

Sample-agnostic Adversarial Perturbation for Vision-Language Pre-training Models

Haonan Zheng

zhenghaonan@mail.nwpu.edu.cn

Northwestern Polytechnical University
School of Electronics and Information
Xi'an, Shaanxi, China

Xinyang Deng*

Northwestern Polytechnical University
School of Electronics and Information
Xi'an, Shaanxi, China

Wen Jiang

Northwestern Polytechnical University
School of Electronics and Information
Xi'an, Shaanxi, China

Wenrui Li

liwr618@163.com

Harbin Institute of Technology
Department of Computer Science and Technology
Harbin, China

Abstract

Recent studies on AI security have highlighted the vulnerability of Vision-Language Pre-training (VLP) models to subtle yet intentionally designed perturbations in images and texts. Investigating multimodal systems' robustness via adversarial attacks is crucial in this field. Most multimodal attacks are sample-specific, generating a unique perturbation for each sample to construct adversarial samples. To the best of our knowledge, it is the first work through multimodal decision boundaries to explore the creation of a universal, sample-agnostic perturbation that applies to any image. Initially, we explore strategies to move sample points beyond the decision boundaries of linear classifiers, refining the algorithm to ensure successful attacks under the top k accuracy metric. Based on this foundation, in visual-language tasks, we treat visual and textual modalities as reciprocal sample points and decision hyperplanes, guiding image embeddings to traverse text-constructed decision boundaries, and vice versa. This iterative process consistently refines a universal perturbation, ultimately identifying a singular direction within the input space which is exploitable to impair the retrieval performance of VLP models. The proposed algorithms support the creation of global perturbations or adversarial patches. Comprehensive experiments validate the effectiveness of our method, showcasing its data, task, and model transferability across various VLP models and datasets. Code: <https://github.com/Libertazz/MUAP>

CCS Concepts

- Security and privacy; • Information systems → Multimedia and multimodal retrieval;

Keywords

Cross-modal Retrieval; Decision Boundary; Universal Perturbation

ACM Reference Format:

Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. 2024. Sample-agnostic Adversarial Perturbation for Vision-Language Pre-training Models. 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, 13 pages. <https://doi.org/xxxxx>

*Corresponding author

1 Introduction

Multimodal contrastive learning [31, 41] aims to leverage raw text as a much broader source of supervision to encourage models to learn more flexible visual representations which align with textual representations. Building on this, by incorporating masked data modelling [7, 1, 28, 37], image-text matching [12], and other advanced pre-training tasks [39], a multimodal learning paradigm based on pre-training plus fine-tuning has gradually formed. This paradigm shift has facilitated the inception of various VLP models [31, 14, 15, 2, 16] featuring meticulously crafted architectures and adaptable pre-trained weights which can be fine-tuned [45, 46] to perform a wide range of downstream tasks [27, 40, 20, 38]. Behind these promising outcomes lie potential security risks [42, 9, 22, 4]. Research indicates that adding imperceptible noise to images or making minor adjustments at the word or letter level in texts can lead to severe performance degradation in VLP models [21, 47, 10, 44]. These maliciously crafted inputs are adversarial examples [9]. Existing work has explored the generation of adversarial examples in VLP models from multiple perspectives. For example, data poisoning can disrupt the training phase, resulting in models that are susceptible to backdoor attacks [10]. During testing, attacks are conducted either in the white-box scenario where the model weights are transparent [42], or in the black-box context where the model is inaccessible causing transferability to be the primary consideration [21]. Given the public availability of pre-trained weights, we aim to design a novel white-box attack method for pre-trained VLP models like CLIP [31], exploring potential security vulnerabilities in the multimodal domain. We further consider the attack patterns from feasibility, immediacy and practicality. (1) **Feasibility:** The differentiability (continuous pixel values) and semantic robustness (minor perturbations don't impact human perception) make image inputs apt carriers for malicious noise. In contrast, attacking text often involves operations like word replacement, inevitably leading to semantic errors. In application scenarios, text often mediates human-computer interactions, and attackers have no access to texts directly provided by users. Conversely, images often automatically

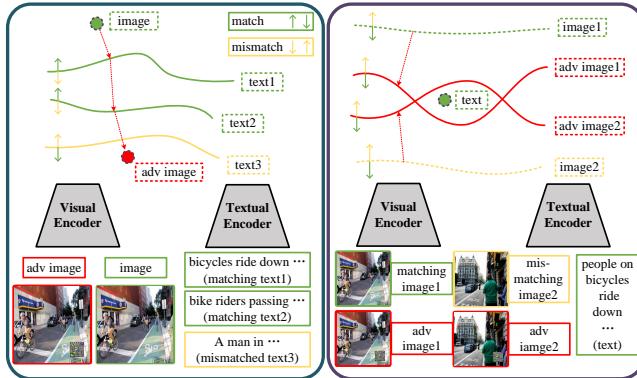


Figure 1: The left part represents attacking text retrieval, involving guiding benign embedding across decision boundaries to generate adversarial images; the right part represents attacking image retrieval, involving distorting benign decision hyperplanes to construct malicious boundaries, corrupting the decision outcomes for benign text.

extracted by intelligent models are accessible for attackers to introduce perturbations. Furthermore, we emphasize the universality of perturbations. Although optimizing them for non-differentiable textual tokens is challenging, it is relatively feasible for image inputs. Therefore, we focus on attacking the visual modality, without neglecting multimodal interactions. (2) **immediacy**: The expeditious process of image acquisition by visual sensors contrasts with sample-specific attacks requiring the generation of adversarial perturbations for images on the spot. Diverging from sample-specific strategies, our approach incorporates decision boundary theory to craft a sample-agnostic perturbation. Irrespective of various images captured at any moment, directly applying a prepared perturbation can create effective adversarial images. Therefore, the primary consideration is universality. (3) **Practicality**: Image perturbation can be categorized into two main types: global noise and adversarial patch. Global noise is impractical for real-world applications due to the infeasibility of directly modifying pixel values in physical scenarios. In contrast, adversarial patches cover only part of an image and can be easily deployed in the physical world. Therefore, our primary focus is patch form, with additional discussion on global perturbation for completeness.

As mentioned above, we focus on designing the universal adversarial patch for image modality in the white-box scenario. To achieve this, we propose an innovative algorithm for generating universal patches based on decision boundary theory, introducing a novel approach in the multimodal domain. Starting with the basic principles of the binary linear classifier, we outline how to calculate the distance and direction from a point to the decision hyperplane. We extend this principle to the linear multi-class classifier, detailing a strategy that forces sample points across multiple decision boundaries simultaneously, ensuring successful attacks under the top k accuracy metric. Subsequently, we apply these theoretical results to the multimodal scenario, in which images and texts can serve interchangeably as sample points and decision boundaries.

This gives rise to two attack methodologies as shown in Fig 1. The key contributions of our research are as follows:

- We pioneer the application of decision boundary theory to the multimodal scenario, providing a theoretical foundation for exploring the robustness of image and visual embeddings within the high-dimensional space.
- We design a novel algorithm for generating universal adversarial perturbation, a singular malicious direction in the image input space that can disrupt retrieval performance.
- Through comprehensive experimentation across various datasets and VLP models, our universal perturbation exhibits remarkable adaptability across samples, highlighting its efficacy in transfer from pre-trained encoders to downstream applications.

2 Related Work

Vision-Language Pre-training Models. Initially, VLP models are developed leveraging pre-trained object detectors, employing region features to acquire vision-language representations [3, 18, 43, 36, 17]. The Vision Transformer (ViT) [8] introduces a powerful transformer-based encoder for visual data, enabling unified feature extraction from multimodal inputs, coupled with multimodal contrastive learning, establishing a straightforward and efficient paradigm for multimodal pre-training. A representative model is CLIP [31], which includes a text and visual encoder. These unimodal encoders extract textual and visual representations and align them within a normalized contrastive space. Subsequently, additional pre-training tasks are introduced into the multimodal domain, such as Masked Language Modeling (MLM) proposed by BERT [7], and Masked Image Modeling (MIM), proposed by BEiT [1] and further developed by BEiT2 [28]. BEiT3 unifies these results into Masked Data Modeling (MDM), which integrates MIM, MLM, and Masked Language-Vision Modeling (MLVM), marking a new paradigm for multimodal pre-training. Based on these, incorporating Multiway Transformer [2] which splits the feed-forward layer into three parallel paths for unimodal feature extraction and feature fusion respectively, BEiT3 achieves better performance across a range of downstream applications.

