GeoDetect: Geometric Adversarial Detection for VLPs

Afsaneh Hasanebrahimi[1], Hanxun Huang[1], Christopher Leckie[1], James Bailey[2], and Sarah Erfani[1]

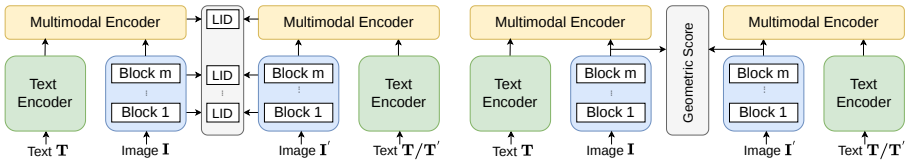
[1] The University of Melbourne, Melbourne, Australia
{a.hasanebrahimi, curtis.huang1, caleckie, sarah.erfani}@unimelb.edu.au
[2] Monash University, Melbourne, Australia
james.a.bailey@monash.edu

Abstract. Vision-language pre-trained models (VLPs) are widely used in real-world applications. However, they remain vulnerable to adversarial attacks. Although adversarial detection methods have demonstrated success in single-modality settings (either vision or language), their effectiveness and reliability in multimodal models such as VLPs remain largely unexplored. In this work, we study the geometry of VLP embedding spaces and observe structured anisotropy that differs from unimodal vision models. Our theoretical analysis shows that under this anisotropic structure, adversarial attacks increase the expected geometric separation between clean and adversarial examples (AEs). Specifically, we demonstrate that AEs consistently exhibit greater expected distances to randomly sampled points than their clean counterparts, indicating that AEs tend to push representations out of manifold regions. Building on these insights, we propose GeoDetect, which leverages these off-manifold deviations via geometric scores to identify AEs. Through comprehensive evaluations, we show that our approach reliably detects AEs across diverse VLP architectures and threat settings, covering unimodal and multimodal attacks as well as adaptive attacks, thereby providing a robust and practical approach to improving the safety and reliability of these models.

Keywords: Adversarial detection · Geometric analysis · Multimodal models

1 Introduction

Vision-language pre-trained models (VLPs) enable the understanding of both visual and textual data by learning joint representations of multimodal inputs. This capability makes them highly effective for tasks requiring a deep understanding of images and text. VLPs have achieved state-of-the-art results across various multimodal tasks [17, 57, 60], including image-text retrieval [10], visual question answering [37], and zero-shot classification [45]. Despite their remarkable success, VLPs remain vulnerable to adversarial examples (AEs) [47, 66], raising concerns about their robustness in real-world, safety-critical applications.



(a) Extraction of LID scores. **(b)** Extraction of k-NN, Mahalanobis, and KDE.

Fig. 1: Pipeline of geometric score extraction for GeoDetect.

Recent research has explored adversarial training as a strategy to enhance the zero-shot robustness of VLPs [40, 48, 54]. However, adversarial training is computationally expensive [39, 55] and often involves a trade-off between model performance and robustness [52, 64]. Detecting AEs presents a more flexible alternative by allowing the model to identify and reject potentially harmful queries, rather than attempting to provide reliable outputs for all inputs.

While existing work has been proposed for detecting AEs in unimodal models [2,16,25,28,38,50], it remains uncertain whether detection signals used in unimodal detectors can transfer to VLPs, since multimodal alignment and fusion can reshape the geometry of clean neighborhoods in embedding space. In contrast to standard unimodal classification models trained with cross-entropy loss to predict discrete labels, VLPs are optimized to align image and text representations within a shared embedding space using contrastive learning objectives [45]. Recent findings by [48] demonstrate that CLIP embeddings experience significant distortion under adversarial attack, as evidenced by substantial shifts in the embedding space. A recent study by [67] proposed Prompt-based Irrelevant Probing (PIP), a task-specific detection method for visual question answering (VQA) that analyzes attention responses to irrelevant probe questions. However, PIP's applicability is limited to architectures employing explicit cross-attention mechanisms, and its reliance on question-conditioned attention constrains it exclusively to VQA tasks. Thus, the absence of a comprehensive and theoretically grounded investigation into the nature of AEs and adversarial detection for VLPs leaves a critical gap in understanding their vulnerabilities and robustness. In this work, we investigate the intrinsic properties of VLPs by revisiting the anisotropic nature of CLIP's embedding space, as observed in prior work [29, 34], and extending this analysis to other VLPs. While existing studies primarily focus on CLIP, we show that this property extends to a broader range of VLPs, forming the foundation for our theoretical assumptions. Anisotropy indicates that representations are unevenly dispersed across embedding dimensions, creating dense and sparse directions. We leverage this property to formulate our central theoretical contribution: *AEs explore off-manifold regions of the embedding space, resulting in a greater expected distance between an AE and a random clean example, compared to the distance between the unperturbed version of the same random clean example.* This motivates us to investigate the geometric properties surrounding data representations, through which we uncover fundamental differences between the regions occupied by adversarial and clean examples.

Building on our theoretical insights, we propose GeoDetect, an effective method for detecting AEs in VLPs. GeoDetect extracts deep representations from VLP encoders and applies classical geometric metrics to compute detection scores, including Local Intrinsic Dimensionality (LID) [21], k-Nearest Neighbours distance (k-NN), Mahalanobis distance [41], and Kernel Density Estimation (KDE). These scores are then used for logistic regression or threshold-based detection of AEs. An overview of GeoDetect is illustrated in Fig. 1. Fig. 1a presents adversarial image detection using LID, while Fig. 1b illustrates detection using the other three studied methods, k-NN, Mahalanobis distance, and KDE. The key difference is that LID operates layer-wise, evaluating the outputs of both multimodal layers and other intermediate layers, while the other three methods operate on the output of the image encoder, making LID more sensitive to perturbations across the multimodal encoder. Built upon solid theoretical foundations, GeoDetect offers a cost-effective and robust method for ensuring the safety of VLPs.

The main contributions can be summarized as follows:

- We analyze the anisotropic structure of VLP embedding spaces and, building on this, theoretically demonstrate that AEs tend to lie in off-manifold regions, resulting in larger geometric deviations from clean reference points.
- We introduce GeoDetect, a novel model-agnostic detection method that employs geometric discrepancies through a family of metrics to identify AEs in VLPs across different downstream tasks (e.g., zero-shot classification and retrieval).
- We comprehensively validate the effectiveness of GeoDetect across various VLP architectures and attacks, achieving consistently high AUC scores, all without requiring fine-tuning. This makes GeoDetect both robust and lightweight compared to existing defense methods.

