

# PIP: Detecting Adversarial Examples in Large Vision-Language Models via Attention Patterns of Irrelevant Probe Questions

Yudong Zhang  
Tsinghua University  
& Tencent  
Beijing, China  
zhangyd16@mails.tsinghua.edu.cn

Ruobing Xie\*  
Tencent  
Beijing, China  
xrbsnowing@163.com

Jiansheng Chen\*  
University of Science and Technology  
Beijing  
Beijing, China  
jschen@ustb.edu.cn

Xingwu Sun  
University of Macau  
& Tencent  
Beijing, China  
sunxingwu01@gmail.com

Yu Wang\*  
Tsinghua University  
Beijing, China  
yu-wang@mail.tsinghua.edu.cn

## Abstract

Large Vision-Language Models (LVLMs) have demonstrated their powerful multimodal capabilities. However, they also face serious safety problems, as adversaries can induce robustness issues in LVLMs through the use of well-designed adversarial examples. Therefore, LVLMs are in urgent need of detection tools for adversarial examples to prevent incorrect responses. In this work, we first discover that LVLMs exhibit regular attention patterns for clean images when presented with probe questions. We propose an unconventional method named PIP, which utilizes the attention patterns of one randomly selected irrelevant probe question (e.g., “Is there a clock?”) to distinguish adversarial examples from clean examples. Regardless of the image to be tested and its corresponding question, PIP only needs to perform one additional inference of the image to be tested and the probe question, and then achieves successful detection of adversarial examples. Even under black-box attacks and open dataset scenarios, our PIP, coupled with a simple SVM, still achieves more than 98% recall and a precision of over 90%. Our PIP is the first attempt to detect adversarial attacks on LVLMs via simple irrelevant probe questions, shedding light on deeper understanding and introspection within LVLMs. The code is available at <https://github.com/btzyd/pip>.

## CCS Concepts

• **Security and privacy** → *Intrusion detection systems*.

## Keywords

Large Vision-Language Model, Detecting Adversarial Example.

## ACM Reference Format:

Yudong Zhang, Ruobing Xie, Jiansheng Chen, Xingwu Sun, and Yu Wang. 2024. PIP: Detecting Adversarial Examples in Large Vision-Language Models

\*Corresponding author.



This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0686-8/24/10  
<https://doi.org/10.1145/3664647.3685510>

via Attention Patterns of Irrelevant Probe Questions. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3664647.3685510>

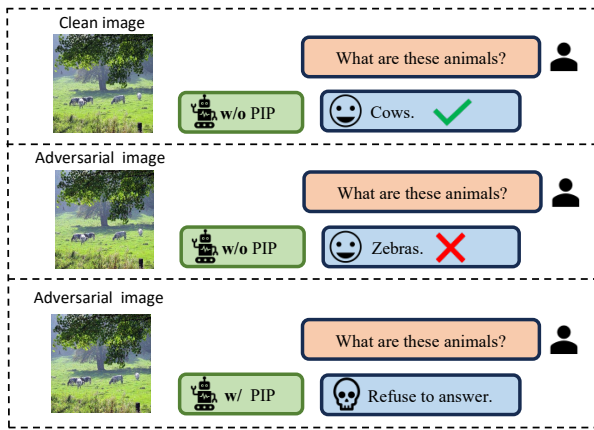
## 1 Introduction

Large vision-language models (LVLMs) have demonstrated their powerful multimodal capabilities across a range of tasks [1, 16, 28]. However, LVLMs continue to confront significant security challenges. Adversaries can perturb the images with elaborate and almost imperceptible noise, leading LVLMs to produce incorrect outputs or even outputs that align with the adversaries’ intentions [2, 4, 8, 24, 31, 34]. Significant security issues have impeded the adoption of LVLMs in critical areas [11, 17, 18, 26, 29, 32, 33].

Adversarial attacks on image modalities are facilitated by their high-dimensional and continuous nature. In recent years, there has been a surge in studies on adversarial attacks on LVLMs. AttackBard [8] manipulates images to make ChatGPT-4 errors on image caption, while Carlini [4] leads LVLMs to produce unethical words in response to adversarial images. In contrast to adversarial attacks, there has been limited research on the detection of adversarial examples. Therefore, there is an urgent need for detection methods for adversarial examples to enhance the safety of LVLMs.

Previous work [21, 22, 30] has primarily focused on the detection of adversarial examples in isolated vision models, *i.e.*, image classifiers based on convolutional neural networks (CNNs), which are not applicable to LVLMs. The detection of adversarial examples on LVLMs faces several challenges: First, LVLMs have more layers and contain interactions between image and text messages, making it difficult to determine where adversarial attacks take effect. Second, traditional CNNs have a more centralized data distribution (*e.g.*, imagenet classifiers detect imagenet adversarial samples), while LVLMs accept data inputs from open scenarios. In addition, LVLMs have a high inference cost and it is inappropriate to introduce too many inference times during the detection process. Few studies have focused on the detection of adversarial examples on LVLMs.

In this paper, we introduce an extremely simple method for the detection of adversarial examples, named **PIP**, that leverages the **attention pattern of irrelevant probe questions**. We initially discovered that for yes/no type questions, LVLM exhibits a *regular*



**Figure 1: Implications for our adversarial example detection method PIP. (Top):** LVLMs can give correct answers for clean images. **(Middle):** LVLMs may give incorrect answers for adversarial images. **(Down):** When detecting adversarial examples through our simple PIP, LVLMs reject answers for adversarial examples to prevent security risks.

*attention pattern* for image tokens when generating text output. Subsequently, we observed markedly distinct, linearly separable attention patterns between clean and adversarial images in response to the same simple yes/no probe questions. Building on this observation, we trained linear classifiers to distinguish between clean and adversarial examples, utilizing their attention patterns on one simple irrelevant probe question. Our PIP does not require the determination of whether an image should respond with “yes” or “no” to irrelevant probe questions, as our focus is exclusively on analyzing the attention patterns of the image in response to these questions, rather than the answers.

The workflow of PIP is illustrated in Fig. 2. We proposed our very simple and straightforward PIP detector could detect the sophisticated well-designed diverse adversarial examples of LVLMs. Our approach does not need to consider the distribution of the images, the type of the questions, the adversarial attack method, *etc.*, but only needs to ask LVLMs one probe question on examples, and trains an SVM [3] in the offline phase by using the attention patterns of clean and adversarial examples on the probe question. Surprisingly, it achieves the detection of adversarial examples for a wide range of questions, even if the probe question is randomly chosen and irrelevant. For the images entered by users in the online phase, the attention maps of images and the probe question are obtained by one additional inference, and the clean and adversarial examples can be distinguished via the lightweight SVM classifier.