Decision Boundary. DeepFool [24] is an adversarial algorithm tailored for neural network image classifiers, leveraging decision boundaries to induce misclassification through minimal perturbations. Although it offers an effective strategy based on decision boundary analysis, its application is limited to image classification tasks. The algorithm aims to achieve minimal perturbation by iteratively exploring the direction of proximal disturbance. This process requires extensive gradient backpropagation, significantly increasing computational time, especially as the number of categories expands, making the approach less feasible. We adopt the decision boundary concept, extend it to the multimodal scenario, and introduce pragmatic solutions, e.g., utilizing cosine similarity to ascertain the nearest decision boundary, effectively reducing redundant gradient backpropagation to enhance efficiency.

Universal Adversarial Perturbation. Universal adversarial perturbations in image classification are extensively discussed and can be divided into sample-dependent and sample-independent. The former kind includes methods such as UAP[23], SGUAP[34],

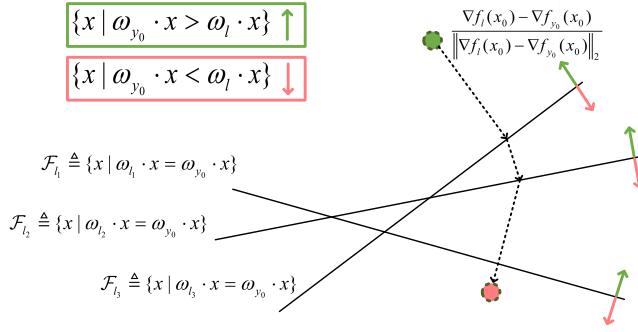


Figure 2: Illustration of the disturbance trajectory that simultaneously crosses the top k decision boundaries.

NAG[34], GAP[30], and FFF[25], which access a set of surrogate images to identify a universal adversarial perturbation, utilizing generative adversarial networks or various optimization strategies. It is noteworthy that these methods focus exclusively on global perturbation. Similarly, our approach requires a batch of proxy data, yet we prioritize the adversarial patch which is more practical in the physical world. The other kind involves methods that do not interact with training images, such as Data-free UAP[26]. This approach generates class impressions for each category based on which it constructs universal perturbations. Since class impressions are tailored specifically for the classification task, they are not applicable in the multimodal context. The work most relevant to ours is AdvCLIP[47], which has developed a universal patch attack for multimodal contrastive learning based on topological structures. It provides valuable insights into addressing the modality gap, boosting attack transferability between pre-trained encoders and downstream tasks, as well as some defence plans. However, it relies on Generative Adversarial Networks (GANs), incurring additional overhead, and the topological structure loss fails to generate stable attack effects. Furthermore, it lacks validation on large-scale datasets and models.

3 Methodology

3.1 Theoretical Basis

Linear Binary Classifier. We consider a sample pair as (x, y) , where $x \in \mathbb{R}^n$ is the input to be classified, and $y \in \{-1, +1\}$ is the true label. A binary classifier is given by $\hat{y} = \text{sign}(f(x))$, where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a linear function $f(x) = \omega \cdot x + b$, and $\hat{y} \in \{-1, +1\}$ is the predicted category. This binary classifier divides the n-dimensional space into two parts: $\{x | f(x) > 0\}$ and $\{x | f(x) < 0\}$, with the decision boundary being the hyperplane $\mathcal{F} \triangleq \{x | \omega \cdot x + b = 0\}$. Considering a sample x_0 where $f(x_0) > 0$, i.e. $\hat{y}_0 = +1$, we try to find a perturbation r which will be added to the clean sample x_0 to obtain the adversarial sample $x_0^{adv} = x_0 + r$. According to the point-to-plane distance formulation, the distance from x_0 to the decision boundary \mathcal{F} is:

$$\frac{f(x_0)}{\|\omega\|_2} = \frac{f(x_0)}{\|\nabla f(x_0)\|_2} \quad (1)$$

where $\nabla f(x_0)$ represents the gradient of the function f with respect to x_0 . The direction of perturbation r is normal to the decision

boundary \mathcal{F} , which is:

$$-\frac{\omega}{\|\omega\|_2} = -\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|_2} \quad (2)$$

Based on this, the perturbation direction r is given by distance times unit normal vector:

$$r = -\frac{f(x_0) \cdot \nabla f(x_0)}{\|\nabla f(x_0)\|_2^2} \quad (3)$$

where $\|\cdot\|_2$ denotes the Euclidean distance.

Linear multi-class classifier. We continue to consider a multi-class classifier (such as a linear layer), where (x, y) is a sample pair, with $y \in \{1, 2, \dots, C\}$. The weight matrix of the linear layer is $W = \{\omega_i\}_{i=1}^C$, where $\omega_i \in \mathbb{R}^n$ is the feature vector corresponding to category i . We define $f_i(x) = \omega_i \cdot x$, where \cdot represents vector inner product. Then, the prediction result of the multi-class classifier is $\hat{y} = \arg \max_i(f_i(x))$. We consider a correctly classified sample pair (x_0, y_0) , which satisfies:

$$\forall i \in \{1, 2, \dots, C\} \wedge i \neq y_0, f_{y_0}(x_0) > f_i(x_0) \quad (4)$$

To attack the multi-class classifier, we similarly need to find an adversarial perturbation r to push the sample x_0 across the decision boundary. In this scenario, there exist $C - 1$ decision boundaries $\mathcal{F}_i \triangleq \{x | f_{y_0}(x) - f_i(x) = 0\} = \{x | (\omega_{y_0} - \omega_i) \cdot x = 0\} (i \neq y_0)$. x_0 crossing any decision boundary will lead to misclassification. Therefore, we first select the nearest decision boundary \mathcal{F}_{l_1} :

$$l = \underset{i \neq y_0}{\operatorname{argmin}} \frac{f_{y_0}(x_0) - f_i(x_0)}{\|\omega_{y_0} - \omega_i\|_2} = \underset{i \neq y_0}{\operatorname{argmin}} \frac{f_{y_0}(x_0) - f_i(x_0)}{\|\nabla f_{y_0}(x_0) - \nabla f_i(x_0)\|_2} \quad (5)$$

Based on this, we can determine that the nearest perturbation direction is normal to the decision boundary \mathcal{F}_l , that is:

$$-\frac{\omega_{y_0} - \omega_l}{\|\omega_{y_0} - \omega_l\|_2} = \frac{\nabla f_l(x_0) - \nabla f_{y_0}(x_0)}{\|\nabla f_{y_0}(x_0) - \nabla f_l(x_0)\|_2} \quad (6)$$

Finally, we can obtain the perturbation r :

$$r = \frac{(\nabla f_l(x_0) - \nabla f_{y_0}(x_0)) \cdot (f_{y_0}(x_0) - f_l(x_0))}{\|\nabla f_{y_0}(x_0) - \nabla f_l(x_0)\|_2^2} \quad (7)$$

Multi-class classifier under top k metric. The described attack strategy for the multi-class classifier faces a problem: the adversarial sample \hat{x}_0 crosses only the nearest decision boundary. If the evaluation is based on the top- k metric (i.e., the classification result is considered correct if the true category is among the top k candidate categories with the highest scores), this attack scheme cannot guarantee a successful attack. Therefore, the perturbation r must guide x_0 across the k nearest decision boundaries simultaneously, ensuring the top k highest-scoring categories exclude the ground truth label. Based on Equation 5, we can determine the k nearest decision boundaries $\{\mathcal{F}_{l_j}\}_{j=1}^k$. We obtain the perturbation r in an iterative process, as shown in Alg 1. We set $\eta = 0.02$ representing the extent to cross the decision boundary in each iteration.

3.2 Multimodal Universal Perturbation

Previously, we analyzed the method of obtaining the adversarial perturbation r for a sample point x_0 in the linear scenario. In this section, based on the theoretical foundation outlined above, we will use the image-text retrieval task to design a novel algorithm

Algorithm 1: Crossing k decision boundaries

Input : Clean sample pair (x, y) , k nearest decision boundaries $\{\mathcal{F}_{l_j}\}_{j=1}^k$, $L = \{l_j\}_{j=1}^k$

Output: Perturbation r .

- 1 Initialize $x_0 \leftarrow x$, $i \leftarrow 0$, $r_0 \leftarrow 0$
- 2 **while** $\exists l \in L$, $f_y(x + (1 + \eta) \cdot r) > f_l(x + (1 + \eta) \cdot r)$ **do**
- 3 $\hat{x} \leftarrow x + r$
- 4 $l_{min} \leftarrow \underset{l \in L}{\operatorname{argmin}} \frac{f_l(\hat{x}) - f_y(\hat{x})}{\|\nabla f_l(\hat{x}) - \nabla f_y(\hat{x})\|_2}$
- 5 $r \leftarrow r + \frac{(\nabla_r f_{l_{min}}(\hat{x}) - \nabla_r f_y(\hat{x})) \cdot (f_y(\hat{x}) - f_{l_{min}}(\hat{x}))}{\|\nabla_r f_y(\hat{x}) - \nabla_r f_{l_{min}}(\hat{x})\|_2^2}$
- 6 **return** $(1 + \eta) \cdot r$

for the multimodal scenario, generating a universal adversarial perturbation for VLP models.