2 Related Work

Vision-Language Pre-Trained Models. Vision-language representation learning outperforms visual representation learning across a wide range of tasks. For instance, CLIP uses a contrastive objective (i.e., InfoNCE loss [44]) to align an image with its corresponding textual description in the feature space. VLPs aim to improve multimodal task performance by pre-training on large-scale image-to-text pairs [30]. Several recent methods utilize pre-trained object detectors with region features as a foundation for obtaining vision-language representations [11]. There are two primary types of VLPs depending on their architectures: fused and aligned [66]. Fused VLPs, such as ALBEF and TCL [59], utilize distinct unimodal encoders to handle token embeddings and visual characteristics separately. They subsequently employ a multimodal encoder to produce integrated multimodal embeddings by combining image and text embeddings. Conversely, aligned VLPs such as CLIP are composed solely of unimodal encoders that have separate embeddings for image and text modalities. This research specifically examines widely used architectures, including both fused and aligned.

Adversarial Attacks and Robustness in VLPs. Adversarial attacks aim to deceive deep learning models into misclassifying an input [51]. While previous work is centered around image classification, recent studies show that VLPs are also vulnerable to adversarial attacks. For example, [58] investigated attacks on visual question-answering models by altering the image modality. [1, 49] focused on disrupting vision-language models through text modality perturbations. [66] offered key insights into the development of multimodal attacks and improving model robustness by exploring VLPs. Building on this, [19, 20, 36] worked on enhancing the transferability of multimodal AEs by leveraging cross-modal interactions, data augmentation, and optimal transport theory. Furthermore, [61, 70] build on [66] by crafting modality-aligned perturbations that improve transferability between downstream tasks. Despite these advances, many attack techniques remain specialized for classification tasks and may not generalize well to retrieval, captioning, or grounding. Therefore, we adopt the adversarial attacks presented in [36, 66] as our attack baselines.

Recent efforts have explored enhancing the adversarial robustness of vision-language models through prompt tuning and training strategies. [32, 65] improve CLIP's resilience by learning robust textual prompts aligned with adversarial image embeddings. [69] introduces adversarial text supervision to balance cross-modal alignment and uni-modal discrimination. [54] adds an auxiliary branch to align adversarial outputs between the target and pre-trained models, reducing overfitting in zero-shot settings. [56] adopts a two-phase adversarial training regime, starting with lightweight pre-training, followed by high-resolution fine-tuning. However, the high computational overhead of these methods poses challenges for scaling to large models and datasets.

Geometric Methods and Geometric Adversarial Detection in Unimodal Models. k-NN [14] is a nonparametric algorithm that classifies points based on the majority label of their nearest neighbors, offering a simple yet powerful method for pattern recognition and regression. LID models the intrinsic dimensionality [3, 21–24] near a point by analyzing the growth rate of nearby data, providing insights into local geometric structures within a dataset. Mahalanobis distance [41] incorporates data covariance to measure similarity, enabling scale-invariant and correlation-sensitive evaluations that are effective for identifying outliers or understanding feature relationships. KDE estimates the probability density function of data in a nonparametric manner, using kernel functions and adaptive bandwidths to achieve smooth and flexible density representations [8].

Several studies have employed geometric approaches to detect AEs in unimodal classification models. [18] introduced the Maximum Mean Discrepancy (MMD), a kernel-based statistical test that distinguishes AEs from a model's training data. As an alternative to KDE, [38] employed LID to evaluate the distance distribution of an input relative to its neighbors, capturing the local complexity of the sample's surrounding space. [28] proposed using the Mahalanobis distance, using Gaussian discriminant analysis to detect out-of-distribution and adversarial samples through a generative classifier, offering a more refined confidence score

than the traditional softmax classifier. [13] further explored k-NN for adversarial detection. While these methods have shown promise in unimodal settings, their effectiveness in VLPs remains unexplored.

3 GeoDetect

In this section, we introduce GeoDetect, a geometric framework for detecting AEs by analyzing the properties of embedding geometry in VLPs. We first provide the formal problem definition in Sec. 3.1, which is followed by a theoretical analysis in Sec. 3.2.

3.1 GeoDetect Framework

Problem Setup. Let $\mathcal{D}_c = \{(x_i, t_i)\}_{i=1}^N$ be a clean dataset of N i.i.d. image-text pairs, where x_i and t_i denote the clean image and text, respectively. The adversarial dataset consists of perturbed samples and is defined as $\mathcal{D}_a = \{(x_i', t_i)\}_{i=1}^N$ when only the image is perturbed, or $\mathcal{D}_a = \{(x_i', t_i')\}_{i=1}^N$ when both modalities are perturbed, with labels $y_i \in \{0, 1\}$ indicating benign ($y_i = 0$) or adversarial ($y_i = 1$). We define the embeddings as $z_{\mathrm{I}} = E_{\mathrm{I}}(x)$ for image, $z_{\mathrm{T}} = E_{\mathrm{T}}(t)$ for text, and in the case of fused VLPs, $z_{\mathrm{M}} = E_{\mathrm{M}}(z_{\mathrm{I}}, z_{\mathrm{T}})$ as the multimodal representation. Clean embeddings are denoted as $Z_c = \{z_i\}_{i=1}^N$, where z_i represents either z_{I} or z_{M}. Our goal is to accurately detect perturbed samples, particularly those in which the image, or both the image and text modalities, have been adversarially modified. Given a query (x_i, t_i) and a reference batch $\{(x_j, t_j)\}_{j=1}^n$, with n denoting the batch size, we evaluate the detection function:

$$f((x_i, t_i), (x_j, t_j)_{j=1}^n) = \mathcal{H}(\mathrm{Metric}(z_i, \{z_j\}_{j=1}^n)), \tag{1}$$

where $\{z_j\}_{j=1}^{n, j\neq i}$ denotes a set of clean reference embeddings, n is the batch size, and $\mathcal{H}$ represents the decision function. The function $\mathrm{Metric}(\cdot, \cdot)$ represents a geometric measure, such as LID, k-NN, KDE, or Mahalanobis distance. In this paper, we adopt the maximum likelihood estimation (MLE) of LID [3], and throughout the paper, we use "LID" to refer to this estimated quantity. The binary classification $f(\cdot, \cdot)$ then determines whether the input is adversarial or clean using the computed score.

GeoDetect Pipeline. GeoDetect comprises three primary steps: generation, extraction, and detection. Following prior work [38], we assume that the defender has access to a subset of the data, and that the initial dataset $\mathcal{D}_c$ is free of AEs.

In the first step of the process, generation, AEs are created from clean samples using different adversarial attacks. Given the clean dataset, we generate perturbed image–text pairs ($x_i^{'}$ and $t_i^{'}$), resulting in a balanced set with equal proportions of clean and adversarial samples. A description of adversarial-sample generation, including the settings for each attack type, is provided in Appendix A.1.

