# Attacking Attention of Foundation Models Disrupts Downstream Tasks

Hondamunige Prasanna Silva
University of Florence
hondamunige.silva1@edu.unifi.it

Federico Becattini
University of Siena
federico.becattini@unisi.it

Lorenzo Seidenari
University of Florence
lorenzo.seidenari@unifi.it

## Abstract

*Foundation models represent the most prominent and recent paradigm shift in artificial intelligence. Foundation models are large models, trained on broad data that deliver high accuracy in many downstream tasks, often without fine-tuning. For this reason, models such as CLIP [36], DINO [35] or Vision Transfomers (ViT) [10], are becoming the bedrock of many industrial AI-powered applications. However, the reliance on pre-trained foundation models also introduces significant security concerns, as these models are vulnerable to adversarial attacks. Such attacks involve deliberately crafted inputs designed to deceive AI systems, jeopardizing their reliability. This paper studies the vulnerabilities of vision foundation models, focusing specifically on CLIP and ViTs, and explores the transferability of adversarial attacks to downstream tasks. We introduce a novel attack, targeting the structure of transformer-based architectures in a task-agnostic fashion. We demonstrate the effectiveness of our attack on several downstream tasks: classification, captioning, image/text retrieval, segmentation and depth estimation.*

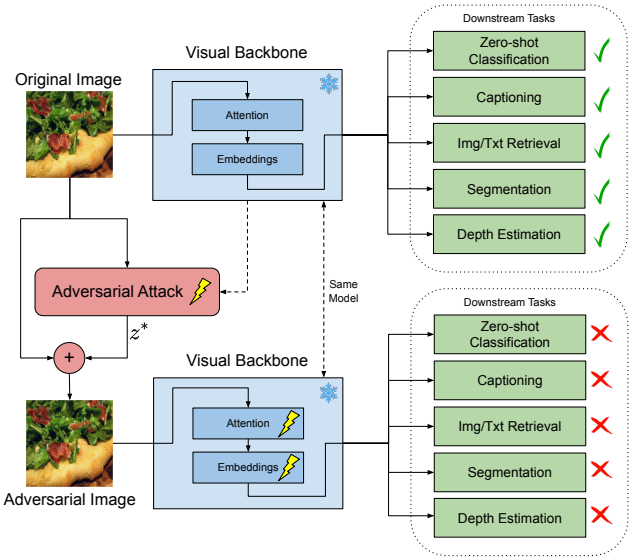


Figure 1. Our method generates adversarial noise $z^*$ by attacking attention and embeddings in visual backbones of foundation models, without any knowledge of downstream tasks. Unlike other attacks, access to other modalities (e.g., text) is not required.

## 1. Introduction

Foundation models are nowadays powering the vast majority of AI-based algorithms and technologies. Foundation models are large models that are trained on broad data and are easily fine-tuned to downstream tasks. A trend started with the first "off-the-shelves" image recognition models pre-trained on Imagenet [39], is now brought forward by remarkable advancements by large language models [4, 8], vision backbones [35] and multimodal models [36].

Among the foundation models, CLIP [36] can be regarded as one of the most prominent players in all modern computer vision applications. It relies on contrastive learning to learn an optimal alignment between textual and visual embeddings, and it is widely used in downstream tasks including classification, retrieval and captioning [25]. Like most foundation models, CLIP relies on attention, which weighs different parts of the input and also plays a role in spatial information handling. Similarly, pretrained DINO models [35] are widely used as feature extractors, especially for fine-grained tasks such as segmentation or depth estimation. However, foundational models, like all deep learning models, are vulnerable to adversarial attacks. In fact, the widespread application of single models throughout an array of different tasks makes attacks potentially more dangerous, as a single attack could affect multiple applications. Nonetheless, attacks on foundation models are currently successfully performed by exploiting knowledge about in-

dividual downstream tasks [29]. When the foundational model is multimodal, e.g. image and text, joint attacks on both modalities can also be conducted [43] to improve their effectiveness. However, this requires access to both modalities, making the attacks not always feasible in practice.