Our main contributions can be summarized as follows:

- We have defined a new task: adversarial example detection for image adversarial attacks on LVLMs, which is essential in practice. We have also shown how to adapt the basic multimedia task evaluation metrics to this task.
- To the best of our knowledge, we are the first to introduce the simple and unconventional PIP, which detects adversarial examples based on the attention pattern of irrelevant probe questions.

Extensive experiments in different settings have validated the effectiveness, universality, and transferability of our PIP.

- PIP can also inspire subsequent work, such as defense and purification against samples. It can also help us understand LVLm in terms of deeper mechanisms.

## 2 Related Works

### 2.1 Large Vision-Language Models

Alignment-based vision-language models generally comprise three modules: a visual encoder, a large language model, and a vision-language alignment module. Generally, LVLms utilize pre-trained visual encoders and large language models, with the vision-language alignment module fine-tuned to enable multimodal capabilities. Recently, a lightweight vision-language alignment module, the Querying Transformer (Q-former), has gained popularity. Recent popular LVLms employing this alignment technique include BLIP-2 [13], InstructBLIP [6], and MiniGPT-4 [35].

### 2.2 Adversarial Attacks and Adversarial Examples

Adversarial attacks generate adversarial examples by introducing almost imperceptible perturbations to images, leading neural networks to respond incorrectly. Previous studies have focused on perturbations of visual modalities, including FGSM [10], PGD [20], JSMA [23], and C&W attack [5], among others. Research has also been conducted on attacks targeting textual modalities, such as Bert-Attack [14] and TextFooler [12]. Large vision-language models are also vulnerable to adversarial examples. For instance, AttackBard [8] induces incorrect captions in Google’s Bard and OpenAI’s ChatGPT-4 through black-box attacks on the image, while Carlini *et al.* [4] prompt LVLms to generate inappropriate speech (*e.g.*, profanity, biased statements, *etc.*) via white-box attacks on the image.

### 2.3 Detecting Adversarial Examples

Since the proposal of adversarial attacks, significant research has been devoted to the detection of adversarial examples, aiming to alert the model to the presence of such examples. This is particularly critical in LVLms because deliberately designed adversarial examples can cause LLMs to generate outputs aligned with the adversary’s intentions, potentially leading to robustness and hallucination issues. More gravely, the model may produce statements that are socially and morally reprehensible.

Previous research has concentrated on detecting adversarial examples for visual modality CNNs, as evidenced by [7, 9, 15, 19]. However, research on detecting adversarial examples in vision-language multimodal models is lacking. This represents a considerable risk when employing LVLms in sensitive domains. LVLms require the capability to detect adversarial examples and, in response, should either refuse to answer or neutralize the adversarial input to provide a clean response.

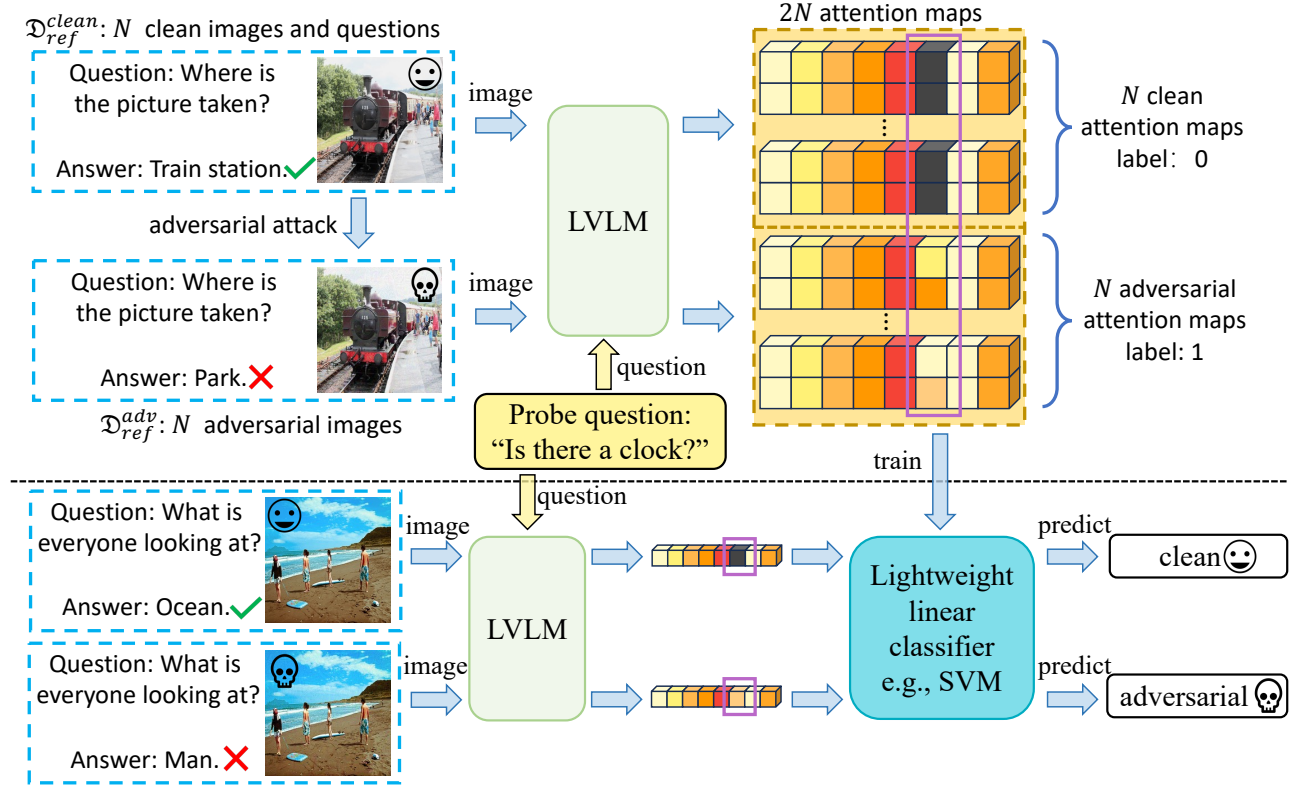


Figure 2: The pipeline of our proposed PIP. The top half is operated offline, while the bottom half is operated online. (Top): We perform adversarial attacks on  $N$  images in  $\mathcal{D}_{ref}^{clean}$  and obtain  $N$  adversarial images, which constitute  $\mathcal{D}_{ref}^{adv}$ . We extract their attention maps (attention of the first word generated by LLM to all image tokens) of the LVLM with the irrelevant probe question “Is there a clock?”, and train a lightweight linear classifier (e.g., SVM) with these  $2N$  attention maps. (Down): For images to be tested from  $\mathcal{D}_{test}$ , we first get their attention maps with the same probe question, and use the classifier to determine whether they are adversarial examples or not. Surprisingly, this simple method PIP functions well in this challenging task.