In the extraction step, we begin by extracting clean image embeddings, z_{I}, and multimodal embeddings, z_{M}, as well as the corresponding adversarial embeddings z'_{I} and z'_{M}. To ensure scalability on large datasets, we use minibatch sampling to estimate local geometric properties, following the approach in [38], which has been shown to provide reliable approximations of neighborhood statistics. To compute geometric scores of a target embedding z_i, we randomly sample a batch of clean embeddings $\{z_j\}_{j=1}^{n}$ as reference points for the computation of different metrics. These reference points can be either z_{I} or z_{M}, depending on whether the detection is performed in the image or multimodal space. Using this reference batch, scores will be computed for both clean z_i and adversarial embeddings z'_i using the following metrics as $\mathrm{Metric}(\cdot,\cdot)$:

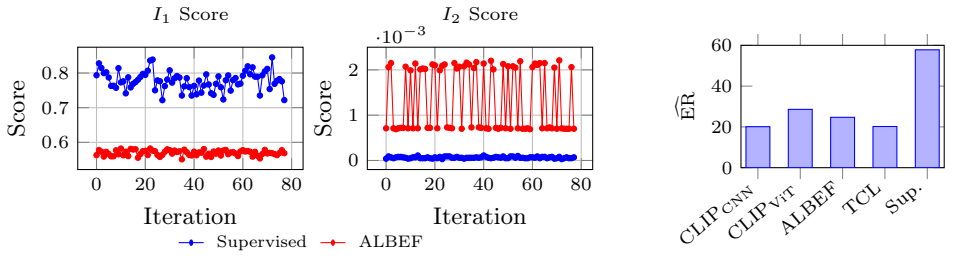
$$\mathrm{Metric}(\cdot,\cdot) = \begin{cases} \text{k-NN}(z_i, \{z_j\}) = \frac{1}{k}\sum_{j=1}^{k} r_j(z_i),\ r_j = \|z_i - z_j\|_2, \\ \widehat{\mathrm{LID}}(z_i, \{z_j\}) = \left(-\frac{1}{k}\sum_{j=1}^{k} \log \frac{r_j}{r_{\max}}\right)^{-1}, \\ \mathrm{KDE}(z_i, \{z_j\}; H) = \frac{1}{n}\sum_{j=1}^{n} K_H(z_i - z_j), \\ \mathrm{Mahal}(z_i) = \sqrt{(z_i-\mu)^T \Sigma^{-1}(z_i-\mu)}. \end{cases}$$

For the Mahalanobis distance, the mean vector μ and covariance matrix Σ are computed using the clean dataset $\mathcal{D}_c$. Similarly, for KDE, the kernel function $K_H(\cdot,\cdot)$ is estimated based on the same clean data. For all metrics, the embedding-level scores are computed as $s_i = \mathrm{Metric}(z_i, \{z_j\}_{j=1}^{n})$ for clean samples, and $s'_i = \mathrm{Metric}(z'_i, \{z_j\}_{j=1}^{n})$ for adversarial samples, where the reference points $\{z_j\}_{j=1}^{n}$ are clean embedding samples. For LID, we follow a layer-wise extraction strategy as proposed by [38]. In fused VLPs, we additionally compute the LID of the multimodal encoder z_{M} as an additional feature alongside the image encoder layers, improving detection performance against multimodal attacks. The complete procedure for computing these values for adversarial image detection is outlined in Algorithm 1 in Appendix A.2. After the extraction step, the extracted clean and adversarial scores are denoted as $s_i \in S_{(N,l)}$ and $s'_i \in S'_{(N,l)}$, where $l = 1$ for k-NN, Mahalanobis, and KDE, and l represents the number of layers for LID. These extracted scores serve as input features for the subsequent detection phase.

We frame adversarial example detection via a decision function $\mathcal{H}$. We split the extracted scores into a calibration subset (to set thresholds or train a classifier) and a test subset. For k-NN, Mahalanobis, and KDE, $\mathcal{H}$ is a threshold rule on the scalar score $s_i = \mathrm{Metric}\big(z_i, \{z_j\}_{j=1}^{n}\big)$: $\mathcal{H}(s_i) = \mathbb{I}(s_i > \tau)$, where τ is chosen on the training split. For LID, $\mathcal{H}$ is a logistic regression trained on multi-layer LID features (Appendix A.2). At test time, given (x_i, t_i), we extract embeddings $(z_{\mathrm{I}}, z_{\mathrm{T}})$, compute the chosen metric with respect to a clean reference batch $\{z_j\}_{j=1}^{n}$, and apply $\mathcal{H}$ (threshold or logistic) to decide if the input is adversarial. The efficiency of GeoDetect is reported in Appendix A.4.

3.2 GeoDetect Theoretical Analysis

In this subsection, we develop GeoDetect's theoretical foundations by formalizing core assumptions and deriving its key technical results.



(a) Isotropy scores (I_1, I_2) across iterations. **(b)** Normalized ER comparison.

Fig. 2: Comparison of isotropy metrics and effective ranking for VLPs vs. supervised model. Iteration refers to the batch index during evaluation.

We begin by empirically verifying that VLP embeddings are anisotropic, concentrated along a few dominant directions, which motivates our assumptions. Building on this, we show that the principal directions of adversarial embeddings differ significantly from those of clean embeddings, effectively pushing them off the manifold. Our theoretical analysis explains why geometric scores are particularly well-suited for adversarial detection in VLPs: they quantify off-manifold deviations that adversarial perturbations inherently induce.

Anisotropic Embedding Space. Understanding the geometric structure of the embedding space is crucial for identifying the fundamental differences between clean and adversarial embeddings in VLPs. [34] showed that CLIP features lie within a low-aperture cone, concentrating in a narrow angular region of high-dimensional space. Building on this, [29] found that CLIP's embedding space forms a double-ellipsoid geometry, with image and text embeddings located on distinct ellipsoidal shells.

Consequently, we expect that other VLPs also exhibit anisotropic embedding spaces, consistent with the patterns observed in CLIP. To formally quantify this, we adopt two established isotropy measures, I_1 and I_2 [53]. Let $Z \in R^{N \times D}$ be the matrix of embedding vectors. The measures are defined as:

$$I_1(Z) = \frac{\min_{v \in V} P(v)}{\max_{v \in V} P(v)}, \quad I_2(Z) = \sqrt{\frac{\sum_{v \in V} \left(P(v) - \bar{P}(v)\right)^2}{|V| \bar{P}(v)^2}}, \tag{2}$$

where V is the set of eigenvectors of $Z^T Z$, and $P : \mathbb{R}^D \to \mathbb{R}^+$ is the partition function $P(v) = \sum_{i=1}^{n} \exp(\langle v, z_i \rangle)$ [43]. For an embedding matrix Z to be isotropic, $P(v)$ should be approximately constant for any unit vector v [4]. Then, the second measure, $I_2(Z)$, is the normalized standard deviation of the partition function $P(v)$, where $\bar{P}(v)$ is the average value of $P(v)$. In this formulation, $I_1(Z)$ is bounded between 0 and 1 ($I_1(Z) \in [0, 1]$), while $I_2(Z)$ is non-negative ($I_2(Z) \geq 0$); lower $I_1(Z)$ and higher $I_2(Z)$ indicate stronger anisotropy.

