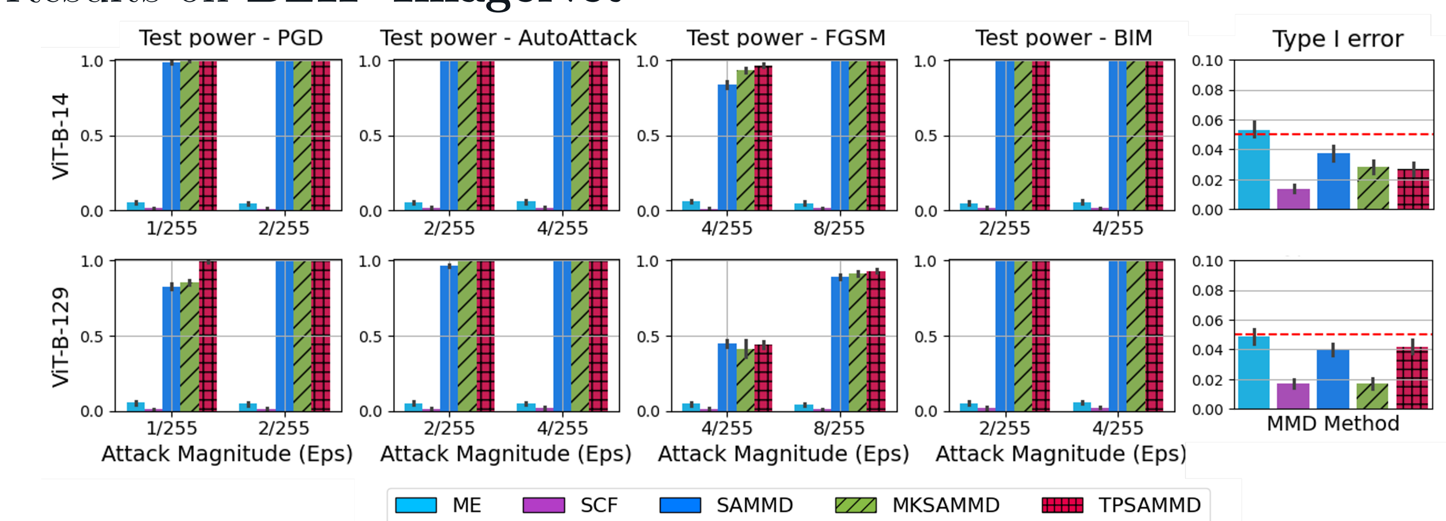
$$k_\tau(x_{\log}, y_{\log}) = \left[(1-\epsilon_0)\,\tau_{\hat{f}}(x_{\log}, y_{\log}) + \epsilon_0\right] \nu_{\hat{f}}(x_{\log}, y_{\log})$$

Experimental results

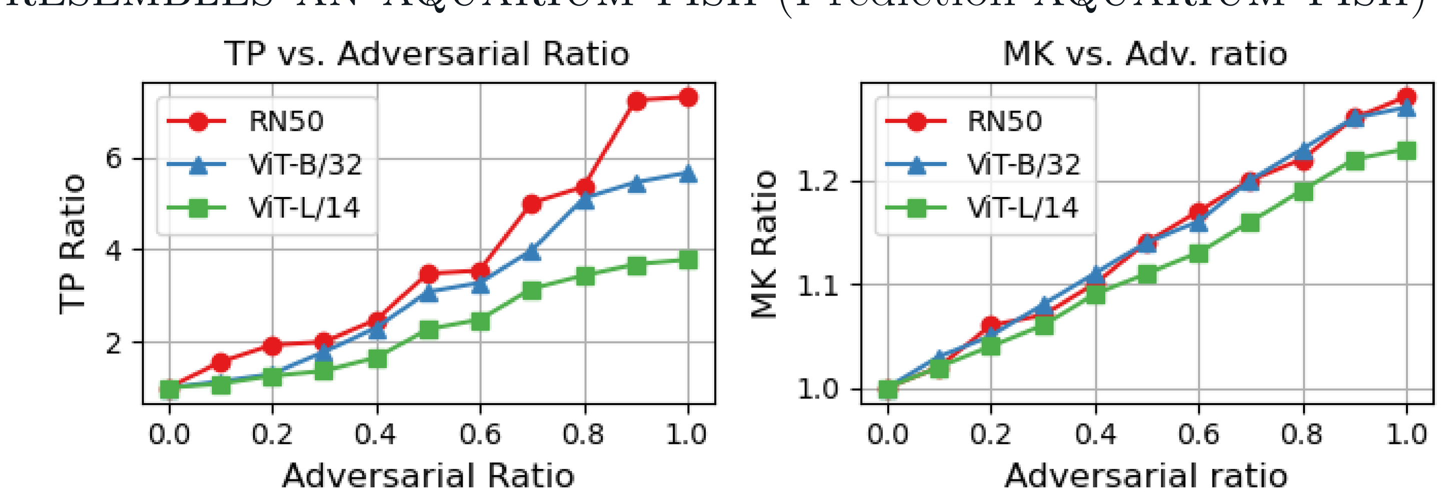
Results on **CLIP-ImageNet**



Results on **BLIP-ImageNet**



Text attacks. Adversary: A photo of an apple that resembles an aquarium fish (Prediction aquarium fish)



Acknowledgements

This work has been assigned LA-UR-25-20561. This research was funded by the Los Alamos National Laboratory (LANL) Laboratory Directed Research and Development (LDRD) program under grants 20230287ER & 20240868PRD3 and supported by LANL's Institutional Computing Program, and by the U.S. Department of Energy National Nuclear Security Administration under Contract No. 89233218CNA000001.