Problem Formulation. Let $V = \{v_i\}_{i=1}^N$ and $T = \{t_i\}_{i=1}^M$ represent an image set and a text set, respectively. E_v and E_t denote the visual encoder and text encoder, where $E_v(v), E_t(t) \in \mathbb{R}^d$ represent the feature vectors normalized by ℓ_2 norm. We primarily consider two image-text downstream tasks: image-text retrieval and image classification. Taking the image-to-text retrieval task as an example, we define the indicator function $I(E_v(v), E_t(T), k)$, where $E_t(T) \in \mathbb{R}^{M \times d}$. First, we compute the cosine similarity between the visual embedding and the M text embeddings. If among the top k texts with the highest cosine similarity to $E_v(v)$, there exists at least one text that matches v , the value of the indicator function is 1; otherwise, it is 0. Similarly, for text-to-image retrieval, the same indicator function is still applicable, denoted as $I(E_t(t), E_v(V), k)$. We aim to identify a universal perturbation δ for the image modality that is sample-agnostic, meaning it applies to all images instead of being specific to one. The attack objective is as follows:

$$\forall v \in V, I(E_v(v + \delta), E(T), k) = 0 \quad (8)$$

$$\forall t \in T, I(E_t(t), E(V + \delta), k) = 0 \quad (9)$$

where $V + \delta = \{v_i + \delta\}_{i=1}^M$ represents adding the universal perturbation δ to all images in the set V .

For the image classification problem, we follow CLIP [31] to construct texts based on category names, forming the text set $T = \{t_i\}_{i=1}^C$, where C is the number of categories, with texts like “a photo of a xx.” For an image, we calculate the cosine similarity between the image embedding and the text embeddings of all categories and classify the image into the most similar category. For the classification task, perturbation δ should satisfy:

$$\forall v \in V, \underset{i}{\operatorname{argmax}}(E_v(v + \delta) \cdot E_t(t_i)) \neq y \quad (10)$$

where $y \in \{1, 2, \dots, C\}$ is the true label of v .

Attacking image-to-text retrieval. We begin with a single image $v \in \mathbb{R}^{c \times h \times w}$, where c , h , and w respectively stand for the number of channels, height, and width. We denote $Y = \{y_j\}_{j=1}^n$, where n represents that there are n texts in T that match v , i.e. $\{t_y\}_{y \in Y}$ match with v . Let $Y' = \{y'_j\}_{j=1}^k$, where Y' contains the indices of k texts that mismatch v , and $Y \cap Y' = \emptyset$. We consider the image-to-text retrieval task as a classification task, denoting $f_i(v) = E_t(t_i) \cdot E_v(v)$, where $E_t(t_i) \in \mathbb{R}^d$ plays the same role as ω_i

Algorithm 2: Attacking image-to-text retrieval

Input : Image set V , text set T , mask matrix m .

Output: Perturbation δ .

- 1 Initialize $\delta \leftarrow 0$
- 2 **for** v in V **do**
- 3 getting $Y = \{y_j\}_{j=1}^n$ and $Y' = \{y'_j\}_{j=1}^k$
- 4 $v \leftarrow v \odot (1 - m) + \delta \odot m$
- 5 $r \leftarrow 0$
- 6 **while** $I(E_v(v + (1 + \eta) \cdot r \odot m), E_t(T), k) == 1$ **do**
- 7 $\hat{v} \leftarrow v + r \odot m$
- 8 $y'_{min} \leftarrow \underset{y' \in Y'}{\operatorname{argmin}} f_{y'}(\hat{v})$
- 9 $y_{max} \leftarrow \underset{y \in Y}{\operatorname{argmax}} f_y(\hat{v})$
- 10 $r \leftarrow$
- 11
$$r + \frac{(\nabla_r f_{y'_{min}}(\hat{v}) - \nabla_r f_{y_{max}}(\hat{v})) \cdot (f_{y_{max}}(\hat{v}) - f_{y'_{min}}(\hat{v}))}{\|\nabla_r f_{y'_{min}}(\hat{v}) - \nabla_r f_{y_{max}}(\hat{v})\|_2^2}$$
- 12 $\delta \leftarrow \operatorname{clamp}_{(0,1)}(\delta + (1 + \eta) \cdot r)$
- 13 **return** δ

in the multi-class classifier. We will traverse the entire image set V , but only optimize one perturbation $\delta \in \mathbb{R}^{c \times h \times w}$. The pseudocode for generating δ is shown in Alg 2. We must clarify four points: (1) Y can be directly obtained through the annotation information of the dataset. (2) Y' serves as the target category for the attack. We can select the k nearest non-matching texts based on the decision boundary according to Equ 5. However, due to the large size of the text set, it will introduce a large amount of gradient backpropagation, severely affecting optimization efficiency. As an alternative, we use the cosine similarity score as the criterion and select the k non-matching texts with the highest similarity to $E_v(v)$. (3) We hypothesize that applying universal adversarial perturbations in patch form to images will be more effective than global perturbations, a theory we will test in subsequent experiments. Therefore, we introduce $m \in \mathbb{R}^{c \times h \times w}$, a binary matrix that specifies the patch's location. \odot denotes the element-wise product. (4) $\operatorname{clamp}_{(0,1)}(\delta)$ represents clipping each element value in δ to be between 0 and 1, obtaining valid pixel values.

Attacking text-to-image retrieval. In this scenario, the most intuitive approach is to construct decision boundaries with $E(V) \in \mathbb{R}^{N \times d}$ and generate adversarial text t_i^{adv} for text t_i to cross the decision boundary, similar to attacking image-to-text retrieval. However, due to the non-differentiability of textual data, generating the universal adversarial perturbation for texts is impractical. Therefore, we continue to create the universal adversarial patch for the image modality, establishing malicious boundaries with $E(V + \delta)$. This implies that the text embedding remains unchanged, and when confronting the benign decision hyperplanes constituted by $E(V)$, the text can accurately retrieve the matching image. While facing the malicious decision boundaries formed by $E(V + \delta)$, the text embedding is placed on an incorrect decision plane, leading to significant confusion. The pseudocode is shown in Alg 3. We need to clarify two points: (1) A text matches only one image, with y

Algorithm 3: Attacking text-to-image retrieval

Input :Image set V , text set T , mask matrix m .
Output:Perturbation δ .

```

1 Initialize  $\delta \leftarrow 0$ 
2 for  $t$  in  $T$  do
3   getting  $y$  and  $Y' = \{y'_j\}_{j=1}^k$ 
4    $\tilde{V} = \{\tilde{v}_i\}_{i \in Y \cup Y'} \leftarrow \{v_i \odot (1 - m) + \delta \odot m\}_{i \in Y \cup Y'}$ 
5    $r \leftarrow 0$ 
6   while  $I(E_t(t), E_v(\tilde{V} + (1 + \eta) \cdot r \odot m), k) == 1$  do
7      $\hat{V} = \{\hat{v}_i\}_{i \in Y \cup Y'} \leftarrow \tilde{V} + r \odot m$ 
8      $y'_{min} \leftarrow \underset{y' \in Y'}{\operatorname{argmin}} f_{y'}(t) // f_i(t) = E_v(\hat{v}_i) \cdot E_t(t)$ 
9      $r \leftarrow r + \frac{(\nabla_r f_{y'_{min}}(t) - \nabla_r f_y(t)) \cdot (f_y(t) - f_{y'_{min}}(t))}{\|\nabla_r f_{y'_{min}}(t) - \nabla_r f_y(t)\|_2^2}$ 
10     $\delta \leftarrow \operatorname{clamp}_{(0,1)}(\delta + (1 + \eta) \cdot r)$ 
11 return  $\delta$ 
```

Table 1: The first column represents three different visual encoder configurations of the CLIP model. The pre-trained weights of CLIP are utilized directly, rather than weights fine-tuned on specific datasets.

Visual Encoder	Dataset	Test Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	Flickr30k	82.20	96.70	98.90	62.12	85.72	91.94
	COCO	52.40	76.78	84.66	33.07	58.35	69.03
ViT-B/32	Flickr30k	78.60	94.90	98.30	58.80	83.56	90.00
	COCO	50.18	75.02	83.50	30.48	55.99	66.87
ViT-L/14	Flickr30k	85.30	97.40	99.20	64.84	87.22	92.04
	COCO	56.34	79.40	86.58	36.51	61.08	71.17

representing the index of that image, obtained from the annotation information. (2) Y' includes k non-matching images with the highest cosine similarity to $E_v(t)$.