To investigate the embedding space of the VLPs (specifically ALBEF here, with other VLPs discussed in Appendix B.1), we empirically evaluate $I_1(Z)$ and

$I_2(Z)$. As a reference, we include a supervised classifier (ResNet-50) trained on ImageNet. We compute $I_1(Z)$ and $I_2(Z)$ across iterations (x-axis) and report the corresponding metrics on the y-axis in Fig. 2a. The results show that $I_1(Z)$ is lower and $I_2(Z)$ is higher for VLPs, shown here for ALBEF, compared to the supervised classifier, indicating stronger anisotropy in VLP embedding spaces. Consistent trends for other VLPs are reported in Appendix B.1, supporting the assumptions underlying our theoretical analysis.

Another metric for evaluation of isotropy is effective rank (ER), a spectral measure of dimensionality that reflects how singular values are distributed, providing a more complex perspective compared to the traditional rank [46]. Mathematically, the ER of a matrix Z is defined based on the spectral entropy of its normalized singular values. Let σ_i be the singular values of Z, and $\hat{\sigma}_i$ represent the normalized singular values $\hat{\sigma}_i = \dfrac{\sigma_i}{\sum_j \sigma_j}$, where $\sum_i \hat{\sigma}_i = 1$. The spectral entropy and the normalized ER, denoted as $\widehat{\mathrm{ER}}$ (scaled by the logarithm of the dimension D), are given by:

$$H = -\sum_i \hat{\sigma}_i \log \hat{\sigma}_i \Rightarrow \widehat{\mathrm{ER}}(Z) = \frac{\exp(H)}{\log D}. \tag{3}$$

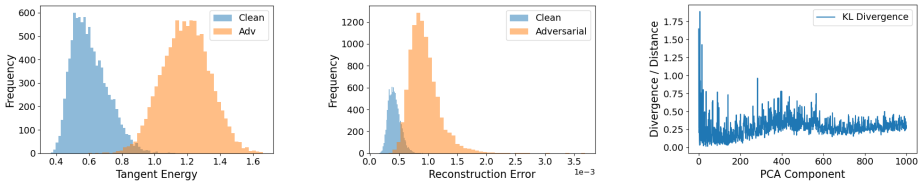
In the isotropic setting, where all singular values are equal, the estimated effective rank $\widehat{\mathrm{ER}}$ approaches the full rank of the matrix. In contrast, when the singular values are concentrated in a few dimensions, $\widehat{\mathrm{ER}}$ is substantially lower, indicating anisotropy. Fig. 2b compares the normalized effective rank of various VLPs with that of a supervised ResNet-50 trained on ImageNet V2. The results show that VLPs consistently exhibit lower $\widehat{\mathrm{ER}}$, confirming that their embedding spaces are more anisotropic than those of supervised models.

Adversarial Perturbation Effect on Geometry. In this subsection, we establish the theoretical assumptions and provide key analyses underpinning our geometric-based detection methodology. Let $\Sigma \in R^{D\times D}$ denote the covariance matrix of the clean embedding distribution. We denote clean embeddings as $z_i \in Z_c$, and adversarial embeddings as $z'_i \in Z_a$. We further represent the distributions of clean and adversarial embeddings by $p(z)$ and $q(z')$, respectively. To characterize the geometric deviation induced by adversarial perturbations, we define the distance of a target embedding (clean or adversarial) to a randomly sampled clean embedding z_u, where $u \neq i$, as:

$$Dist_{\mathrm{c}} = \|z_i - z_u\| \quad \text{and} \quad Dist_{\mathrm{a}} = \|z'_i - z_u\|.$$

We analyze the expected distances $\mathbb{E}[Dist_{\mathrm{c}}]$ and $\mathbb{E}[Dist_{\mathrm{a}}]$, and formally demonstrate that adversarial embeddings have higher expected distances, an observation underlying the effectiveness of geometric metrics in adversarial detection. We first state two assumptions that ground our theoretical analysis:

Assumption 1 (Anisotropic Covariance)**.** *The covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$ of the clean embedding space is positive-definite and anisotropic, specifically*



(a) Comparison of energy in normal direction.

(b) Comparison of reconstruction error.

(c) Divergence per PCA dimension (ordered by variance).

Fig. 3: Verification of Lemma 2: adversarial data in VLPs are off-manifold.

$\Sigma \neq cI$ for any scalar constant c, indicating anisotropy property. Consequently, its eigenvalues vary significantly across dimensions ($\sigma_1 \gg \sigma_2 \gg ... \gg \sigma_D$).

Assumption 2 (Manifold Proximity)**.** *Clean embeddings reside on a manifold $\mathcal{M}$, such that the distance of a data point z_i from the manifold satisfies a proximity condition $\|z_i - \mathcal{M}\| \leq \alpha$. Additionally, given the high dimensionality of the embeddings, we assume Gaussian distributions for both clean and adversarial embeddings: $p(z) \sim \mathcal{N}(\mu_z, \Sigma)$, $\quad q(z') \sim \mathcal{N}(\mu_{z'}, \Sigma')$.*

Assumption 1, verified in Sec. 3.2, explains the anisotropic geometry observed in VLP embeddings. Assumption 2, built on the manifold hypothesis [6], distinguishes clean embeddings that lie near the manifold from adversarial embeddings that deviate from it. To justify Assumption 2, we note that although data may be globally non-Gaussian or manifold-valued, it is standard to analyze them through local neighborhoods: non-Gaussian structures often appear approximately Gaussian when viewed locally, and curved manifolds can be locally approximated by Euclidean spaces [27]. As emphasized by [68], even globally non-Gaussian or manifold-valued data exhibit locally Gaussian behavior, since any curved manifold is locally Euclidean. Building on these assumptions, and inspired by Theorem 1 in [63], we derive the optimal adversarial embedding.

Lemma 1. *Following Assumption 2, let clean and adversarial embeddings follow $p(z) \sim \mathcal{N}(\mu_z, \Sigma)$ and $q(z') \sim \mathcal{N}(\mu_{z'}, \Sigma')$, respectively, where Σ is a fixed positive-definite covariance. Maximizing the KL divergence $\mathrm{KL}\big(q\|p\big)$ is approximately equivalent to maximizing the quadratic form of $(z_i' - z_i)^\top \Sigma^{-1}(z_i' - z_i)$, which can be transformed into a Lagrangian minimization optimization problem, which has an optimal closed-form solution:*

$$z_i'^* = (\Sigma + \lambda I)^{-1}\lambda z_i, \quad \lambda > 0, \tag{4}$$

where λ is the Lagrange multiplier.