### 3 A New Task: Detecting Adversarial Examples in Large Vision-Language Models

#### 3.1 Definition of Our Adversarial Examples Detection Task

In this section, we introduce a novel task aimed at detecting adversarial examples in LVLMs.

Let  $f$  represent a large vision-language model, which takes an image  $I$  and a question  $Q$  as input, and produces an answer  $A = f(I, Q)$ . The task will provide two datasets,  $\mathcal{D}_{ref}$  and  $\mathcal{D}_{test}$ , where  $\mathcal{D}_{ref}$  comprises  $N$  samples from a large clean dataset  $\mathcal{D}$ , serving as a reference for clean examples. Additionally, we randomly select  $M$  images and questions from  $\mathcal{D}$  to create the dataset  $\mathcal{D}_{clean}$ , and then execute adversarial attacks on these  $M$  images to generate the dataset  $\mathcal{D}_{adv}$ . By randomly combining clean and adversarial examples in a ratio of  $M_{clean} : M_{adv}$ , we generate the test dataset  $\mathcal{D}_{test}$ . The task’s objective is to train a classifier  $h(I)$  that discerns whether each  $I \in \mathcal{D}_{test}$  originates from the clean dataset  $\mathcal{D}_{clean}$  or the adversarial dataset  $\mathcal{D}_{adv}$ . Note that for this task, only  $\mathcal{D}_{ref}$ ,  $f$ , and the test dataset  $\mathcal{D}_{test}$  are provided. If  $I$  originates from  $\mathcal{D}_{clean}$ , the ground truth  $g(I)$  should be 1, otherwise 0, as delineated in

Eq. (1).

$$g(I) = \begin{cases} 1 & \text{if } I \in \mathcal{D}_{adv} \\ 0 & \text{if } I \in \mathcal{D}_{clean} \end{cases} \quad (1)$$

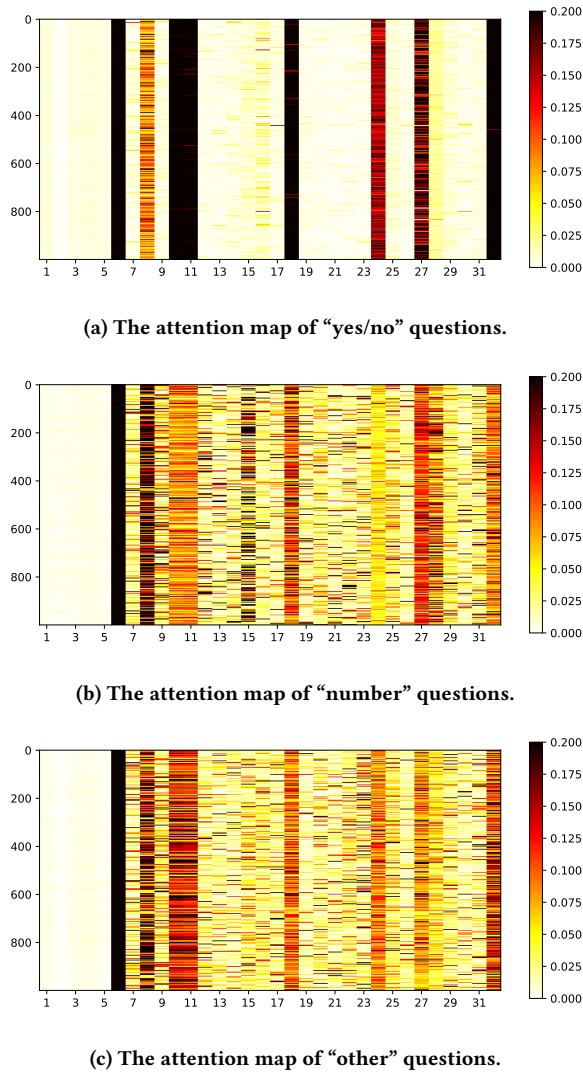
The adversarial attacks on LVLMs are more diverse in terms of attack methods and targets, and more challenging to detect.

#### 3.2 Evaluation of Our Detection Task

To evaluate the performance of the classifier  $h$ , which determines the likelihood of  $I$  being an adversarial example, we employ the metrics outlined in Eq. (2).

$$\begin{aligned} TP &= \sum_{I \in \mathcal{D}_{test}} g(I) \cdot h(I), & FN &= M_{adv} - TP \\ FP &= \sum_{I \in \mathcal{D}_{test}} (1 - g(I)) \cdot h(I), & TN &= M_{clean} - FP \end{aligned} \quad (2)$$

Upon defining True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), we compute the precision, recall, accuracy, and F1-score to comprehensively assess the performance of classifier  $h$ .



**Figure 3: The attention maps of different types of questions on 1,000 randomly-selected images and questions. Due to space limitations, we select only one layer (the 16th layer of the LLM) and display the maximum value in the multi-head attention. The attention map of (a) “yes/no” is more regular than (b) “number” and (c) “other”, indicating that the simple “yes/no” is a more suitable probe question.**

## 4 Explore the Use of Our PIP to Detecting Adversarial Examples

### 4.1 LVLMs Have Regularized Attention Patterns of Clean Examples to Yes/No Questions

We begin by selecting a popular LVM InstructBLIP Vicuna-7B (decoder-only) to explore attention patterns, then observe other LVLMs if there is a similar phenomenon. It is based on the Q-former architecture, and the LLM receives 32 image tokens and  $P$  question tokens as inputs. Initially, the LLM computes self-attention on the

$32 + P$  tokens, followed by the insertion of the start token  $eos_{start}$  into the LLM to generate the output sequence sequentially. We examine the attention map from the first token following the start token to the 32 image tokens. The attention map has a dimension of  $[32, 32, 32]$ , with the first 32 representing the number of layers, the second 32 indicating the number of multi-heads, and the last 32 denoting the index of image tokens. To display the attention map of multiple images within a single figure, we select a certain layer and focus on the largest head within the multi-head attention.