In summary, the inner loop concentrates on a single sample to determine the perturbation direction r . In the outer loop, we continuously accumulate r onto δ , ultimately obtaining the universal adversarial perturbation applicable to all images. Our method can be adapted to create universal adversarial global perturbations (Appendix C). We don't specifically design the attack strategy for zero-shot classification task for two reasons: (1) Universal adversarial perturbation for the classification problem has been extensively studied in the unimodal field [23]. (2) Our attack strategy demonstrates excellent transferability between datasets, meaning that δ constructed based on the image-text retrieval task can be directly applied to classification datasets with good attack effectiveness.

4 Experiments

In this section, we begin by presenting the performance of the CLIP model on the Flickr30K[29] and MS COCO[19] datasets' test

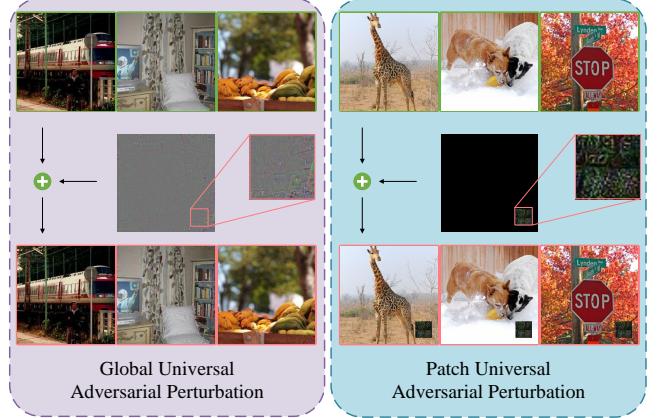


Figure 3: The top row shows the clean image, the middle row shows the universal adversarial perturbation (both global and patch), and the bottom row shows the adversarial image with the perturbation applied.

Table 2: The effects of universal patch attack with three different configurations.

Visual Encoder	Config	Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	TIRA	0.10	0.60	1.00	0.06	0.36	0.54
	TRA	1.60	4.00	7.40	25.92	51.24	62.36
	IRA	0.00	0.00	0.20	0.08	0.22	0.60
ViT-B/32	TIRA	0.50	1.20	2.50	3.72	7.80	10.90
	TRA	0.70	2.20	4.80	8.94	32.30	45.58
	IRA	1.00	2.80	5.50	3.00	6.44	9.18
ViT-L/14	TIRA	0.10	0.30	0.90	0.52	1.40	2.50
	TRA	0.40	3.70	6.90	1.56	6.12	10.76
	IRA	0.00	0.00	0.30	0.10	0.22	0.36

split for image-to-text retrieval task under the normal condition, in Tab 1. This represents the baseline retrieval performance of the CLIP model. In an adversarial environment, the retrieval performance of CLIP will significantly deteriorate. The greater the decline, the more effective the attack.

4.1 Exploring the Optimal Configuration

In this section, we explore some configurations when generating universal adversarial patches based on the CLIP_{vit/B16}, CLIP_{vit/B32} and CLIP_{vit/L14} models. We use the first 1000 images from the Flickr30k training set and the corresponding 5000 texts to construct $V = \{v_i\}_{i=1}^{1000}$ and $T = \{t_i\}_{i=1}^{5000}$, meaning that only 1000 images are visible during the optimization process of generating the universal adversarial patch. we use the mask matrix m to ensure that the patch area is 3% of the total image area. Since the input size of the CLIP model is 224×224 , the patch size is 38×38 , and it is fixed in the bottom right corner of the image, as shown in Fig 3. We set k to 10, as we will evaluate the R@10 metric. During the evaluation, We

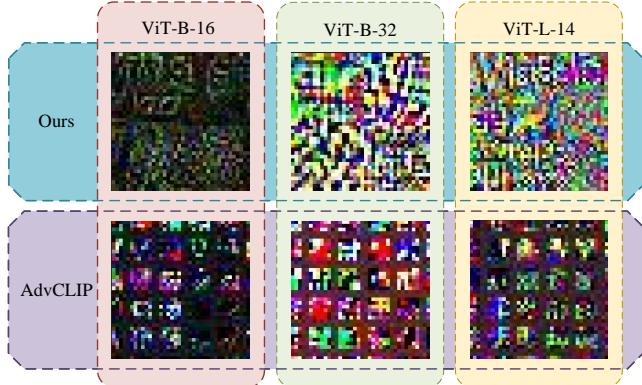


Figure 4: Visualization of universal adversarial patches generated by two methods on three models.

apply the universal patch to all images in the Flickr30k test split and conduct the image-text retrieval task using the CLIP model.

In Alg 2 and Alg 3, we have provided detailed descriptions of how to attack the text retrieval and image retrieval tasks. These two attack strategies can be used independently (denoted as **IRA** and **TRA** for short), or in conjunction (“Text retrieval and Image Retrieval Attack” is denoted as **TIRA**). See Appendix A for details. We conduct a comparison of three configurations based on patch form. The results are shown in Tab 2, and we find that using TRA alone results in poor attack performance on the image retrieval (i.e., text-to-image retrieval) sub-task. However, employing IRA alone can achieve good attack performance on both sub-tasks. During model training, one image corresponds to multiple texts, while one text corresponds to only a single image. Besides the cross-modal contrastive representation learning between images and texts, the textual modality implicitly includes within-modality contrastive representation clustering, which does not exist in the visual modality. When using TRA to drive image representations away from similar texts, it does not disrupt the topological structure among the image representations. That is, from the perspective of text representations, all matching image representations are moving away, but the similarity ranking remains largely unchanged. Therefore, the performance decline in the image retrieval sub-task is minimal. Based on this, we speculate that introducing contrastive learning between images during the visual-language pre-training process might enhance model robustness, a hypothesis that awaits exploration in future work.

4.2 Comparing with SOTA

In this section, we compare our method with AdvCLIP[47], which is currently the best method for generating universal adversarial patches against VLP models. We denote our proposed universal patch attack (using TIRA) as $\text{UAP}_{\text{TIRA}}^{\text{patch}}$. See Appendix B for more experiments about $\text{UAP}_{\text{IRA}}^{\text{patch}}$. For fairness, both approaches utilize the same 1000 images from Flickr30K as the surrogate images and maintain identical patch size and position mentioned above. Based on three CLIP models, the six universal adversarial patches generated by two methods are shown in Fig 4.

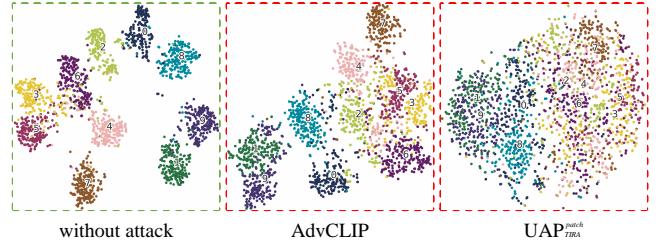


Figure 5: Based on the CIFAR-10 dataset and the $\text{CLIP}_{\text{vit}/\text{B16}}$ model, t-SNE feature dimensionality reduction visualization reveals that upon the addition of the universal adversarial patch generated by $\text{UAP}_{\text{TIRA}}^{\text{patch}}$, the features extracted by the visual encoder become severely disordered, leading directly to a collapse in model performance.

Image-text Retrieval. The effects of two attack methods are presented in Tab 3. We observe that: (1) The universal adversarial patches generated by the Advclip perform poorly in the image retrieval sub-task. (2) Our patches outperform Advclip across all metrics. (3) Our patches, generated based on Flickr30K images demonstrate strong attack effectiveness even when applied to the MS COCO dataset, indicating that our method possesses good **cross-dataset transferability**.

Zero-shot Classification. We perform this task on three commonly used classification datasets: ImageNet1k[6], CIFAR-100[13], and CIFAR-10[13]. It’s important to note that we continue to use the six universal adversarial patches from Fig 4, instead of generating unique patches for each dataset. We resize the images to 224×224 to meet the requirement of the CLIP models. The effects of the attacks are presented in Tab 4. It can be observed that: (1) Our patches consistently outperform Advclip in terms of attack effectiveness. (2) The dataset with fewer categories is less susceptible to attack because the fewer the categories, the sparser the decision boundaries, making samples farther from these boundaries and crossing them more difficult. (3) Our patches exhibit good **cross-task transferability**. We employ t-SNE[35] to reduce the dimensionality and visualize features extracted by the ViT-B/16 visual encoder on the CIFAR-10 dataset, as shown in Fig 5. Our method evidently results in a more chaotic feature distribution.