The proof of Lemma 1 appears in Appendix B.2.

Lemma 2. *Following Assumption 1 and Lemma 1, let $\mathcal{M}$ represent the data manifold formed by clean embeddings within a local batch. When the data is perturbed, and assuming that clean embeddings lie close to the manifold $\mathcal{M}$,*

the resulting embeddings deviate from the manifold $\mathcal{M}$, thus characterized as off-manifold. Specifically, given the optimal adversarial embedding $z_i'^* = (\Sigma + \lambda I)^{-1}\lambda z_i, \quad \lambda > 0$, the deviation from the manifold satisfies $\|z_i' - \mathcal{M}\| \geq \gamma$, where $\gamma > \alpha \geq 0$ defines the minimum separation threshold for off-manifold data.

Lemma 2 shows that adversarial embeddings leave the clean manifold by suppressing tangent components and amplifying normal components (proof in Appendix B.3). We empirically verify that adversarial perturbations move samples off the clean manifold using ImageNet-V2 and CLIP_{CNN}. Fig. 3a shows that adversarial embeddings concentrate more energy in the lowest-singular-value directions (rarely used by clean data), consistent with motion into manifold-normal space. Fig. 3b further shows higher low-rank PCA reconstruction residuals for adversarial samples (top K=10 clean PCs), indicating increased distance from the clean low-dimensional subspace. Finally, Fig. 3c reports larger KL divergence between clean and adversarial distributions along leading PCA axes (most pronounced in the top 1–20 components), confirming a significant shift in the most informative representation directions.

Theorem 1 (Expected Distance Gap). *Following Lemmas 1 and 2, let $z_i \sim p(\cdot)$ be a clean embedding satisfying $\|z_i - \mathcal{M}\| \leq \alpha$, and let z_i' be an adversarial embedding satisfying $\|z_i' - \mathcal{M}\| \geq \gamma$ with $\gamma > 3\alpha$. Then*

$$\mathbb{E}_{z_u \sim p(\cdot)}\big[\|z_i' - z_u\|\big] > \mathbb{E}_{z_u \sim p(\cdot)}\big[\|z_i - z_u\|\big]. \tag{5}$$

Theorem 1 implies that given a query embedding, a large distance to randomly sampled clean embeddings strongly indicates adversarial perturbation. This justifies the effectiveness of geometric-based metrics for adversarial detection. Proof of Theorem 1 is included in Appendix B.4, and details on the connection between Theorem 1 and Lemma 2 to different geometric-based approaches are provided in Appendix B.5. Specifically, we demonstrate that under mild assumptions, AEs are expected to exhibit higher LID, k-NN, and Mahalanobis scores, along with lower KDE scores. Empirical evidence supporting Lemma 2 and Theorem 1 is provided in Appendix B.6.

4 Experiments

We evaluate GeoDetect on standard VLP tasks, including zero-shot classification and image-text retrieval. We compare against MCM [42], which is designed for CLIP-based classification and detects out-of-distribution inputs via softmax-normalized similarity scores. MCM is a score-based zero-shot baseline that can be directly applied to AE detection in multimodal models. While relevant for classification-based VLP settings, its reliance on discrete class labels makes it incompatible with retrieval tasks.

4.1 Experimental Setup

Datasets and Models. We evaluate zero-shot classification with ImageNet [15], CIFAR10, CIFAR100 [26], STL-10 [12], and Food-101 [7], as the standard datasets

for zero-shot classification. Following [45], we use class prompts of the form "a photo of a c", where c is the name of the class. For image-text retrieval, we conduct experiments on commonly used datasets, Flickr30K [62] and MS-COCO [35]. We consider two types of VLPs: aligned and fused. For aligned VLPs, we evaluate CLIP_{ViT} (using ViT-B/16) and CLIP_{CNN} (using ResNet-50) [45]. For fused VLPs, we examine ALBEF [31] and TCL [59], which consist of separate image, text, and multimodal encoders. For image-text retrieval experiments on MSCOCO and Flickr30k, we use the publicly released fine-tuned checkpoints of ALBEF and TCL.

Threat Models and Evaluation Metrics. We follow Sep-Attack and Co-Attack methods [66] due to their applicability to different models and tasks. The [CLS] embedding is widely used in pre-trained models for downstream tasks; therefore, we use it as the attack target in our evaluation. Sep-Attack perturbs each modality independently, while Co-Attack jointly targets both modalities. For image-focused attacks, we evaluate two variants of Sep-Attack: Sep_{uni}, which targets unimodal embeddings, and $\text{Sep}_{\text{multi}}$, which targets the fused multimodal representation (applicable only to fused VLPs). For image attacks, consistent with [66], we adopt iterative adversarial perturbations constrained in the ℓ_∞ norm, and a BERT-style [33] attack strategy for text attack. The maximum perturbation ϵ_i is set to $8/255$, with a step size of 1.25 for 10 iterations. For text, the perturbation budget is set to 1 token. We also evaluated the SGA attack [36], an improved version of Co-Attack, in Appendix D.2. Detailed attack configurations and success rates are reported in Appendix A.1 and Appendix A.5, respectively. We assess performance using two standard metrics: (1) the false positive rate at 95% true positive rate (FPR95), and (2) the area under the receiver operating characteristic curve (AUC).

Settings and Layers. For CLIP_{CNN} and CLIP_{ViT}, we use batch size 128 with k=100 for LID, k=10 for k-NN, and a Gaussian KDE bandwidth of 0.1. For ALBEF and TCL, batch size is 64 with k=40 for LID, k=10 for k-NN, and the same KDE bandwidth. Adversarial examples are generated from the entire test set. The resulting mixed dataset (clean and adversarial) is then randomly split into 80% for calibration (threshold fitting / LID training) and 20% for evaluation. Detailed layer selection is provided in Appendix A.3, with layer sensitivity in Appendix C.3. [3]

4.2 GeoDetect Performance

Performance of GeoDetect in Zero-Shot Classification. As shown in Tab. 1, our geometric approaches consistently outperform the MCM method across all datasets in CLIP_{CNN}, achieving lower FPR and higher AUC. This highlights GeoDetect's effectiveness for AE detection. Among the evaluated

[3] The code is publicly available in `https://github.com/AfsanehEB/GeoDetect`.

Table 1: Results on zero-shot classification performance using the area under the receiver operating characteristic curve (AUC) and the false positive rate at 95% true positive rate (FPR95). Higher AUC (↑) and lower FPR95 (↓) values indicate more accurate detection.