We randomly selected 1000 questions of each type (yes/no, number, and other) from VQA v2. For each of these question types, we generated the attention maps for layer 16 of the LLM in InstructBLIP Vicuna-7B, as illustrated in Fig. 3, where each row is an image and each column is the 32 image tokens fed into the LLM, with the color indicating attention values. The attention map for yes/no questions exhibits a more regular pattern compared to those for number and other question types. This insight offers a novel perspective on distinguishing between clean and adversarial examples through the analysis of regular attention patterns.

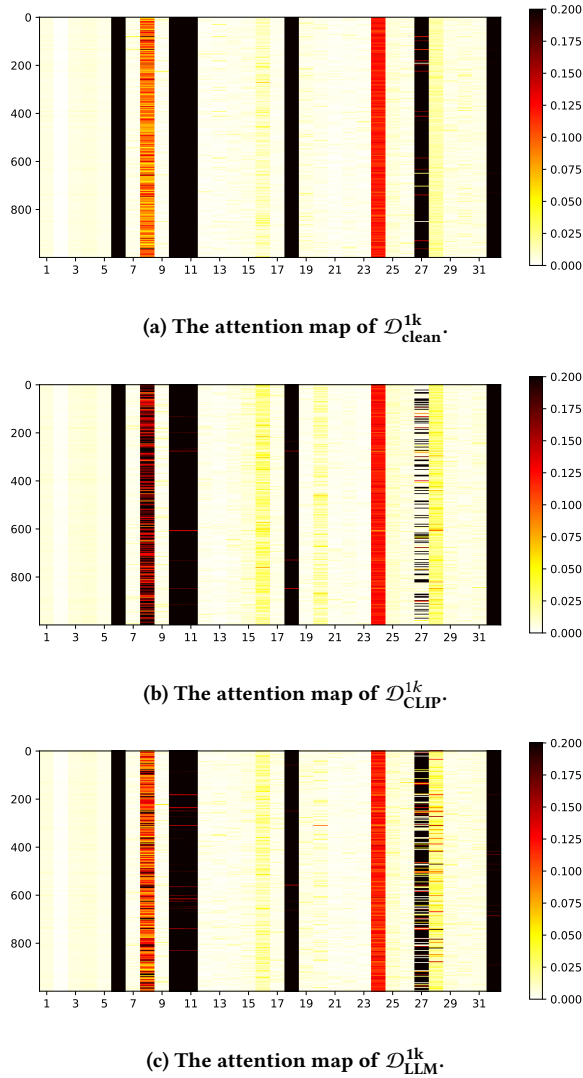
### 4.2 The Attention Patterns of “yes/no” Probe Questions between Clean and Adversarial Examples are Clearly Distinguishable

As previously identified in Sec. 4.1, yes/no type questions exhibit regular attention patterns. This section explores the differences in attention patterns between clean and adversarial examples for yes/no questions.

We randomly selected 1000 images and questions from dataset  $\mathcal{D}$  to create dataset  $\mathcal{D}_{clean}^{1k}$ . Using PGD on the white-box model, we attack  $\mathcal{D}_{clean}^{1k}$  to generate the adversarial example datasets  $\mathcal{D}_{CLIP}^{1k}$  and  $\mathcal{D}_{LLM}^{1k}$ . For the PGD attack, the number of steps was set to 20, with a perturbation size per step of  $\alpha = 2/255$ , and a total perturbation limit of  $\epsilon_{\infty} = 8/255$ . We consider two methods of attack. For  $\mathcal{D}_{LLM}^{1k}$ , an untargeted attack was conducted on the LLM’s logit cross-entropy. Obtaining the LLM component of an LVM can be challenging, whereas accessing its visual encoder (typically CLIP [25] or EVA-CLIP [27]) is comparatively easier for an adversary. Therefore, for  $\mathcal{D}_{CLIP}^{1k}$ , the mean square error (MSE) loss function was employed to conduct an untargeted attack on the output feature of the CLIP or EVA-CLIP visual encoder.

We randomly select a yes/no question as the probe question, for example, “Is there a clock?”. Images from  $\mathcal{D}_{clean}^{1k}$ ,  $\mathcal{D}_{CLIP}^{1k}$  and  $\mathcal{D}_{LLM}^{1k}$ , along with the probe question, were fed into InstructBLIP Vicuna-7B, with their attention maps displayed in Fig. 4. Clearly, Figure 4a exhibits significant differences when compared to Fig. 4b and Fig. 4c. Specifically, the 27th token in Fig. 4a is predominantly black, whereas Figure 4c shows a reduced percentage of black, and Fig. 4b contains very little black. Furthermore, the attention for the 28th token in Fig. 4b and Fig. 4c is more pronounced than in Fig. 4a.

Although Fig. 4 illustrates results from just one layer of the LLM, a significant difference between the clean and adversarial examples is already evident. A similar phenomenon is widely observed across other layers of the LLM. It indicates that the attention patterns of clean and adversarial examples are straightforward and likely to



**Figure 4: The attention maps of  $\mathcal{D}_{\text{clean}}^{1k}$ ,  $\mathcal{D}_{\text{CLIP}}^{1k}$  and  $\mathcal{D}_{\text{LLM}}^{1k}$ .** Due to space limitations, we select only one layer (the 16th layer of the LLM) and take the maximum value in the multi-head attention. The probe questions are all “Is there a clock?”. The attention maps of adversarial examples differed significantly from those of the clean examples on certain sensitive tokens (e.g., the 27th and 28 tokens), which are good indicators.

be linearly separable, thus motivating the usage of simple linear classifiers for discrimination as our detector.