Evaluated on Fine-tuned Models. We select three datasets, namely NUS-WIDE[5], Pascal-Sentence[33], and Wikipedia[32], that support both image-text retrieval and image classification. The data format is a triplet (v, t, y) , consisting of matching image v and text t with their category y . Based on the three datasets, two downstream tasks, and three CLIP pre-trained models, we obtained a total of 18 fine-tuned models. Considering the powerful zero-shot capability of the CLIP model, we follow the setup of AdvCLIP[47], freezing the weights of CLIP encoders and only fine-tuning the subsequent nonlinear layers. For the image-text retrieval, we add a nonlinear projection head after both the text and visual encoders respectively to obtain feature vectors, and then perform classification supervised by category annotation information through the same linear layer. During the testing phase, we discard the linear classification head and calculate the cosine similarity between embeddings output by the nonlinear projection heads to perform the

Table 3: Comparing between AdvCLIP and UAP^{patch}_{TIRA}, based on the Flickr30k and MS COCO datasets. Recall rates are reported.

Visual Encoder	Method	Flickr30k(1K test set)						MS COCO(5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	Advclip	2.30	7.00	11.20	19.98	48.08	64.38	1.58	4.44	6.54	8.10	23.55	34.49
	UAP ^{patch} _{TIRA}	0.10	0.60	1.00	0.06	0.36	0.54	0.02	0.04	0.10	0.18	0.51	0.75
ViT-B/32	Advclip	10.80	23.40	32.80	25.68	54.94	67.60	7.18	17.77	24.34	11.66	29.50	40.84
	UAP ^{patch} _{TIRA}	0.50	1.20	2.50	3.72	7.80	10.90	0.38	1.30	2.00	2.00	5.35	8.20
ViT-L/14	Advclip	4.80	14.40	19.20	18.28	44.18	55.82	0.84	3.18	5.06	6.84	19.32	28.87
	UAP ^{patch} _{TIRA}	0.10	0.30	0.90	0.52	1.40	2.50	0.12	0.20	0.36	0.32	0.77	1.14

Table 4: Comparing between AdvCLIP and UAP^{patch}_{TIRA}, based on the ImageNet, CIFAR-100 and CIFAR-10 datasets. The accuracy is reported. Lower is better.

Visual Encoder	Method	ImageNet		CIFAR100		CIFAR10	
		Top1	Top5	Top1	Top5	Top1	Top5
ViT-B/16	w/o atk	62.42	86.74	66.56	88.49	90.10	99.07
	Advclip	5.98	14.80	8.91	45.04	81.90	97.01
	UAP ^{patch} _{TIRA}	0.12	0.48	1.62	5.82	10.67	53.04
ViT-B/32	w/o atk	57.50	83.57	62.27	86.98	88.34	99.24
	Advclip	7.98	27.40	31.92	51.84	64.45	96.51
	UAP ^{patch} _{TIRA}	1.68	5.36	7.22	17.76	15.10	68.69
ViT-L/14	w/o atk	69.74	90.15	75.72	93.06	95.19	99.52
	Advclip	1.54	7.61	7.42	32.30	73.87	97.95
	UAP ^{patch} _{TIRA}	0.58	1.61	6.57	15.44	20.76	70.55

image-text retrieval task. We only measure the R@1 metric, with text retrieval R@1 and image retrieval R@1 abbreviated as TR@1 and IR@1, respectively. For the image classification, we measure the Top1 and Top5 accuracy. Note, the adversarial patches remain the six patches from Fig 4, aimed at observing the attack effectiveness of perturbations generated based on pre-trained models when transferred to fine-tuned models. From Tab 5 we observed that our method consistently maintains better and more stable attack performance, demonstrating its excellent **cross-model transferability**. Due to the shared parameters between the encoder of the pre-trained model and the encoder of the fine-tuned model, it does not constitute a transfer attack in the absolute sense.

We recognize that our approach requiring precise perturbation distance and direction determination, involves extracting an individual adversarial direction for each image and obtaining a universal perturbation ultimately. As a result, our method consumes more optimization time than AdvCLIP. But these additional time overheads occur before the perturbation is deployed, not affecting the real-time nature of the attack strategy, and are therefore entirely acceptable. Due to the absence of batch processing, our method's effectiveness

is not influenced by batch size, in contrast to AdvCLIP, which exhibits significant performance fluctuation under different batch-size values according to its original paper[47]. Furthermore, our method requires few GPU computational resources. For instance, utilizing the CLIP model to iterate over 1,000 Flickr30k surrogate images 50 times on a single NVIDIA RTX4090 GPU necessitates 2430MB of memory and approximately 18 hours of optimization time.

4.3 Extend to Larger VLP models

Considering the CLIP pre-trained models do not exhibit state-of-the-art performance in the non-adversarial environment, we chose the recently open-sourced VLP model BEiT3[37], which has a larger parameter size and more advanced image-text interactive capabilities, to continue validating the effectiveness of our attack method. D_1 and D_2 represent the MS COCO and Flickr30k datasets. M_1 and M_2 denote the BEiT3 models fine-tuned on D_1 and D_2 , respectively. $Patch_1$ and $Patch_2$ are universal adversarial patches generated based on M_1 using 1000 training images from MS COCO and 1000 training images from Flickr30k, respectively. M_2 is reserved for testing the cross-model transferability, with experimental results shown in Tab 6. The experimental results validate that our universal adversarial patches possess stable and potent attack effectiveness, as well as good cross-dataset and cross-model transferability. See Appendix D for more experiments on BEiT3.

4.4 Global Perturbation

Our method can also generate universal adversarial perturbations in global form. See Appendix C for technical details. We use 1000 Flickr30k images as proxy images, utilizing only the IRA (Image Retrieval Attack) under constraints based on the ℓ_2 norm ($\epsilon_{\ell_2} = 2000$) and ℓ_∞ norm ($\epsilon_{\ell_\infty} = 10$), following [23]. Fig 6 shows the performance degradation of the CLIP models when universal global perturbations are applied to the entire Flickr30k test dataset. Fig 7 shows the results of transferring global perturbations to the COCO dataset to demonstrate their cross-dataset transferability. We observe that under ℓ_2 norm constraint, universal global perturbations generally achieve better attack effectiveness. Visualization of global perturbations can be found in Appendix C.

Table 5: Comparing between AdvCLIP and UAP^{patch}_{TIRA}, based on the NUS-WIDE, Pascal-Sentence and Wikipedia datasets.

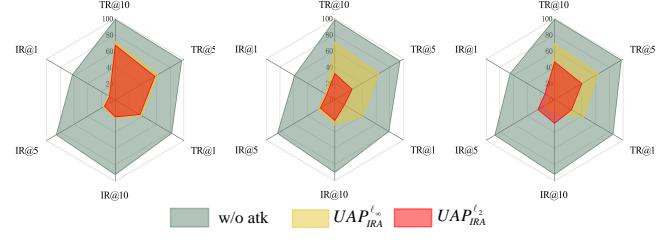
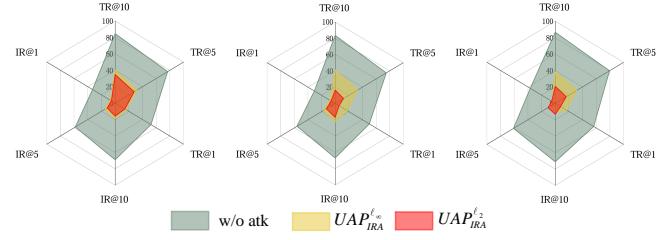
Encoder	Metric	NUS-WIDE			Pascal-Sentence			Wikipedia		
		w/o atk	AdvCLIP	UAP ^{patch} _{TIRA}	w/o atk	AdvCLIP	UAP ^{patch} _{TIRA}	w/o atk	AdvCLIP	UAP ^{patch} _{TIRA}
ViT-B/16	TR@1	66.90	36.75	11.50	67.00	7.00	5.00	61.69	14.72	6.93
	IR@1	61.80	51.25	16.80	68.50	10.00	7.50	66.67	42.64	23.16
	AVG	64.35	44.00	14.15	67.75	8.50	6.25	64.18	28.68	15.05
	Top1	77.85	49.65	11.25	77.00	5.50	5.50	64.01	27.95	7.09
	Top5	98.75	72.95	47.55	96.00	39.50	29.00	94.35	73.15	52.75
ViT-B/32	TR@1	69.25	34.15	23.25	72.00	35.00	10.50	58.23	31.60	17.97
	IR@1	60.25	47.60	39.10	70.50	47.00	17.00	62.77	48.70	18.61
	AVG	64.75	40.88	31.18	71.25	41.00	13.75	60.5	40.15	18.29
	Top1	77.65	32.60	28.15	72.00	52.00	9.00	62.09	35.33	19.87
	Top5	98.55	87.05	66.00	94.50	88.00	42.00	93.91	81.30	62.10
ViT-L/14	TR@1	67.50	37.15	12.80	65.50	17.00	4.00	62.55	40.48	19.48
	IR@1	61.15	50.10	48.55	68.00	9.50	2.00	62.77	50.43	33.98
	AVG	64.325	43.625	30.68	66.75	13.25	3.00	62.66	45.455	26.73
	Top1	76.35	16.15	11.30	75.00	22.00	5.00	67.09	35.36	17.99
	Top5	98.20	76.05	57.55	94.20	69.50	26.00	94.23	74.63	52.25

Table 6: The effects of universal adversarial patches on BEiT3.