In this work, we propose an attack for foundation models, specifically focusing on CLIP and Vision Transformer backbones, by taking a different angle on the problem. Unlike existing methodologies, which are often multimodal and task-specific, we design our attack to target only the visual domain and to be agnostic to the downstream task. In fact, we believe that to make the attack more disruptive, foundational models should be attacked independently from the downstream tasks. Our method, by jointly damaging attention and the final model embeddings, can attack with the same perturbation all downstream applications, as shown in Fig. 1. We show that tasks like zero-shot classification, retrieval, captioning, segmentation and depth estimation are all affected by our simple attack. These results highlight the potential fragility of foundation models. Moreover, attacks in the image domain are harder to spot especially when conducted with low budgets, compared to attacks that perturb both images and text. Interestingly, attacks targeting specific parts of attention layers have different impacts, depending on the nature of the downstream task: fine-grained tasks, as captioning or segmentation/depth estimation, are better addressed by disrupting visual attention, whereas more semantic tasks, as classification and retrieval, are better targeted by attacks on embeddings. In summary, our main contributions are the following:

- We propose an effective adversarial perturbation algorithm targeting attention of vision backbones in foundational models.
- Our method is task agnostic, i.e. can attack samples without their label or textual description.
- Extensive experimentation shows that our method can affect downstream tasks such as captioning, classification, retrieval, segmentation and depth estimation, reporting state-of-the-art results in terms of attack success rate.

## 2. Related works

**Vision Foundation Models** Vision Foundation Models, have become pivotal in advancing multimodal AI by effectively addressing downstream tasks across various domains. These models, trained on large-scale image-text datasets, learn representations that can be fine-tuned for specific tasks, significantly improving performance. Whereas there are examples of purely visual foundation models [10, 21, 35], most of these models are multi-modal, exploiting a Vision-Language Pre-training (VLP) that jointly learns to process images and text. CLIP [36] is a landmark model in this field, enabling zero-shot open-vocabulary image classification by aligning images with textual descriptions. This model set a new benchmark for generalization in vision-language tasks. Following CLIP, ALIGN [18] scaled this approach, improving the performance on tasks like image retrieval and cross-modal retrieval by leveraging billions of image-text pairs. Florence [47] exemplifies the application of VLPs in multi-task learning. Trained on a vast dataset, it excels in object detection, semantic segmentation, and image captioning, demonstrating the benefits of joint vision-language optimization. Regarding purely visual backbones, DINO [7] and its improved version DINOv2 [35], introduced a self-supervised approach to train Vision Transformers, yielding extremely versatile backbones that can be used across tasks even without fine-tuning. In this work, we focus on attacking CLIP and Vision Transformer backbones as they are largely employed as core building blocks of a plethora of downstream applications.

**Adversarial Attacks on ViT** With the extensive use of Vision Transformers (ViTs) [10] as the core of various vision foundation models [21, 36], there is a growing interest in evaluating their robustness to pinpoint their weaknesses. The Pay No Attention (PNA) [44] method adapts the Skip Gradient Method (SGM) for ViTs, specifically omitting the gradient of the attention block during back-propagation to enhance the transferability of adversarial examples through gradient regularization. The attack also introduces the PatchOut strategy, which randomly selects a subset of patches to compute the gradient in each attack iteration, serving as an image transformation technique for transferable adversarial attacks on ViTs. Another notable method is the Token Gradient Regularization (TGR) [49]: leveraging the structural characteristics of ViTs, TGR reduces the variance of back-propagated gradients on a token-wise basis and utilizes the regularized gradient to generate adversarial samples. Contrasting with these strategies, Naseer et al. [33] introduced the Self-Ensemble (SE) method and the Token Refinement (TR) module to boost the transferability of adversarial examples generated by ViTs. SE leverages the classification token at each ViT layer with a shared classification head for feature-level attacks. Building on SE, TR further refines the classification token through fine-tuning to improve attack performance.