### 4.3 Distinguishing Attention Maps via Lightweight Support Vector Machine

As discussed in Sec. 4.2, the use of an irrelevant probe question has been identified as a viable method for distinguishing between clean

---

#### Algorithm 1: The steps of PIP using SVM.

---

**Input:** Clean reference dataset  $\mathcal{D}_{\text{ref}}^{\text{clean}}$ ,  $\mathcal{D}_{\text{test}}$  to be tested  
**Model:** The large vision-language model  $f$ , where  $f_{\text{att}}(I, Q)$  outputs the LLM’s attention map  
**Data:** The irrelevant probe question  $Q_p$   
**Output:** The predict results of  $\mathcal{D}_{\text{test}}$ , where 0 denotes clean example and 1 denotes adversarial examples

- 1  $\mathcal{D}_{\text{ref}}^{\text{adv}} = \{\};$
- 2 **for** image  $I_j$  and question  $Q_j \in \mathcal{D}_{\text{ref}}^{\text{clean}}$  **do**
  - /\* AA denotes a adversarial attack which returns the adversarial image \*/
  - 3  $I'_j = \text{AA}(I_j, Q_j);$
  - 4  $\mathcal{D}_{\text{ref}}^{\text{adv}} = \mathcal{D}_{\text{ref}}^{\text{adv}} + \{I'_j\};$
- 5 **end**
  - /\* The train data ( $X$ ) and label ( $y$ ) of SVM \*/
  - 6  $X = \{\};$
  - 7  $y = \{\};$
  - 8 **for** image  $I_j \in \{\mathcal{D}_{\text{ref}}^{\text{clean}} + \mathcal{D}_{\text{ref}}^{\text{adv}}\}$  **do**
    - 9  $X = X + \{f_{\text{att}}(I_j, Q_p)\};$
    - /\* Adversarial examples are labeled 1, otherwise 0 \*/
    - 10  $y = y + \{\mathbb{I}(I_j \in \mathcal{D}_{\text{ref}}^{\text{adv}})\};$
  - 11 **end**
    - /\* Train the classifier  $h$  with SVM \*/
    - 12  $h = \text{SVM}(X, y);$
    - /\* Using the classifier  $h$  to predict \*/
    - 13  $Y_{\text{pred}} = \{\};$
    - 14 **for**  $I_j \in \mathcal{D}_{\text{test}}$  **do**
      - 15  $Y_{\text{pred}} = Y_{\text{pred}} + \{h(f_{\text{att}}(I_j, Q_p))\};$
    - 16 **end**
    - 17 **return**  $Y_{\text{pred}};$

---

and adversarial examples. A simple linear classifier, precisely a support vector machine (SVM) [3], has been employed to differentiate between clean and adversarial examples.

The steps of our PIP method using SVM are outlined in Algorithm 1 and Fig. 2. We conducted an attack using PGD on the cross-entropy loss function of the LLM’s logit output on  $\mathcal{D}_{\text{ref}}^{\text{clean}}$ , resulting in the dataset  $\mathcal{D}_{\text{ref}}^{\text{adv}}$ . These clean and adversarial images, along with an irrelevant probe question, were fed into the LLM to extract their attention maps. Utilizing these attention maps, we trained a support vector machine (SVM)  $h$  to discern whether the tested examples are adversarial or not.

Acknowledging that adversarial examples are likely less prevalent than clean examples in practical scenarios (due to fewer malicious users), the mixing ratios  $M_{\text{clean}} : M_{\text{adv}}$  were established at 1000 : 100 and 1000 : 1000. The clean reference dataset  $\mathcal{D}_{\text{ref}}$  comprises 5000 samples ( $N = 5000$ ). The PGD attack encompasses a 20-step iteration, with a step learning rate of  $2/255$  and an overall perturbation limit of  $\epsilon_{\infty} = 8/255$ .

**Table 1: Results of using PIP with SVM to detect adversarial examples. In this table,  $\mathcal{D}_{ref}$  and  $\mathcal{D}_{test}$  are from the same COCO dataset.**

Attack	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
$\mathcal{D}_{adv}^{CLIP}$	1000/1000	90.91	100.00	95.00	95.24
	1000/100	50.00	100.00	90.91	66.67
$\mathcal{D}_{adv}^{LLM}$	1000/1000	98.97	96.50	97.75	97.72
	1000/100	90.48	95.00	98.64	92.68

The results of detecting adversarial examples using PIP with SVM are presented in Tab. 1. For  $\mathcal{D}_{adv}^{LLM}$ , given that the SVM was trained while also attacking the output of LLM, it demonstrates excellent capability in recognizing adversarial examples, achieving very high performance. For  $\mathcal{D}_{adv}^{CLIP}$ , despite training the SVM with a different adversarial attack, it remains capable of recognizing adversarial examples. However,  $\mathcal{D}_{adv}^{CLIP}$  exhibits a high false alarm rate (10%) due to the disparity in attack methods used during training SVM and testing examples. With a clean-to-adversarial example ratio of 10 : 1, examples with false alarms constitute half of the total alarms, resulting in reduced precision. It is astonishing that our simple PIP achieves impressive results in adversarial example detection.

#### 4.4 Exploring the PIP’s Decision-making Process with Decision Trees

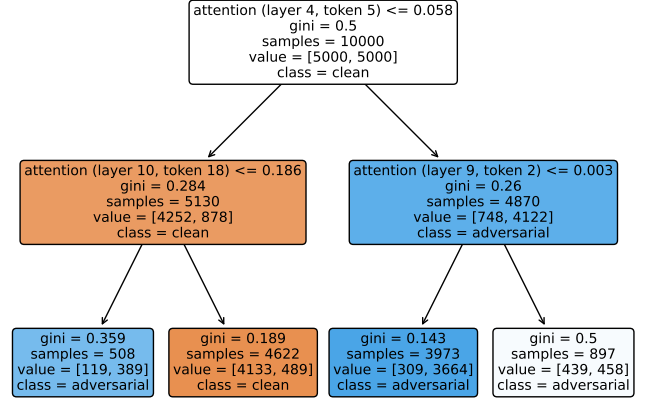
In previous sections, PIP with SVM was used to detect adversarial samples. However, visualizing the SVM’s decision-making process is challenging due to the high-dimensional space of the attention maps. In this section, the decision tree (DT) is used as an intuitive alternative to SVM.

Decision tree operates by recursively partitioning a dataset into increasingly smaller subsets and facilitates the generation of easily understandable rules. To visualize the decision-making process, a DT with a depth of 2 is trained. The input for the DT consists of attention maps with 1024 dimensions (32 layers  $\times$  32 tokens, with the maximum value in the multi-head attention dimension). The DT(depth=2) makes a linear decision based on two dimensions, as illustrated in Fig. 5. Despite its limited depth of 2, the DT successfully detects adversarial examples, as evidenced in Tab. 2. Increasing the DT’s depth could enhance its performance. However, this work only focuses on visualizing the decision-making process using DT(depth=2) to demonstrate PIP’s effectiveness.