Model	Method	Attack	CIFAR10		CIFAR100		ImageNet1k		STL10		Food101	
			AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95	AUC	FPR95
CLIP_{CNN}	MCM	Sep_{uni}	65.47	82.88	41.13	94.15	86.10	60.35	95.82	17.92	91.70	40.18
		Co-Attack	67.10	79.54	43.99	93.21	80.83	68.38	94.10	25.64	82.14	64.38
	GeoDet-LID	Sep_{uni}	100	0.00	100	0.00	99.31	1.87	100	0.00	99.98	0.06
		Co-Attack	100	0.00	100	0.00	99.50	1.62	100	0.00	99.95	0.08
	GeoDet-k-NN	Sep_{uni}	100	0.00	100	0.00	99.65	1.62	100	0.00	100	0.00
		Co-Attack	100	0.00	100	0.00	99.67	0.89	100	0.00	100	0.00
	GeoDet-Mah.	Sep_{uni}	100	0.00	100	0.00	96.62	9.32	99.88	0.33	99.79	1.16
		Co-Attack	100	0.00	100	0.00	97.28	7.97	99.80	0.59	99.38	2.32
	GeoDet-KDE	Sep_{uni}	100	0.00	100	0.00	98.72	7.24	99.87	0.26	100	0.00
		Co-Attack	100	0.00	100	0.00	99.33	2.81	99.85	0.33	100	0.00
ALBEF	MCM	Sep_{uni}	91.20	29.02	82.80	49.19	92.15	25.38	96.83	16.02	90.26	37.03
		$\text{Sep}_{\text{multi}}$	47.43	98.23	33.55	99.56	63.03	97.59	65.32	86.60	41.98	99.46
		Co-Attack	93.34	24.64	82.67	47.27	92.14	24.94	96.45	19.70	81.29	74.70
	GeoDet-LID	Sep_{uni}	100	0.00	99.97	0.05	91.85	29.41	99.64	1.62	99.87	0.67
		$\text{Sep}_{\text{multi}}$	99.96	0.20	99.85	0.44	78.77	67.68	96.63	15.65	92.31	33.27
		Co-Attack	100	0.00	99.98	0.05	93.85	20.07	99.85	0.69	99.92	0.42
	GeoDet-k-NN	Sep_{uni}	100	0.00	100	0.00	98.60	7.23	99.97	0.19	99.98	0.04
		$\text{Sep}_{\text{multi}}$	99.27	3.05	99.21	3.25	51.92	93.61	75.95	75.75	86.46	50.96
		Co-Attack	100	0.00	100	0.00	98.64	7.33	99.96	0.19	99.98	0.04
	GeoDet-Mah.	Sep_{uni}	100	0.00	100	0.00	99.94	0.20	100	0.00	100	0.00
		$\text{Sep}_{\text{multi}}$	100	0.00	100	0.00	81.41	64.82	99.25	3.19	99.16	3.92
		Co-Attack	100	0.00	100	0.00	99.93	0.25	100	0.00	100	0.00
	GeoDet-KDE	Sep_{uni}	99.38	0.71	100	0.00	96.78	16.93	99.70	1.06	99.95	0.16
		$\text{Sep}_{\text{multi}}$	99.24	0.86	99.85	0.81	66.83	81.75	88.63	66.19	87.61	49.27
		Co-Attack	99.38	0.76	100	0.00	96.67	18.09	99.72	1.00	99.94	0.16

metrics, k-NN surpasses other metrics (particularly Mahalanobis and KDE) in CLIP_{CNN}, with LID showing comparable performance to k-NN in this context. On ALBEF, Mahalanobis slightly outperforms other metrics, particularly KDE, highlighting its sensitivity to image-level perturbations, while LID performs comparably in multimodal attacks, emphasizing the value of incorporating multimodal embeddings. Despite ALBEF's multimodal design, image perturbations still yield detectable shifts in the embedding space, which Mahalanobis effectively captures by modeling the covariance of clean image features. Extended results for CLIP_{ViT} and TCL in Appendix E.1 exhibit consistent patterns with those in this subsection.

Performance of GeoDetect in Image-Text Retrieval. We also evaluate image-text retrieval to demonstrate that GeoDetect applies beyond classification, without labeled data. Due to the lack of labels, we evaluate only LID and k-NN distance, as they do not require class labels. As shown in Tab. 2, the performance of all models is comparable to their classification results. For both the COCO and Flickr30k datasets, each image is annotated with five captions. To maintain

Table 2: Results on image-text retrieval with Flickr30k and COCO dataset evaluated using the area under the receiver operating characteristic curve (AUC) and the false positive rate at 95% true positive rate (FPR95). Higher AUC (↑) and lower FPR95 (↓) values indicate more accurate detection.

(a) Results for CLIP_{CNN} and ALBEF Models

Model	Method	Attack	Dataset			
			Flickr30k		COCO	
			AUC	FPR95	AUC	FPR95
CLIP_{CNN}	LID	Sep_{uni}	98.45	4.52	99.54	1.46
		Co-Attack	98.90	4.52	99.50	1.56
	k-NN	Sep_{uni}	99.99	0.00	99.97	0.00
		Co-Attack	99.97	0.00	99.95	0.02
ALBEF	LID	Sep_{uni}	94.99	23.83	91.80	35.88
		$\text{Sep}_{\text{multi}}$	74.26	78.75	79.85	64.51
		Co-Attack	93.80	27.98	91.49	35.58
	k-NN	Sep_{uni}	99.75	1.05	98.54	5.67
		$\text{Sep}_{\text{multi}}$	54.84	92.46	57.02	88.27
		Co-Attack	99.88	0.50	98.73	6.84

(b) Results for CLIP_{ViT} and TCL Models

Model	Method	Attack	Dataset			
			Flickr30k		COCO	
			AUC	FPR95	AUC	FPR95
CLIP_{ViT}	LID	Sep_{uni}	99.37	1.51	99.98	0.20
		Co-Attack	96.55	25.63	99.06	5.08
	k-NN	Sep_{uni}	100.00	0.00	100.00	0.00
		Co-Attack	99.59	0.50	99.51	1.27
TCL	LID	Sep_{uni}	90.88	40.93	89.32	42.52
		$\text{Sep}_{\text{multi}}$	84.72	56.47	83.95	58.16
		Co-Attack	90.76	37.31	88.25	43.79
	k-NN	Sep_{uni}	96.10	19.60	98.01	11.24
		$\text{Sep}_{\text{multi}}$	32.89	95.48	33.89	95.41
		Co-Attack	96.59	15.07	98.05	11.73

consistency, as Co-Attack requires a matching prompt to simultaneously attack both the image and the associated text, we use the first caption as the target text.

4.3 Evaluation of Adaptive Attacks

We evaluate GeoDetect under strong white-box adaptive attacks that explicitly incorporate the detector into the perturbation optimization. Following prior work on adaptive adversarial attacks [5, 9], we consider two types of adaptive attacks: (i) attacks generated from a different batch distribution than the detection batch, and (ii) Selective gradient descent adaptive attacks that balance misclassification and detection objectives. In the following attack setting, the batch size for CLIP_{CNN} and CLIP_{ViT} is set to 128, with $k = 100$ for LID and $k = 10$ for k-NN. For ALBEF and TCL, the batch size is set to 32, with $k = 20$ for LID and $k = 10$ for k-NN.