MD	DS	Patch	Text Retrieval			Image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
M1	D1	w/o atk	78.98	94.38	97.20	61.36	84.61	90.74
		Patch1	0.22	0.72	1.12	1.66	4.72	6.74
		Patch2	0.20	0.58	1.00	2.12	5.29	7.53
	D2	w/o atk	93.60	99.30	99.80	82.90	96.54	98.46
		Patch1	0.70	2.70	4.10	5.42	12.42	17.10
		Patch2	0.40	1.00	2.00	1.92	4.64	7.12
	M2	w/o atk	71.00	90.16	94.44	52.79	77.12	85.07
		Patch1	1.74	4.48	6.20	5.61	13.97	19.38
		Patch2	0.86	2.76	4.46	5.10	13.40	19.40
	D2	w/o atk	96.30	99.70	100.00	86.14	97.68	98.82
		Patch1	4.10	12.10	16.90	17.92	37.02	46.80
		Patch2	1.70	3.80	6.10	11.58	24.64	32.06

5 CONCLUSION

To the best of our knowledge, this is the first work to apply decision boundary theory to the multimodal scenario. We provide a theoretical foundation for exploring the robustness of multimodal embeddings within the high-dimensional space. Based on this, a new method is proposed to improve understanding of VLP models through the development of sample-agnostic perturbations. Our method differs from sample-specific perturbation techniques by using decision boundaries to create universal perturbations. It investigates a unique interaction between visual and textual modalities, with each serving as the decision boundary for the other. We further validated the universality of the perturbation based on larger models and the global perturbation form. Experimental results

**Figure 6: Universal adversarial global perturbation results based on Flickr30k. From left to right, we utilize CLIP_{vit/B16}, CLIP_{vit/B32}, and CLIP_{vit/L14}, respectively.****Figure 7: Universal adversarial global perturbation results based on MS COCO, for verifying cross-dataset transferability.**

demonstrate promising transferability across images, tasks, and models, providing a scalable and efficient strategy to explore the security of multimedia and multimodal technologies. We hope our code can be beneficial to the community.

Algorithm 4: TIRA (combining IRA and TRA)

Input :Image set V , text set T , mask matrix m .
Output:Perturbation δ .

```

1 Initialize  $\delta \leftarrow 0$ 
// getting a batch of images each time  $\tilde{V}$ 
2 for  $\tilde{V}$  in  $V$  do
3   get matching text set  $\tilde{T}$  of  $\tilde{V}$ 
4    $r \leftarrow 0$ 
5   for  $v$  in  $\tilde{V}$  do
6     traverse  $\tilde{V}$  to execute IRA
7      $\delta \leftarrow \text{clamp}_{(0,1)}(\delta + (1 + \eta) \cdot r)$ 
8      $r \leftarrow 0$ 
9   for  $t$  in  $\tilde{T}$  do
10    traverse  $\tilde{T}$  to execute TRA
11     $\delta \leftarrow \text{clamp}_{(0,1)}(\delta + (1 + \eta) \cdot r)$ 
12 return  $\delta$ 
```

A Combining IRA and TRA

Image Retrieval universal perturbation Attack (IRA) and Text Retrieval universal perturbation Attack (TRA) can be used independently or in combination (TIRA), as shown in Alg 4. During TIRA, for each iteration, a subset \tilde{V} is extracted from the image set V . Concurrently, based on annotation information, all texts that match \tilde{V} are collected to construct the text subset \tilde{T} . IRA and TRA are executed separately based on \tilde{V} and \tilde{T} . During IRA, benign decision boundaries are constructed with $E_t(\tilde{T})$, and the optimization goal is to use universal adversarial perturbation to push $v \in \tilde{V}$ across these decision boundaries, compromising the performance of text retrieval. During TRA, malicious boundaries are established with $E_t(\tilde{V}')$ with the text input remaining unchanged. \tilde{V}' represents the subset of images where each $v \in \tilde{V}$ has been modified with the universal adversarial perturbation.

B Additional experiments based on CLIP

In the main text, $\text{UAP}_{\text{TIRA}}^{\text{patch}}$ and AdvCLIP [47] have been thoroughly compared. In fact, using IRA alone can also achieve stable attack effects. Therefore, we compare $\text{UAP}_{\text{IRA}}^{\text{patch}}$ with AdvCLIP, keeping all other experimental settings unchanged. The experimental results are shown in Tab 7, Tab 8, and Tab 9. Except for a few cases in Tab 9 where $\text{UAP}_{\text{IRA}}^{\text{patch}}$ is slightly inferior to AdvCLIP, in most situations, $\text{UAP}_{\text{IRA}}^{\text{patch}}$ achieves more effective and stable attack effects.

C Generating Global Universal Adversarial Perturbation

This section explains how to generate universal global adversarial perturbations and provides the visualizations.

C.1 Noise Constraints

Global adversarial attacks can modify any pixel in an image (equivalently understood as adding noise to the image) [9, 22]. Still, modifying pixels (the magnitude of noise) is constrained to ensure it is

difficult for human detection. And it does not alter the original semantic information of the image. Considering the average ℓ_2 and ℓ_∞ norms of images in the ImageNet validation set are approximately 5×10^4 and 250, we set $\epsilon = 2000$ for ℓ_2 norm and $\epsilon = 10$ for ℓ_∞ norm to obtain perturbations whose norm is significantly smaller than the original image norm [24]. The valid range for pixel values here is $[0, 255]$, without considering normalization. ϵ represents the noise constraint for global perturbations, i.e., $\|\delta\|_p \leq \epsilon$. For any given noise δ , to ensure $\|\delta\|_p \leq \epsilon$, a projection operation should be performed [22]. The projection strategy based on the ℓ_2 norm is:

$$P_{\ell_2}(\delta, \epsilon) = \begin{cases} \epsilon \cdot \frac{\delta}{\|\delta\|_2} & \text{if } \|\delta\|_2 > \epsilon \\ \delta & \text{if } \|\delta\|_2 \leq \epsilon \end{cases} \quad (11)$$

Algorithm 5: Global TRA

Input :Image set V , text set T , constraint ϵ , norm ℓ_p .
Output:Perturbation δ .

```

1 Initialize  $\delta \leftarrow 0$ 
2 for  $v$  in  $V$  do
3   getting  $Y = \{y_j\}_{j=1}^n$  and  $Y' = \{y'_j\}_{j=1}^k$ 
4    $v \leftarrow v + \delta$ 
5    $r \leftarrow 0$ 
6   while  $I(E_v(v + (1 + \eta) \cdot r), E_t(T), k) == 1$  do
7      $\hat{v} \leftarrow v + r$ 
8      $y'_{min} \leftarrow \underset{y' \in Y'}{\text{argmin}} f_{y'}(\hat{v})$ 
9      $y_{max} \leftarrow \underset{y \in Y}{\text{argmax}} f_y(\hat{v})$ 
10     $r \leftarrow$ 
11       $r + \frac{(\nabla_r f_{y'_{min}}(\hat{v}) - \nabla_r f_{y_{max}}(\hat{v})) \cdot (f_{y_{max}}(\hat{v}) - f_{y'_{min}}(\hat{v}))}{\|\nabla_r f_{y'_{min}}(\hat{v}) - \nabla_r f_{y_{max}}(\hat{v})\|_2^2}$ 
12     $\delta \leftarrow P_{\ell_p}(\delta + (1 + \eta) \cdot r, \epsilon)$ 
13 return  $\delta$ 
```

Algorithm 6: Global IRA

Input :Image set V , text set T , constraint ϵ , norm ℓ_p .
Output:Perturbation δ .

```

1 Initialize  $\delta \leftarrow 0$ 
2 for  $t$  in  $T$  do
3   getting  $y$  and  $Y' = \{y'_j\}_{j=1}^k$ 
4    $\tilde{V} = \{\tilde{v}_i\}_{i \in y \cup Y'} \leftarrow \{v_i + \delta\}_{i \in y \cup Y'}$ 
5    $r \leftarrow 0$ 
6   while  $I(E_t(t), E_v(\tilde{V} + (1 + \eta) \cdot r), k) == 1$  do
7      $\hat{v} = \{\hat{v}_i\}_{i \in y \cap Y'} \leftarrow \tilde{V} + r$ 
8      $y'_{min} \leftarrow \underset{y' \in Y'}{\text{argmin}} f_{y'}(t) // f_i(t) = E_v(\hat{v}_i) \cdot E_t(t)$ 
9      $r \leftarrow r + \frac{(\nabla_r f_{y'_{min}}(t) - \nabla_r f_y(t)) \cdot (f_y(t) - f_{y'_{min}}(t))}{\|\nabla_r f_{y'_{min}}(t) - \nabla_r f_y(t)\|_2^2}$ 
10     $\delta \leftarrow P_{\ell_p}(\delta + (1 + \eta) \cdot r, \epsilon)$ 
11 return  $\delta$ 
```

Table 7: Comparing between AdvCLIP and UAP_{IRA}^{patch}, based on the Flickr30k and MS COCO datasets, the recall rate is reported. Lower is better.