In general, the literature presents conflicting conclusions regarding the adversarial robustness of transformers, compared to the one of Convolutional Neural Networks (CNNs). One school of thought posits that transformers demonstrate superior robustness compared to CNNs. Recent studies [1, 19, 38] attribute this robustness to the differing reliance on input features. Specifically, CNNs depend heavily on high-frequency information, whereas transformers do not. Consequently, transformers are believed to be more resilient to gradient-based attacks like Fast Gradient Sign Method (FGSM) [14], Projected Gradient De-

scent (PGD) [30], and Carlini & Wagner (C&W) [5] attacks. Furthermore, the severe nonlinearity inherent in the input-output relationships of transformers may contribute to their enhanced robustness [20]. In scenarios involving adversarial training, Vision Transformers (ViTs) exhibit superior generalization and robustness compared to CNNs, as highlighted by Liu et al. [27].

Contrarily, another body of research contends that transformers are just as susceptible to adversarial attacks as CNNs. The findings of Mahmood et al. [31] indicate that ViTs do not outperform ResNet architectures in terms of robustness against various attack methods including FGSM [14], PGD [30], and C&W [5]. Similarly, there is evidence that both CNNs and transformers exhibit comparable vulnerabilities to natural and adversarial perturbations [3], and a direct comparison under a unified training setup revealed that both architectures possess similar levels of adversarial robustness [2]. Additional vulnerabilities of Vision Transformers are exposed through targeted patch attacks [13, 15]. Recently, EmbedAttack [22] was proposed as the first task agnostic attack on ViT backbones, by simply applying PGD on the embeddings rather than on the targets. In this work, differently from EmbedAttack, we target ViT backbones of foundation models by attacking its core component: attention. Our attack crafts adversarial noise to make the model focus on irrelevant parts of the image, yielding performance drops across several downstream tasks, without knowing a-priori which task the backbone could be used for. In addition, we also show that attacking different parts of the attention layer (e.g., the attention matrix of the final embedding), yields different degrees of efficacy depending on the nature of the downstream task.

**Attacks on Multimodal Foundation Models** Unlike standard attacks, which typically target the task by exploiting the relationship between a single instance and its label, attacks on vision-language models must account for multi-modal data relationships. Fort [11] demonstrated that "standard" adversarial attacks are effective on CLIP models when customized for a particular downstream task. Other studies [12, 34] examined attacks on multi-modal CLIP neurons by placing text stickers on images, prompting the CLIP model to "read" the text and ignore the rest of the image. These attacks adhere to a different threat model compared to traditional budget-based attacks, as the adversarial manipulations are quite evident. Notably, [6] recently revealed that poisoning attacks on web-scale datasets are feasible and can significantly affect the training process of foundational models like CLIP. Zhang et al. [48] introduced the Collaborative Multimodal Adversarial Attack (Co-Attack) to target various pre-trained models, such as CLIP, ALBEF [24], and TCL [45]. Co-Attack generates multi-modal adversarial perturbations by increasing the embedding distance between adversarial examples and their original data pairs. The Set-level Guidance Attack (SGA) [29] uses data augmentation to boost data diversity, thereby creating multi-modal adversarial perturbations by minimizing the similarity between them and their matched data from another modality. Additionally, the Self-augment-based Transfer Attack (SA-Attack) [16] applies further data augmentations to both the original data and the adversarial examples to improve attack transferability across models. While Co-Attack, SGA, and SA-Attack typically generate adversarial perturbations sequentially, first for one modality and then for another, methods for learning multi-modal adversarial perturbations simultaneously also exist [43]. All aforementioned attacks depend on task-specific characteristics, hindering their broader applicability. In contrast, our proposed method makes a significant advancement by adopting a task-agnostic approach. It simultaneously disrupts both the attention mechanism and the image embedding within attention-based models, without relying on specific task labels or characteristics. This dual disruption effectively shifts the embedding away from its correct representation and manipulates the attention distribution across the input, making it applicable to a wide range of downstream tasks. By targeting the foundational backbone rather than the task-specific components, our method provides a more robust, scalable, and universally applicable adversarial strategy. It also simplifies the attack pipeline on multimodal models, by solely attacking the vision branch. Whereas leveraging a joint attack could yield more effective disruptions, not accessing other modalities greatly improves the attack applicability, as no additional information on the data other than the image is required.