Different-Distribution Adaptive Attacks. We first evaluate adaptive attacks where the batch used for attack generation differs from the batch used for detection. In this adaptive setting, the attacker is assumed to know the VLP, the detector family, the calibration rule, and the geometric score used during optimization. However, the final evaluation uses a disjoint clean reference batch from the one used during attack generation. This prevents the attacker from exploiting the specific local neighborhood structure of a single batch and avoids misleading gradients caused by unstable k-nearest-neighbor estimates [5]. Adaptive perturbations are optimized using

$$\mathcal{L}_{adaptive}(z_i, z_i') = \mathcal{L}_{main}(z_i, z_i') - \zeta \cdot \text{Metric}(z_i', \{z_j\}_{j=1}^{n}), z_i \in B_g \neq B_d. \quad (6)$$

Table 3: GeoDetect discrimination power (AUC score) comparison between different distribution adaptive and non-adaptive attacks for Image-Retrieval Task with Flickr30k and COCO dataset in aligned VLPs (CLIP_{CNN} and CLIP_{ViT}), and fused VLPs (ALBEF and TCL) (Note: 'N-adaptive' refers to the Non-adaptive method.)

(a) Effect of k-NN adaptive Attacks

Model	Attack	Dataset			
		Flickr30k		COCO	
		N-adap.	Adap.	N-adap.	Adap.
CLIP_{CNN}	Sep_{uni}	99.99	67.82	99.97	74.47
	Co-Attack	99.97	67.32	99.95	74.28
CLIP_{ViT}	Sep_{uni}	100	61.16	100	68.44
	Co-Attack	99.59	59.68	99.51	67.68
ALBEF	Sep_{uni}	99.75	51.38	98.54	71.53
	$\text{Sep}_{\text{multi}}$	54.84	49.82	57.02	69.88
	Co-Attack	99.88	51.02	98.73	71.53
TCL	Sep_{uni}	96.10	51.27	98.01	74.19
	$\text{Sep}_{\text{multi}}$	32.89	50.57	33.89	74.04
	Co-Attack	96.59	51.18	98.05	74.00

(b) Effect of LID adaptive Attacks

Model	Attack	Dataset			
		Flickr30k		COCO	
		N-adap.	Adap.	N-adap.	Adap.
CLIP_{CNN}	Sep_{uni}	98.45	83.81	99.54	93.14
	Co-Attack	98.90	82.81	99.50	94.67
CLIP_{ViT}	Sep_{uni}	99.37	31.56	99.98	83.82
	Co-Attack	96.55	52.24	99.06	86.96
ALBEF	Sep_{uni}	94.99	71.78	91.80	89.86
	$\text{Sep}_{\text{multi}}$	74.26	89.83	79.85	92.31
	Co-Attack	93.80	70.12	91.49	90.11
TCL	Sep_{uni}	90.88	77.15	89.32	88.70
	$\text{Sep}_{\text{multi}}$	84.72	91.83	83.95	94.64
	Co-Attack	90.76	78.10	88.25	90.02

Here, $\text{Metric}(z_i', \{z_j\}_{j=1}^n)$ represents the LID or k-NN function that computes the score for AE embeddings z_i' relative to the clean sample embeddings z_i. The batch used for attack generation, B_g, is different from the batch used for detection, B_d. We set $\zeta = 0.1$, $k = 20$ for LID, and $k = 10$ for k-NN in optimization of attacks. The results presented in Tab. 3 show that across aligned VLPs (CLIP) and fused VLPs (ALBEF and TCL) on Flickr30k and COCO, GeoDetect remains robust under this challenging setting. In particular, GeoDetect-LID maintains strong discrimination power with high AUC scores even when the attacker has full white-box access to the detector objective. While adaptive attacks reduce performance compared to non-adaptive settings, GeoDetect continues to reliably distinguish adversarial from clean samples across all models.

Selective Gradient Adaptive Attacks. We also evaluate selective gradient descent adaptive attacks [9], which alternate optimization between misclassification and detection objectives to avoid local minima. These attacks provide an even stronger adaptive setting. Detailed explanation and experimental results for this attack type are reported in Appendix D.3.

4.4 Extended Evaluation and Ablation Study

Appendix B.6 provides empirical verification of GeoDetect, including visualizations showing clear separability between clean and adversarial samples via geometric scores. Sensitivity analyses over neighborhood size, sample availability, batch size, layer choice, and multimodal layers are presented in Appendix C,

demonstrating robustness to these variations. Appendix D evaluates generalization to diverse attack backbones, the SGA attack [36], and selective gradient adaptive attacks; GeoDetect remains robust. Appendix E reports extended evaluations on additional models (TCL, CLIP_{ViT}) and a comparison with PIP [67] (a VQA-specific adversarial detector), showing that GeoDetect maintains its performance across models and achieves superior results to PIP.

5 Conclusion

In this paper, we propose the first task-agnostic, theoretically grounded framework for detecting AEs in VLPs. By leveraging the anisotropic structure of VLP embedding spaces, we show through theoretical analysis that adversarial perturbations push embeddings into off-manifold regions, leading to fundamental geometric differences between clean and perturbed samples. Building on this insight, we introduce GeoDetect, a lightweight, model-agnostic detection method that applies simple geometric metrics to image or joint representations. GeoDetect generalizes across multiple tasks and VLP architectures, and achieves strong detection performance against a range of state-of-the-art adversarial attacks. Notably, GeoDetect remains effective even under adaptive attack settings, where adversaries are aware of the detection strategy and attempt to bypass it. This robustness, combined with its independence from task-specific logits or labels, makes GeoDetect well-suited for both classification and retrieval scenarios.

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services, the Petascale Campus Initiative, and the Spartan HPC facilities. This facility was established with the assistance of LIEF Grant LE170100200. Moreover, this research was supported by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005), and partially funded by the Australian Government through the Australian Research Council.