Visual Encoder	Method	Flickr30k(1K test set)						MS COCO(5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	AdvCLIP	2.30	7.00	11.20	19.98	48.08	64.38	1.58	4.44	6.54	8.10	23.55	34.49
	UAP _{IRA} ^{patch}	0.00	0.00	0.20	0.08	0.22	0.60	0.00	0.10	0.16	0.18	0.37	0.55
ViT-B/32	AdvCLIP	10.80	23.40	32.80	25.68	54.94	67.60	7.18	17.77	24.34	11.66	29.50	40.84
	UAP _{IRA} ^{patch}	1.00	2.80	5.50	3.00	6.44	9.18	0.60	2.34	3.74	2.39	6.02	8.83
ViT-L/14	AdvCLIP	4.80	14.40	19.20	18.28	44.18	55.82	0.84	3.18	5.06	6.84	19.32	28.87
	UAP _{IRA} ^{patch}	0.00	0.00	0.30	0.10	0.22	0.36	0.06	0.08	0.14	0.11	0.27	0.39

Table 8: Comparing between AdvCLIP and UAP_{IRA}^{patch}, based on the ImageNet, CIFAR-100 and CIFAR-10 datasets, the accuracy is reported. Lower is better.

Visual Encoder	Method	ImageNet		CIFAR100		CIFAR10	
		Top1	Top5	Top1	Top5	Top1	Top5
ViT-B/16	w/o atk	62.42	86.74	66.56	88.49	90.10	99.07
	AdvCLIP	5.98	14.80	8.91	45.04	81.90	97.01
	UAP _{IRA} ^{patch}	0.13	0.54	0.94	5.66	9.88	47.03
ViT-B/32	w/o atk	57.50	83.57	62.27	86.98	88.34	99.24
	AdvCLIP	7.98	27.40	31.92	51.84	64.45	96.51
	UAP _{IRA} ^{patch}	2.45	7.05	8.90	22.57	10.85	81.12
ViT-L/14	w/o atk	69.74	90.15	75.72	93.06	95.19	99.52
	AdvCLIP	1.54	7.61	7.42	32.30	73.87	97.95
	UAP _{IRA} ^{patch}	0.21	0.74	4.52	14.96	22.71	73.67

where P_{ℓ_2} ensures that if $\|\delta\|_2 > \epsilon$, i.e., the noise exceeds the valid boundary, it will be projected back into the defined noise range. The projection strategy based on the ℓ_∞ norm is:

$$P_{\ell_\infty}(\delta, \epsilon) = \text{clamp}_{(-\epsilon, \epsilon)}(\delta) \quad (12)$$

where $\text{clamp}_{(-\epsilon, \epsilon)}$ represents limiting the modification range of each pixel by clipping each element value in δ to be between $-\epsilon$ and ϵ .

C.2 Global Universal Attack Algorithms

We revise the pseudocode for the IRA and TRA to generate universal adversarial global perturbations, as shown in Alg 5 and Alg 6, with three main modifications: (1) When adding the perturbation δ or r to an image, the masking matrix m is unnecessary which is used to determine the location of the patch. Instead, the image and the perturbation are directly added together. (2) The adversarial patch directly replaces pixel values in a specified area of the image, it only needs to ensure that the pixel values of the patch block are

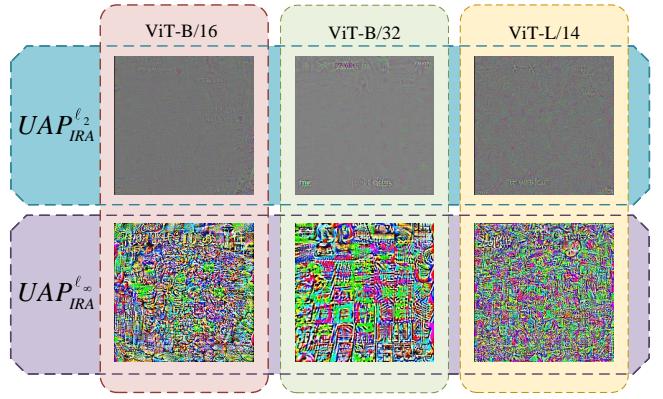


Figure 8: Visualization of six universal adversarial global perturbations.

within the valid range [11]. The global perturbation modifies all pixel values of the original image, i.e. the noise is superimposed on the original image, thus necessitating a projection function P_{ℓ_p} used to limit the extent of modifications. (3) During the testing phase (i.e., after the optimization of the universal perturbation is complete and it enters the usage stage), directly adding the universal global perturbation to each image cannot guarantee that the pixel values of each adversarial image are within the valid range. Due to this, the universal global perturbation added to each clean image undergoes varying degrees of clipping. This variability in modification might partially account for the lesser effectiveness of universal global perturbations compared to universal adversarial patches.

C.3 Visualization

We use 1000 images and 5000 texts from the Flickr30k [29] dataset as surrogate data to generate six universal global adversarial perturbations based on three CLIP [31] pre-trained models and two norm constraints. We provide the results of the attack effectiveness, as seen in Tab 10, along with visualization of the perturbations in Fig 8.

Table 9: Comparing between AdvCLIP and UAP_{IRA}^{patch}, based on the NUS-WIDE, Pascal-Sentence and Wikipedia datasets.

Visual Encoder	Metric	NUS-WIDE			Pascal-Sentence			Wikipedia		
		w/o atk	AdvCLIP	UAP _{IRA} ^{patch}	w/o atk	AdvCLIP	UAP _{IRA} ^{patch}	w/o atk	AdvCLIP	UAP _{IRA} ^{patch}
ViT-B/16	TR@1	66.90	36.75	12.65	67.00	7.00	6.50	61.69	14.72	9.96
	IR@1	61.80	51.25	38.15	68.50	10.00	8.50	66.67	42.64	32.68
	AVG	64.35	44.00	25.4	67.75	8.50	7.50	64.18	28.68	21.32
	Top1	77.85	49.65	12.15	77.00	5.50	8.00	64.01	27.95	8.41
	Top5	98.75	72.95	55.75	96.00	39.50	32.00	94.35	73.15	55.60
ViT-B/32	TR@1	69.25	34.15	31.85	72.00	35.00	14.50	58.23	31.60	19.91
	IR@1	60.25	47.60	49.95	70.50	47.00	15.00	62.77	48.70	42.64
	AVG	64.75	40.88	40.90	71.25	41.00	14.75	60.5	40.15	31.28
	Top1	77.65	32.60	23.90	72.00	52.00	10.00	62.09	35.33	21.99
	Top5	98.55	87.05	65.75	94.50	88.00	48.50	93.91	81.30	62.61
ViT-L/14	TR@1	67.50	37.15	12.10	65.50	17.00	1.50	62.55	40.48	10.17
	IR@1	61.15	50.10	18.40	68.00	9.50	5.50	62.77	50.43	26.41
	AVG	64.325	43.625	15.25	66.75	13.25	3.50	62.66	45.455	18.29
	Top1	76.35	16.15	11.15	75.00	22.00	5.00	67.09	35.36	8.98
	Top5	98.20	76.05	51.00	94.20	69.50	26.00	94.23	74.63	42.31

Table 10: The effects of six universal adversarial global perturbations.