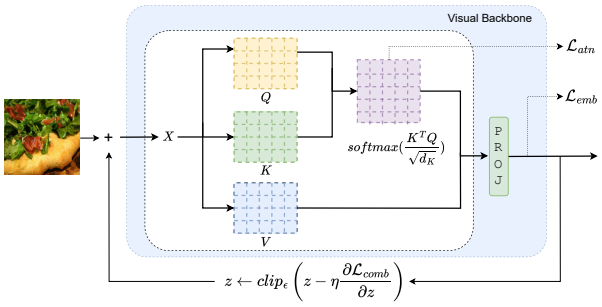


Figure 2. Our attack disrupts both the attention mechanism and image embedding in attention-based models.

## 3. Attacking Foundation Models

We introduce a novel adversarial attack that simultaneously targets the attention mechanism and image embedding in attention-based models, as shown in Fig. 2. The core objective of our approach is to disrupt the input's representation by shifting it away from its original embedding and altering

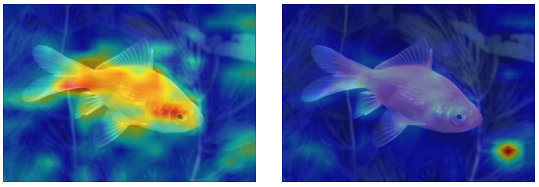

Figure 3. Attention on clean (left) and perturbed (right) image.

how the model distributes attention across different parts of the input. Although simple, what makes our attack strategy appealing is that it is completely agnostic to ground truth labels and, since it is tailored to attack foundational backbones, it is also agnostic to the nature of any downstream task. Our proposed attack is versatile and can be applied to any vision transformer or attention-based model. However, motivated by its agnostic nature, we focus on foundational models, which are frequently leveraged either by tuning only task-specific heads or in zero- or few-shot settings. The main idea of our method is an attack that can target said backbones, and by damaging the original model, also disrupts the performance of any downstream task that builds upon its learned representations.

### 3.1. Optimizing the Perturbation

Given a model $f(x)$ with input $x$, our objective is to generate a perturbation $z$ such that the adversarial example $x' = x + z$ causes the model output to diverge from its intended outcome. To achieve this, we optimize the perturbation $z$ by considering the gradients of both the attention mechanism and the embedding space with respect to the input. The optimization problem can be formally defined as

$$z^* = \arg\min_z \mathcal{L}_{comb}(f(x+z)) \tag{1}$$

where $\mathcal{L}_{comb}$ is a combined loss that integrates both attention and embedding losses, as described further. The perturbation is iteratively updated using gradient descent:

$$z \leftarrow \text{clip}_\epsilon\left(z - \eta \frac{\partial \mathcal{L}_{comb}}{\partial z}\right) \tag{2}$$

where $\eta$ is the learning rate, computed using Weighted Adam [28] and $\epsilon$ is the perturbation budget. The function $\text{clip}_\epsilon(\cdot)$ is the clipping function that ensures that adversarial noise does not exceed the budget at every iteration, that is, $\|z\|_\infty < \epsilon$. The optimization alters the embedding of the perturbed input while simultaneously altering the attention distribution.

### 3.2. Attention Manipulation

Our approach manipulates the attention matrix in transformers to redirect the model's focus from significant tokens to less relevant ones, thereby impairing its performance, as shown in Fig. 3. The attention mechanism in transformers projects the input into three embedding vectors in $\mathbb{R}^{d_k}$, the query $\mathbf{Q}$, the key $\mathbf{K}$ and the value $\mathbf{V}$, and computes an attention matrix $\mathbf{A}$ as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right) \tag{3}$$

which is then multiplied by $\mathbf{V}$ to compute the output values.

To disrupt the attention of the model, we define a loss $\mathcal{L}_{atn}$ that minimizes the alignment between the clean attention matrix $\mathbf{A}^{gt}$ and the adversarial attention matrix $\mathbf{A}^{adv}$ for a given layer $l$. This loss is formulated as:

$$\mathcal{L}_{atn} = \sum_{h=1}^{N_h} \frac{1}{(N_t-1)^2} \sum_{r=2}^{N_t}\sum_{c=2}^{N_t} A^{gt}_{h,r,c} \cdot A^{adv}_{h,r,c} \tag{4}$$

where $N_h$ is the number of attention heads, and $N_t$ is the number of tokens. The terms $A^{gt}_{h,r,c}$ and $A^{adv}_{h,r,c}$ represent the attention values at head $h$, row $r$, and column $c$ for the ground truth and adversarial attention matrices, respectively. The indices $r, c \geq 2$ indicate rows and columns of each attention map of shape $N_t \times N_t$, where $N_t$ corresponds to the number of visual tokens in each row/column, excluding the CLS token. In our experiments, we attack the attention map from the last layer. In a set of preliminary experiments, we found that this strategy performs best, with earlier layers yielding slightly worse results.