### 5 In-depth Analyses on PIP

Having initially validated PIP’s effectiveness, we then explored its generalizability across various settings and endeavored to understand and analyze our detection method and the intrinsic mechanisms underlying the differences in attention patterns.

We obtained 5,000  $\mathcal{D}_{ref}^{adv}$  images from adversarial attacks on  $\mathcal{D}_{ref}^{clean}$ . Afterward, an SVM was trained using the attention pattern of 10,000 images on the probe question. In Sec. 5.1, Sec. 5.2 and Sec. 5.3, we try the results of this SVM on different test data distributions (training SVM on COCO while test adversarial examples on ImageNet), attack methods (PGD to C&W), and attack parameters



**Figure 5: PIP with the decision-making process of decision trees. the DT(depth=2) linearly distinguishes between clean and adversarial examples only by the two feature dimensions of the attention maps.**

( $\epsilon_\infty$ ). Section 5.4 explores the results of PIP on other LVLMs (different models require different SVMs to be trained because of different attention patterns), and Section 5.5 explores the use of multiple SVM fusion decisions for the same model. Finally, we explore a real-world black-box attack scenario in Sec. 5.6. It is important to emphasize that for all experiments (except Sec. 5.5), we trained the SVMs using the **same data** (PGD  $\epsilon_\infty = 8/255$  on 5,000 images from COCO  $\mathcal{D}_{ref}$  targets the output of LLM) and the **same probe question** “Is there a clock?” to extract attention maps (Section 5.5 used more probe questions than one). Note that using the same unified PIP detector for different settings is extremely challenging.

#### 5.1 Generalization of Our Adversarial Examples Detection Method across Datasets

In Sec. 4.3, a portion of the adversarial examples was successfully detected in the dataset  $\mathcal{D}_{test}$ . However, there exists a potential issue as both  $\mathcal{D}_{ref}$  and  $\mathcal{D}_{test}$  derive from  $\mathcal{D}$  (i.e., COCO dataset). In practical applications, access to the dataset of user input images may be unavailable, making it crucial to assess the generalization performance when  $\mathcal{D}_{ref}$  and  $\mathcal{D}_{test}$  do not align.

In this section,  $\mathcal{D}_{test}$  is replaced with the ImageNet dataset, while maintaining  $\mathcal{D}_{ref}$  as is. Specifically, an image from each of ImageNet’s 1000 classes was randomly selected, and its label was used to generate a corresponding question using the template “Is there a/an {label}?”. This process forms the dataset  $\mathcal{D}_{clean}$ , and then, following the previous method, these 1000 images and questions were attacked to create the dataset  $\mathcal{D}_{adv}$ .  $\mathcal{D}_{clean}$  and  $\mathcal{D}_{adv}$  were mixed according to  $M_{clean} : M_{adv}$ , with the detection of adversarial examples performed using the method described in Sec. 4.3. The results are presented in Tab. 3. Despite the inconsistency between  $\mathcal{D}_{ref}$  (from COCO) and  $\mathcal{D}_{test}$  (from ImageNet-1K), PIP maintains high accuracy (over 90%). This indicates that PIP’s effectiveness is

**Table 2: Results of using PIP with DT(depth=2) to detect adversarial examples. The purpose of this table is not performance, but to show that PIP does indeed find linearly separable features between clean and adversarial examples, even through DT(depth=2). Its decision-making process is shown in Fig. 5.**

Attack	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
$\mathcal{D}_{adv}^{LLM}$	1000/1000	85.74	90.20	87.60	87.91
	1000/100	37.76	91.00	85.55	53.37

**Table 3: Results of using PIP with SVM to detect adversarial examples. In this table,  $\mathcal{D}_{ref}$  and  $\mathcal{D}_{test}$  are from COCO and ImageNet-1K, respectively.**

Attack	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
$\mathcal{D}_{adv}^{CLIP}$	1000/1000	90.91	100.00	95.00	95.24
	1000/100	50.00	100.00	90.91	66.67
$\mathcal{D}_{adv}^{LLM}$	1000/1000	90.37	93.80	91.90	92.05
	1000/100	48.72	95.00	90.45	64.41

**Table 4: Generalizability of PIP over different attack method C&W. In this table, we only consider attacks the LLM’s outputs (i.e.,  $\mathcal{D}_{adv}^{LLM}$ ).**

$\mathcal{D}_{test}$	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
COCO	1000/1000	98.99	97.80	98.40	98.39
	1000/100	90.83	99.00	99.00	94.74
ImageNet-1K	1000/1000	90.55	95.80	92.90	93.10
	1000/100	49.24	97.00	90.64	65.32

not contingent on a specific dataset distribution and can be generalized across different datasets. In a practical scenario,  $\mathcal{D}_{ref}$  can be generated from a given dataset, with PIP employed to detect adversarial examples in  $\mathcal{D}_{test}$ , irrespective of  $\mathcal{D}_{test}$ ’s distribution during the generation of  $\mathcal{D}_{ref}$ .

## 5.2 Generalizability of PIP over Different Attack Methods

In prior experiments, the generalizability of PIP across various attack targets, such as  $\mathcal{D}_{adv}^{CLIP}$  and  $\mathcal{D}_{adv}^{LLM}$  was established. However, these experiments were exclusively based on the PGD attack method. The ability of SVMs trained on PGD-generated adversarial examples to generalize to unknown attack methods in real-world scenarios remains a critical aspect of PIP’s utility.

This section examines the effectiveness of PIP against another popular attack method C&W beyond PGD. For the C&W attack, the process iterated over 50 steps with a perturbation of 0.01 per step, equivalent to 2.55/255. Table 4 displays the experimental results. Even against unknown attacks (C&W) and on datasets distinct from  $\mathcal{D}_{ref}$  (ImageNet-1K), PIP demonstrates superior performance.