References

1. Agrawal, A., Batra, D., Parikh, D., Kembhavi, A.: Don't just assume; look and answer: Overcoming priors for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
2. Aldahdooh, A., Hamidouche, W., Déforges, O.: Revisiting model's uncertainty and confidences for adversarial example detection. Applied Intelligence **53**(1), 509–531 (2023)
3. Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M.E., Kawarabayashi, K.i., Nett, M.: Estimating local intrinsic dimensionality. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015)

4. Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: A latent variable model approach to pmi-based word embeddings. Transactions of the Association for Computational Linguistics **4**, 385–399 (2016)
5. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Proceedings of the International Conference on Machine Learning (ICML) (2018)
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **35**(8), 1798–1828 (2013)
7. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European Conference on Computer Vision (ECCV) (2014)
8. Botev, Z.I., Grotowski, J.F., Kroese, D.P.: Kernel density estimation via diffusion. Annals of Statistics **38**(5), 2916–2957 (2010)
9. Bryniarski, O., Hingun, N., Pachuca, P., Wang, V., Carlini, N.: Evading adversarial example detection defenses with orthogonal projected gradient descent. In: Proceedings of the International Conference on Learning Representations (ICLR) (2022)
10. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
11. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: European Conference on Computer Vision (ECCV) (2020)
12. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (2011)
13. Cohen, G., Sapiro, G., Giryes, R.: Detecting adversarial samples using influence functions and nearest neighbors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
14. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13**(1), 21–27 (1967)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
16. Feinman, R., Curtin, R.R., Shintre, S., Gardner, A.B.: Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410 (2017)
17. Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion **91**, 424–444 (2023)
18. Grosse, K., Manoharan, P., Papernot, N., Backes, M., McDaniel, P.: On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)
19. Han, D., Jia, X., Bai, Y., Gu, J., Liu, Y., Cao, X.: OT-Attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. arXiv preprint arXiv:2312.04403 (2023)
20. He, B., Jia, X., Liang, S., Lou, T., Liu, Y., Cao, X.: SA-Attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. arXiv preprint arXiv:2312.04913 (2023)

21. Houle, M.E.: Dimensionality, discriminability, density and distance distributions. In: IEEE International Conference on Data Mining Workshops (2013)
22. Houle, M.E.: Local intrinsic dimensionality i: an extreme-value-theoretic foundation for similarity applications. In: Similarity Search and Applications (2017)
23. Houle, M.E., Kashima, H., Nett, M.: Generalized expansion dimension. In: IEEE International Conference on Data Mining Workshops (2012)
24. Karger, D.R., Ruhl, M.: Finding nearest neighbors in growth-restricted metrics. In: Proceedings of the Annual ACM Symposium on Theory of Computing (2002)
25. Kherchouche, A., Fezza, S.A., Hamidouche, W., Déforges, O.: Detection of adversarial examples in deep neural networks with natural scene statistics. In: International Joint Conference on Neural Networks (IJCNN) (2020)
26. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
27. Lee, J.M.: Riemannian manifolds: an introduction to curvature. Springer Science & Business Media (2006)
28. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems (NeurIPS) (2018)
29. Levi, M.Y., Gilboa, G.: The double-ellipsoid geometry of CLIP. In: Proceedings of the International Conference on Machine Learning (ICML) (2025)
30. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the International Conference on Machine Learning (ICML) (2022)
31. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems (NeurIPS) (2021)
32. Li, L., Guan, H., Qiu, J., Spratling, M.: One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
33. Li, L., Ma, R., Guo, Q., Xue, X., Qiu, X.: BERT-attack: Adversarial attack against BERT using BERT. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2020)
34. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems (NeurIPS) (2022)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
36. Lu, D., Wang, Z., Wang, T., Guan, W., Gao, H., Zheng, F.: Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
37. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems (NeurIPS) (2019)
38. Ma, X., Li, B., Wang, Y., Erfani, S.M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M.E., Bailey, J.: Characterizing adversarial subspaces using local intrinsic dimensionality. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018)

39. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018)
40. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. In: Proceedings of the International Conference on Learning Representations (ICLR) (2023)
41. McLachlan, G.J.: Mahalanobis distance. Resonance **4**(6), 20–26 (1999)
42. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in Neural Information Processing Systems (NeurIPS) (2022)
43. Mu, J., Viswanath, P.: All-but-the-top: Simple and effective postprocessing for word representations. In: Proceedings of the International Conference on Learning Representations (ICLR) (2018)
44. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML) (2021)
46. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: European Signal Processing Conference (EUSIPCO) (2007)
47. Schlarmann, C., Hein, M.: On the adversarial robustness of multi-modal foundation models. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW) (2023)
48. Schlarmann, C., Singh, N.D., Croce, F., Hein, M.: Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
49. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
50. Sotgiu, A., Demontis, A., Melis, M., Biggio, B., Fumera, G., Feng, X., Roli, F.: Deep neural rejection against adversarial examples. EURASIP Journal on Information Security **2020**, 1–10 (2020)
51. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations (ICLR) (2014)
52. Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A.: Robustness may be at odds with accuracy. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
53. Wang, L., Huang, J., Huang, K., Hu, Z., Wang, G., Gu, Q.: Improving neural language generation with spectrum control. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
54. Wang, S., Zhang, J., Yuan, Z., Shan, S.: Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
55. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: Proceedings of the International Conference on Learning Representations (ICLR) (2020)
56. Wang, Z., Li, X., Zhu, H., Xie, C.: Revisiting adversarial training at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

57. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **45**(10), 12113–12132 (2023)
58. Xu, X., Chen, X., Liu, C., Rohrbach, A., Darrell, T., Song, D.: Fooling vision and language models despite localization and attention mechanism. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
59. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
60. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
61. Yin, Z., Ye, M., Zhang, T., Du, T., Zhu, J., Liu, H., Chen, J., Wang, T., Ma, F.: VLATTACK: Multimodal adversarial attacks on vision-language tasks via pre-trained models. Advances in Neural Information Processing Systems (NeurIPS) (2023)
62. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
63. Zhang, C., Jin, M., Yu, Q., Liu, C., Xue, H., Jin, X.: Goal-guided generative prompt injection attack on large language models. In: International Conference on Data Mining (ICDM) (2024)
64. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: Proceedings of the International Conference on Machine Learning (ICML) (2019)
65. Zhang, J., Ma, X., Wang, X., Qiu, L., Wang, J., Jiang, Y.G., Sang, J.: Adversarial prompt tuning for vision-language models. In: European Conference on Computer Vision (ECCV) (2024)
66. Zhang, J., Yi, Q., Sang, J.: Towards adversarial attack on vision-language pre-training models. In: Proceedings of the ACM International Conference on Multimedia (ACM MM) (2022)
67. Zhang, Y., Xie, R., Chen, J., Sun, X., Wang, Y.: PIP: Detecting adversarial examples in large vision-language models via attention patterns of irrelevant probe questions. In: Proceedings of the ACM International Conference on Multimedia (ACM MM) (2024)
68. Zhao, D., Lin, Z., Tang, X.: Laplacian pca and its applications. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2007)
69. Zhou, Y., Xia, X., Lin, Z., Han, B., Liu, T.: Few-shot adversarial prompt learning on vision-language models. Advances in Neural Information Processing Systems (NeurIPS) (2024)
70. Zhou, Z., Hu, S., Li, M., Zhang, H., Zhang, Y., Jin, H.: AdvCLIP: Downstream-agnostic adversarial examples in multimodal contrastive learning. In: Proceedings of the ACM International Conference on Multimedia (ACM MM) (2023)