By minimizing the dot product between the clean attention $\mathbf{A}^{gt}$ and the adversarial attention $\mathbf{A}^{adv}$, our loss function effectively alters adversarial attention making it orthogonal to clean attention. The main idea is that the original attention patterns are to be pushed away from the directions of high importance identified by the model.

This effect is applied to each token in the attention matrix. Specifically, for each token $t$, minimizing the dot product between the attention row corresponding to $t$ in $\mathbf{A}^{gt}$ and $\mathbf{A}^{adv}$ forces the adversarial perturbation to produce an orthogonal vector. As a consequence, where the clean attention would have a high value for a particular token, the adversarial perturbation induces a low value, thereby diminishing the attention on that token. Due to the nature of the softmax operation applied after the multiplication of query ($\mathbf{Q}$) and key ($\mathbf{K}$), this reduction in attention causes a redistribution of the attention scores across other tokens.

This mechanism is particularly effective because it not only changes the distribution of attention allocation but also directly impacts the magnitude of attention weights. By altering both aspects, the attack ensures that the attention is redistributed from the originally highly attended tokens to those that were previously less significant. This redistribution impairs the model ability to focus on relevant parts of the input, thereby degrading its overall performance and making the attack more powerful.

### 3.3. Embedding Space Disruption

To enhance the effectiveness of our adversarial attack, we not only manipulate the attention mechanism but also introduce a disruption in the embedding space. This approach allows us to directly alter the learned representations that are crucial for the model's performance across various tasks. By targeting the embedding space, we affect the foundational representation that informs subsequent decision-making layers.

Existing attacks in the embedding domain have access not just to the image but also to the text. Moreover, attacks are often carried out jointly on the two modalities, altering image and text [16, 29, 48]. CO-Attack [48] crafts a multimodal embedding combining the two modalities and relies on the KL-Divergence to distance adversarial embeddings from unattacked image embeddings. SGA [29] maximizes the misalignment, in terms of cosine distance, between text and image embedding. Finally, SA-Attack [16] extends SGA through augmentation, thus leveraging input diversity to disrupt the alignment. In contrast, we attack a single modality, minimizing the dot product between adversarial embeddings and clean embeddings. Even without accessing or modifying the text embeddings, lowering the alignment of clean image embeddings and adversarial embeddings implicitly affects the alignment with textual embedding which are aligned by construction in models like CLIP [36]. We define the embedding loss, $\mathcal{L}_{emb}$, as the $L_2$ norm between the clean and adversarial embeddings:

$$\mathcal{L}_{emb} = \left\| E^{gt} - E^{adv} \right\|_2 \tag{5}$$

To calculate the adversarial embedding $E^{adv}$, we first create an adversarial example by adding a perturbation $z^*$ to the input:

$$X^{adv} = X + z^* \tag{6}$$

Then, we pass $X^{adv}$ through the model to obtain $E^{adv}$:

$$E^{adv} = f_l(X^{adv}) = f_l(X + z^*) \tag{7}$$

This gives us the adversarial embedding $E^{adv}$ from the transformer layer $l$ inside the vision encoder.