**Table 5: Generalizability of PIP over different methods than PGD attacks. In this table, we only consider attacks the LLM’s outputs (i.e.,  $\mathcal{D}_{adv}^{LLM}$ ) on COCO.**

$\epsilon_{\infty}$	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
2/255	1000/1000	97.45	38.20	68.60	54.89
	1000/100	80.00	40.00	93.64	53.33
4/255	1000/1000	98.81	82.80	90.90	90.10
	1000/100	89.25	83.00	97.55	86.01
16/255	1000/1000	98.99	98.50	98.75	98.75
	1000/100	90.83	99.00	99.00	94.74

## 5.3 Generalization Performance of PIP on Different Attack Parameters

In the training process of SVM and prior experiments involving PGD attacks, a consistent perturbation strength of  $\epsilon_{\infty} = 8$  was employed. The aim is to ascertain PIP’s generalizability across weaker and stronger PGD attacks. Table 5 examines PIP’s efficacy on  $\mathcal{D}_{test}$  examples under three perturbation levels  $\epsilon_{\infty} = \{2, 4, 16\}$ , revealing that PIP maintains a 40% recall even with a minimal perturbation of  $\epsilon_{\infty} = 2$ . With an  $\epsilon_{\infty} = 4$  attack, PIP achieves over 80% recall. Under a significantly strong attack of  $\epsilon_{\infty} = 16$ , PIP is able to detect nearly all adversarial examples. Table 5 demonstrates PIP’s generalizability to both weaker and stronger attacks.

## 5.4 Results of PIP on other LVLMs

In Tab. 1 and Tab. 3, InstructBLIP Vicuna-7B is utilized as the LVLm. Assessing whether other LVLms can employ PIP to detect adversarial examples is crucial for evaluating PIP’s generalizability. This section examines other versions of BLIP-2 and InstructBLIP to evaluate PIP’s performance on various LVLms. For each model, PIP initially extracts the attention maps from  $\mathcal{D}_{ref}$ , followed by SVM training. PIP’s transferability across LVLms is impossible due to varying attention map configurations and dimensions among LVLms, such as [32, 32] for Vicuna-7B, [40, 40] for Vicuna-13B, [24, 32] for FlanT5<sub>XL</sub>, [24, 64] for FlanT5<sub>XL</sub>, with the first number representing the LLM’s layers and the second the multi-head attention count. Therefore, a separate PIP detector must be trained on each LVLm. For BLIP-2 and InstructBLIP, the analysis includes both decoder-only LLMs (e.g., OPT, Vicuna) and encoder-decoder LLMs (e.g., FlanT5). In decoder-only LLMs, we focus on the attention maps of 32 image tokens during the generation of the first word. In encoder-decoder LLMs, we focus on the cross-attention maps of 32 image tokens at the first word’s generation. Within the multi-head attention dimension, only the head with the highest attention is selected to reduce the attention map dimensions. Table 6 presents the results of applying PIP with SVM across a broader range of LVLms and demonstrates the effectiveness of PIP on different LVLms.

## 5.5 Alleviating the High False Alarm Rate Issue

The primary advantage of PIP is its high recall rate, which is critical for LVLms. Adversaries can manipulate the model using adversarial examples, potentially leading to significant public opinion risks if the model generates content that violates morality or law. However,

**Table 6: Results of PIP on other LVLMS besides InstructBLIP Vicuna-7B. In this table, we only consider attacks the LLM’s outputs (i.e.,  $\mathcal{D}_{adv}^{LLM}$ ) on COCO.**

LVLM	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
InstructBLIP (FlanT5 <sub>XL</sub> )	1000/1000	96.19	96.00	96.10	96.10
	1000/100	71.85	97.00	96.27	82.55
InstructBLIP (FlanT5 <sub>XXL</sub> )	1000/1000	97.48	96.90	97.20	97.19
	1000/100	79.67	98.00	97.55	87.89
InstructBLIP (Vicuna-13B)	1000/1000	98.79	98.20	98.50	98.50
	1000/100	89.19	99.00	98.82	93.84
BLIP-2 (OPT-2.7B)	1000/1000	96.36	95.20	95.80	95.77
	1000/100	72.93	97.00	96.45	83.26
BLIP-2 (OPT-6.7B)	1000/1000	95.78	95.40	95.60	95.59
	1000/100	69.12	94.00	95.64	79.66
BLIP-2 (FlanT5 <sub>XL</sub> )	1000/1000	94.73	91.70	93.30	93.19
	1000/100	64.58	93.00	94.73	76.23

**Table 7: Results of PIP with multiple SVMs. AR (Alarm Rules)  $i/j$  means that among  $j$  SVMs of  $j$  probe questions, the alarmed SVMs  $\geq i$  are determined as adversarial examples.**

$\mathcal{D}_{test}$	AR	Attack	$M_{clean}/M_{adv}$	Precision	Recall	Accuracy	F1-score
COCO	2/3	$\mathcal{D}_{adv}^{CLIP}$	1000/1000	98.81	100.00	99.40	99.40
			1000/100	89.29	100.00	98.91	94.34
	$\mathcal{D}_{adv}^{LLM}$	1000/1000	98.78	96.80	97.80	97.78	
		1000/100	88.99	97.00	98.64	92.82	
IN-1K	3/3	$\mathcal{D}_{adv}^{CLIP}$	1000/1000	94.61	100.00	97.15	97.23
			1000/100	63.69	100.00	94.82	77.82
	$\mathcal{D}_{adv}^{LLM}$	1000/1000	94.06	90.30	92.30	92.14	
		1000/100	61.74	92.00	94.09	73.90	

as indicated in Tab. 1, PIP experiences a high false alarm rate (about 10%) when the attack method used to train the SVM differs from that used on the test examples. This false alarm rate is particularly high when the ratio of clean to adversarial examples is 1000 : 100, resulting in half of the alarms being false. Such a high false alarm rate could jeopardize PIP’s practical application, as normal user images may be incorrectly flagged, degrading the user experience.

Our fundamental approach involves conducting a focused review of samples triggering alerts. A straightforward method is to apply three instead of one probe question to these targeted samples. In scenarios with low false alarm rates, like in Tab. 1, an image is classified as an adversarial example if it triggers at least two alarms out of three probe questions. In cases with high false alarm rates, such as in Tab. 3, an image is deemed an adversarial example only if alarms are triggered on all three probe questions. Besides the existing probe question “Is there a clock?”, two additional questions, “Is this in the United States?” and “Is this photo an action shot?” were randomly selected, and two more SVMs were trained accordingly.

The outcomes are detailed in Tab. 7, which, compared to Tab. 1 and Tab. 3, exhibits a notable reduction in the false alarm rate, as indicated by a substantial increase in precision. Moreover, Table 7 maintains an exceptionally high recall rate, thus reducing the false alarm rate while preserving the detection of adversarial examples.