Visual Encoder	Method	Flickr30k(1K test set)						MS COCO(5K test set)					
		TR			IR			TR			IR		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/16	w/o atk	82.20	96.70	98.90	62.12	85.72	91.94	52.40	76.78	84.66	33.07	58.35	69.03
	UAP _{IRA} ^{ℓ_∞}	38.80	62.30	71.80	9.40	17.31	22.56	15.76	31.80	40.00	6.73	14.39	18.78
	UAP _{IRA} ^{ℓ₂}	36.00	57.70	67.4	8.36	15.64	21.04	14.30	27.72	34.78	5.45	11.62	15.80
ViT-B/32	w/o atk	78.60	94.90	98.30	58.80	83.56	90.00	50.18	75.02	83.50	30.48	55.99	66.87
	UAP _{IRA} ^{ℓ_∞}	41.00	64.00	70.70	13.22	24.46	30.86	17.70	32.60	40.96	7.59	16.70	22.35
	UAP _{IRA} ^{ℓ₂}	13.70	25.10	32.20	12.00	21.12	26.04	5.80	12.26	16.40	6.06	12.92	17.10
ViT-L/14	w/o atk	85.30	97.40	99.20	64.84	87.22	92.04	56.34	79.40	86.58	36.51	61.08	71.17
	UAP _{IRA} ^{ℓ_∞}	42.40	64.10	69.30	8.20	14.44	18.60	16.32	31.34	38.48	4.43	8.55	10.78
	UAP _{IRA} ^{ℓ₂}	24.90	39.80	46.70	14.12	24.22	29.16	7.76	16.16	20.74	5.19	10.52	13.37

D Additional experiments based on BEiT3

We further conduct detailed experiments on BEiT3 [37]. First, some configurations need to be determined: (1) We consistently use TIRA, in which TRA and IRA are set with $k = 10$ because the Text Retrieval R@10 and Image Retrieval R@10 metrics will be evaluated. (2) For generating universal adversarial patches, we set the patch size to 66×66 (BEiT3 requires an input image size of 384×384 , thus the patch covers 3% of the image area). (3) The patch is located at the bottom right corner of the image, with the bottom right corner of the patch having an offset of $(x, y) = (-24, -24)$ relative to the bottom right corner of the image. (5) For generating universal global perturbations, we use ℓ_2 norm to constraint the noise, setting $\epsilon = 2000$.

The “BEiT3-base” model is a pre-trained model obtained through the Masked Data Modeling (MDM) pre-training task. It is further fine-tuned on the Image-Text Contrastive (ITC) task to obtain the “BEiT3-base-itc” pre-trained model. Building on “BEiT3-base-itc”, the model is fine-tuned separately on the MS COCO dataset [19] and the Flickr30k dataset [29], resulting in two models specifically for image-text retrieval task, which we denote as M_1 and M_2 , respectively. We consistently conduct the white-box attack to generate universal perturbation based on the M_1 . M_2 is reserved for testing the cross-model transferability of the universal perturbation.

Table 11: The effects of universal global perturbations on BEiT3.

MD	DS	Patch	Text Retrieval			Image Retrieval		
			R@1	R@5	R@10	R@1	R@5	R@10
M1	D1	w/o atk	78.98	94.38	97.20	61.36	84.61	90.74
		Noise1	17.72	29.90	35.48	21.85	40.62	50.00
		Noise2	18.94	30.96	36.68	22.92	42.15	51.55
	D2	w/o atk	93.60	99.30	99.80	82.90	96.54	98.46
		Noise1	44.20	63.10	69.20	43.00	67.10	76.10
		Noise2	40.80	56.30	63.80	44.66	68.50	76.56
M2	D1	w/o atk	71.00	90.16	94.44	52.79	77.12	85.07
		Noise1	22.34	36.86	43.92	20.30	39.17	49.20
		Noise2	22.62	37.36	45.00	19.70	37.29	46.71
	D2	w/o atk	96.30	99.70	100.00	86.14	97.68	98.82
		Noise1	56.60	75.10	79.60	57.58	82.08	87.84
		Noise2	53.10	69.70	76.00	56.24	78.72	85.90

D.1 Generating Universal Adversarial Patch

We use 1000 training images from the MS COCO dataset and their matching texts, as well as 1000 training images and their matching texts from the Flickr30k dataset, to generate two universal adversarial patches, respectively denoted as *Patch1* and *Patch2*. A comprehensive evaluation of their attack effectiveness is presented in Tab 6. The first column indicates the model under test, and the second row specifies the dataset under test, where *D1* represents the MS COCO test set, and *D2* represents the Flickr30k test set. The third row indicates which patch is being applied to the entire set of test images. The experimental results validate that our universal adversarial patches possess stable and potent attack effectiveness, as well as good cross-dataset and cross-model transferability.

D.2 Generating Universal Global Perturbation

We use 1000 training images from the MS COCO dataset and their matching texts, as well as 1000 training images and their matching texts from the Flickr30k dataset, to generate two universal adversarial global perturbations, respectively denoted as *Noise1* and *Noise2*. A comprehensive evaluation of their attack effectiveness is presented in Tab 11. The experimental results validate that our universal adversarial global perturbations possess stable and potent attack effectiveness, as well as good cross-dataset and cross-model transferability.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*.
- [2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhajit Som, Songhao Piao, and Furu Wei. 2022. VLMO: unified vision-language pre-training with mixture-of-modality-experts. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 32897–32912.
- [3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: universal image-text representation learning. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 104–120.
- [4] Zhuangzhuang Chen, Jin Zhang, Zhuonian Lai, Jie Chen, Zun Liu, and Jianqiang Li. 2022. Geometry-aware guided loss for deep crack recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4693–4702.
- [5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, 1–9.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Li Kai, and Fei-Fei Li. 2009. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*.
- [8] Alexey Dosovitskiy et al. 2023. An image is worth 16x16 words: transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.
- [10] Liang Jiawei, Liang Siyuan, Luo Man, Liu Aishan, Han Dongchen, Chang Ee-Chien, and Cao Xiaochun. 2024. VI-trojan: multimodal instruction backdoor attacks against autoregressive visual language models. In *arXiv preprint arXiv:2402.13851*.
- [11] Danny Karmon, Daniel Zoran, and Yoav Goldberg. 2018. LaVAN: localized and visible adversarial noise. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2507–2515.
- [12] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: vision-and-language transformer without convolution or region supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 5583–5594.
- [13] A. Krizhevsky and G. Hinton. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*.
- [14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 12888–12900.
- [15] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: vision and language representation learning with momentum distillation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*, 9694–9705.
- [16] Wenrui Li, Zhengyu Ma, Liang-Jian Deng, Penghong Wang, Jinqiao Shi, and Xiaopeng Fan. 2023. Reservoir computing transformer for image-text retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 5605–5613.
- [17] Wenrui Li, Penghong Wang, Ruiqin Xiong, and Xiaopeng Fan. 2024. Spiking tucker fusion transformer for audio-visual zero-shot learning. *IEEE Transactions on Image Processing*, 1–1.
- [18] Xiujun Li et al. 2020. Oscar: object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 121–137.
- [19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proceedings of the 13th European Conference on Computer Vision (ECCV)*, 740–755.
- [20] Fangyu Liu, Rémi Lebret, Didier Orel, Philippe Sordet, and Karl Aberer. 2020. Upgrading the newsroom: an automated image selection system for news articles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 1–28.
- [21] 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 (ICCV)*, 102–111.
- [22] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.
- [23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1765–1773.
- [24] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–2582.
- [25] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. 2017. Fast feature fool: a data independent approach to universal adversarial perturbations. In *arXiv preprint arXiv:1707.05572*.

- [26] Konda Reddy Mopuri, Phani Krishna Uppala, and R Venkatesh Babu. 2018. Ask, acquire, and attack: data-free uap generation using class impressions. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 19–34.
- [27] Liang Peng, Yang Yang, Zheng Wang, Zi Huang, and Heng Tao Shen. 2020. Mra-net: improving vqa via multi-modal relation attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 318–329.
- [28] Zhihang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. BEiT v2: masked image modeling with vector-quantized visual tokenizers. In *arXiv preprint arXiv:2208.06366*.
- [29] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2641–2649.
- [30] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 742–751.
- [31] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 8748–8763.
- [32] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon’s mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s Mechanical Turk (NAACL-HLTW)*, 139–147.
- [33] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Covillo, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia (ACM MM)*, 251–260.
- [34] Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. 2020. Universal adversarial training. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, 5636–5643.
- [35] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*.
- [36] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. 2022. Vlmixer: unpaired vision-language pre-training via cross-modal cutmix. In *Proceedings of the International Conference on Machine Learning (ICML)*, 22680–22690.
- [37] Wenhui Wang et al. 2023. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19175–19186.
- [38] Jiwei Wei, Yang Yang, Xing Xu, Jingkuan Song, Guoqing Wang, and Heng Tao Shen. 2023. Less is better: exponential loss for cross-modal matching. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 5271–5280.
- [39] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15671–15680.
- [40] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, and Alexander G Hauptmann. 2018. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Transactions on Multimedia (TMM)*, 1276–1288.
- [41] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faletta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6995–7004.
- [42] 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 (ACM MM)*, 5005–5013.
- [43] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yeqin Choi, and Jianfeng Gao. 2021. Vinvl: revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5579–5588.
- [44] Haonan Zheng, Xinyang Deng, Wen Jiang, and Wenrui Li. 2024. A unified understanding of adversarial vulnerability regarding unimodal models and vision-language pre-training models. (2024). arXiv: 2407.17797 [cs.CV].
- [45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16816–16825.
- [46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2337–2348.
- [47] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. 2023. AdvCLIP: downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*, 6311–6320.