### 3.4. Combined Loss Function

To generate a robust adversarial perturbation, we combine the attention and embedding losses into a single objective:

$$\mathcal{L}_{comb} = \alpha \mathcal{L}_{atn} + \beta \mathcal{L}_{emb} \tag{8}$$

where $\alpha$ and $\beta$ are hyperparameters that balance the contributions of each loss. By optimizing this combined loss, our approach effectively perturbs both the attention mechanism and embedding space, resulting in a powerful adversarial attack that compromises the integrity of foundational models and their downstream applications. The method optimizes the adversarial attack using Weighted Adam with an initial learning rate of $0.01$ for 250 iterations. The losses $\mathcal{L}_{emb}$ and $\mathcal{L}_{atn}$ were balanced at every iteration by setting $\alpha = 1; \beta = \alpha |\frac{\mathcal{L}_{atn}}{\mathcal{L}_{emb}}|$.

| Attacks | ViT-B/16 | ViT-B/32 | ViT-L/14 | Label |
|---|---|---|---|---|
| **FGSM** [14] | 99.90 % | 100% | 99.70% | ✓ |
| **PGD** [30] | 99.90% | 100% | 99.80% | ✓ |
| **Ours** ($\mathcal{L}_{emb}$) | 97.90% | 96.50% | 98.10% | - |
| **Ours** ($\mathcal{L}_{atn}$) | 86.40% | 85.50% | 83.50% | - |
| **Ours** ($\mathcal{L}_{comb}$) | **98.40%** | **97.40%** | **98.30%** | - |

Table 1. Attack success rate for classification on ImagenetV2 test set. $\epsilon = 2/255$

## 4. Experiments

We evaluate our approach by covering a wide range of downstream tasks, assessing the impact of our attack. We analyze the effects on classification, retrieval, captioning, zero-shot classification, depth estimation and semantic segmentation. Note that, when different experiments share the same backbone (e.g., ViT-B/16), a single attack suffices to affect all of the described downstream tasks.

**Classification** We evaluate our attack on image classification using Vision Transformers (ViTs) with different architectures. Specifically, we focus on ViT-B/16, ViT-B/32, and ViT-L/14 [10], comparing the attack against two widely used targeted adversarial techniques: Fast Gradient Sign Method (FGSM) [14] and Projected Gradient Descent (PGD) [30]. Tab. 1 shows that our attack yields slightly lower success rates. Nonetheless, differently from FGSM and PGD, our method does not require label information, highlighting its effectiveness and broader applicability.

**Image/Text Retrieval** In Tab. 2, we present a comparison of our method with state-of-the-art approaches on the image/text retrieval tasks using the COCO [26] and Flickr30K [46] datasets by attacking CLIP (ViT-B/16). As evaluation metric, we use attack success rate at **R@1**, **R@5**, and **R@10**. The term **TR-R@K** refers to the percentage of adversarial images for which the correct caption (among the ground truth captions) is not retrieved in the top-K by the VLP model. Similarly, **IR-R@K** indicates the percentage of texts for which the correct image is not found in the top-K results when the attack is performed. To enable this attack with our architecture, we attack all images in the gallery, leaving the input text unaltered. Columns labeled "T", "I" and "J" indicate whether the attack targets text (T), image (I), or both jointly (J). For both the SGA attack [29] and Co-Attack [48] baselines, results are provided for two distinct