**Table 8: Results of PIP on detecting adversarial examples generated by black-box Attack-Bard on NIPS2017 dataset. The  $M_{clean}/M_{adv}$  is 200 : 200.**

LVLM	Precision	Recall	Accuracy	F1-score
InstructBLIP FlanT5 <sub>XL</sub>	90.09	100.00	94.50	94.79
InstructBLIP FlanT5 <sub>XXL</sub>	90.32	98.00	93.75	94.00
InstructBLIP Vicuna-7B	94.79	100.00	97.25	97.32
InstructBLIP Vicuna-13B	96.10	98.50	97.25	97.28

Employing additional SVMs will increase computational demands. For a user-input image, one SVM judgment necessitates only one extra inference beyond the user-input question to derive the attention map, while three SVM judgments necessitate three extra inferences. Fortunately, not all images require multiple inferences. It suffices to infer all images once and selectively focus on certain suspicious images.

## 5.6 Generalization Performance of PIP on Actual Black-box Attacks

In earlier experiments, adversarial examples were generated through the white-box attacks. However, in real-world scenarios, obtaining the model’s weights and executing white-box attacks is challenging for users. Consequently, numerous studies have investigated black-box attacks. Exploring PIP’s effectiveness against black-box attack-generated adversarial examples is worthwhile.

Attack-Bard [8] used black-box attacks on the NIPS2017 dataset to generate adversarial examples, successfully compromising major commercial models like ChatGPT-4V, Google’s Bard, Bing Chat, and ERNIE Bot (with approximately 45% attack success on ChatGPT-4V). This section involves selecting 200 original images from the NIPS2017 dataset along with corresponding adversarial images generated by Attack-Bard to create the dataset  $\mathcal{D}_{test}$ , using PIP to detect adversarial examples on InstructBLIP.

This constitutes a comprehensive evaluation of PIP, as in this experimental setup, the constant factor is the model used in both training and detection phases (necessary due to varying attention patterns across models). Beyond this, PIP remains uninformed about other aspects like the distribution of user input data, attack methods, parameters, targets, and the models used for the attack. Table 8 affirms PIP’s generalizability in authentic black-box attack scenarios. In the context of black-box attacks, PIP maintains a recall rate exceeding 95% and a precision greater than 80%.

## 6 Conclusion

In this paper, we introduce PIP, a new and simple method for detecting adversarial examples in LVLMS. Although PIP is simple and whimsical, it achieves impressive results on recent, popular LVLMS like BLIP-2 and InstructBLIP, achieving high recall rates of adversarial examples with low false alarms among clean examples. For detected adversarial examples, post-processing measures such as focusing on alert examples, denying answers, and implementing adversarial defenses can enhance the security and robustness of LVLMS, thereby mitigating public and legal risks.



## Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376024, 62171313), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001) and Beijing National Research Center for Information Science and Technology (BNRist).

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [2] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. 2023. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236* (2023).
- [3] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. 144–152.
- [4] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. 2024. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems* 36 (2024).
- [5] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 39–57.
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).
- [7] Sumanth Dathathri, Stephan Zheng, Tianwei Yin, Richard M Murray, and Yisong Yue. 2018. Detecting adversarial examples via neural fingerprinting. *arXiv preprint arXiv:1803.03870* (2018).
- [8] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How Robust is Google's Bard to Adversarial Image Attacks? *arXiv preprint arXiv:2309.11751* (2023).
- [9] Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. 2021. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*. PMLR, 3564–3575.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
- [11] Yihao Huang, Liangru Sun, Qing Guo, Felix Juefei-Xu, Jiayi Zhu, Jincao Feng, Yang Liu, and Geguang Pu. 2023. ALA: Naturalness-aware Adversarial Lightness Attack. In *Proceedings of the 31st ACM International Conference on Multimedia (Ottawa ON, Canada) (MM '23)*. Association for Computing Machinery, New York, NY, USA, 2418–2426. <https://doi.org/10.1145/3581783.3611914>
- [12] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932* 2 (2019), 10.
- [13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [14] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984* (2020).
- [15] Xin Li and Fuxin Li. 2017. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE international conference on computer vision*. 5764–5772.
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [17] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 102–111.
- [18] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. 2023. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*.
- [19] Chen Ma, Chenxu Zhao, Hailin Shi, Li Chen, Junhai Yong, and Dan Zeng. 2019. Metaadvdet: Towards robust detection of evolving adversarial attacks. In *Proceedings of the 27th ACM International Conference on Multimedia*. 692–701.
- [20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
- [21] Dongyu Meng and Hao Chen. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. 135–147.
- [22] Tianyu Pang, Chao Du, Yinpeng Dong, and Jun Zhu. 2018. Towards robust detection of adversarial examples. *Advances in neural information processing systems* 31 (2018).
- [23] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 372–387.
- [24] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 21527–21536.
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [26] Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, Huimin Ma, and Yong Li. 2023. Enhancing Adversarial Robustness of Multi-modal Recommendation via Modality Balancing. In *Proceedings of the 31st ACM International Conference on Multimedia (Ottawa ON, Canada) (MM '23)*. Association for Computing Machinery, New York, NY, USA, 6274–6282. <https://doi.org/10.1145/3581783.3612337>
- [27] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).
- [28] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems* 36 (2024).
- [29] Yuxuan Wang, Jiakai Wang, Zixin Yin, Ruihao Gong, Jingyi Wang, Aishan Liu, and Xianglong Liu. 2022. Generating Transferable Adversarial Examples against Vision Transformers. In *Proceedings of the 30th ACM International Conference on Multimedia (Lisboa, Portugal) (MM '22)*. Association for Computing Machinery, New York, NY, USA, 5181–5190. <https://doi.org/10.1145/3503161.3547989>
- [30] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).
- [31] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. Vllattck: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *arXiv preprint arXiv:2310.04655* (2023).
- [32] Jiaming Zhang, Qi Yi, and Jitao Sang. 2022. Towards Adversarial Attack on Vision-Language Pre-training Models. In *Proceedings of the 30th ACM International Conference on Multimedia (Lisboa, Portugal) (MM '22)*. Association for Computing Machinery, New York, NY, USA, 5005–5013. <https://doi.org/10.1145/3503161.3547801>
- [33] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. 2024. Universal Adversarial Perturbations for Vision-Language Pre-trained Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (Washington DC, USA) (SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 862–871. <https://doi.org/10.1145/3626772.3657781>
- [34] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems* 36 (2024).
- [35] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).