| Attacks | COCO (Text Retrieval) | | | Flickr30K (Text Retrieval) | | | COCO (Image Retrieval) | | | Flickr30K (Image Retrieval) | | | T | I | J |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | | | |
| **Sep Attack** [48] | 69.10% | 52.20% | 43.64% | 49.57% | 23.99% | 15.65% | 78.64% | 66.89% | 60.66% | 66.98% | 51.83% | 45.31% | ✓ | ✓ | - |
| **SGA Attack** [29] | 99.85% | 99.42% | 98.89% | 99.14% | 97.40% | 95.73% | 99.85% | 99.37% | 98.88% | 99.10% | 97.64% | 96.16% | ✓ | ✓ | ✓ |
| **SA Attack** [16] | 99.58% | 98.74% | 98.01% | 98.65% | 96.26% | 93.09% | 99.62% | 99.06% | 98.35% | 99.00% | 96.87% | 94.88% | ✓ | ✓ | ✓ |
| **Co-Attack** [48] | 97.98% | 94.94% | 93.00% | 93.25% | 84.88% | 78.96% | 98.80% | 96.83% | 95.33% | 95.86% | 90.83% | 87.36% | ✓ | ✓ | ✓ |
| **Uni$_A$**$^*$ [50] | 95.50% | 91.42% | 88.82% | 91.90% | 82.87% | 78.66% | 94.24% | 90.11% | 87.65% | 91.14% | 81.65% | 79.96% | - | ✓ | - |
| **Mul$_A$**$^*$ [50] | 95.50% | 91.60% | 88.87% | 92.02% | 82.04% | 78.66% | 96.13% | 93.50% | 91.76% | 94.85% | 90.42% | 86.22% | ✓ | ✓ | - |
| **ETU**$^*$ [50] | 93.55% | 89.01% | 85.88% | 88.47% | 78.61% | 73.07% | 94.25% | 91.30% | 89.31% | 92.69% | 87.50% | 84.89% | ✓ | ✓ | - |
| **SGA Attack** [29] | 97.30% | 94.02% | 91.17% | 94.11% | 88.89% | 83.64% | 97.72% | 94.65% | 91.86% | 95.91% | 90.10% | 85.98% | - | ✓ | - |
| **Co-Attack** [48] | 94.93% | 90.19% | 86.52% | 87.73% | 78.09% | 72.05% | 95.88% | 91.58% | 88.55% | 91.72% | 83.32% | 78.67% | - | ✓ | - |
| **Ours** ($\mathcal{L}_{emb}$) | **99.81%** | 99.50% | **99.29%** | **99.26%** | 98.03% | 97.36% | 99.42% | **98.72%** | **98.21%** | 98.61% | 97.01% | 95.81% | - | ✓ | - |
| **Ours** ($\mathcal{L}_{atn}$) | 97.10% | 93.54% | 91.33% | 89.82% | 80.48% | 74.80% | 95.36% | 90.63% | 87.26% | 89.79% | 78.37% | 72.20% | - | ✓ | - |
| **Ours** ($\mathcal{L}_{comb}$) | **99.81%** | **99.53%** | **99.29%** | 99.14% | **98.55%** | **97.66%** | **99.43%** | 98.61% | 98.11% | **98.87%** | **97.41%** | **96.16%** | - | ✓ | - |

Table 2. Attacks to CLIP on retrieval on COCO and Flickr30K. Universal perturbation attack with a 12/255 budget are marked with$^*$. Remaining methods use $\epsilon = 2/255$. Best results not using joint information are in **bold**. Best results using joint information are underlined.

settings: the image-only attack and the joint image-and-text attack. The image-only setting applies perturbations solely to the image (I), while the joint setting perturbs both the image and text modalities (I, T).

In the text retrieval task, our proposed method achieves a high attack success rate, surpassing joint modality methods such as SGA Attack [29], SA Attack [16], and Co-Attack [48]. Furthermore, when SGA and Co-Attack operate in the image-only setting, their performance drops significantly. Our method consistently outperforms them, with the only exception being that SGA achieves a slightly higher attack success rate at R@1 on COCO. Similarly, for image retrieval, our method performs on par with the joint modality competitors, achieving high attack success rates, with R@1 values of 99.43% on the COCO dataset and 98.87% on the Flickr30K dataset. Again, in the image-only setting, our method outperforms both SGA and Co-Attack.

**Image Captioning** Table 3 presents the impact of adversarial attacks on image captioning for the BLIP-2 model[1] [25], evaluated on COCO and Flickr30k. Performance is measured using BLEU (1–4), METEOR, ROUGE, CIDEr, and SPICE, which collectively measure the quality and accuracy of the generated captions compared to reference captions. Our attack causes a severe performance drop across all metrics. On COCO, BLEU-2 falls from 0.6212 to 0.2004, signaling a collapse in bigram precision, while CIDEr, which quantifies similarity to human captions, drops from 1.2912 to 0.0716. The degradation extends to Flickr30k, further demonstrating the vulnerability of captioning models, especially those relying on foundation models like CLIP, under adversarial attacks. A key distinction between our method and existing adversarial approaches lies in what is being attacked. Even though state-of-art methods, such as SGA [29] and Co-Attack [48], optimize adversarial perturbations by jointly targeting both the

| Metric | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDER | SPICE |
|---|---|---|---|---|---|---|---|---|
| **COCO Dataset** | | | | | | | | |
| **Clean** | 0.7595 | 0.6212 | 0.4877 | 0.3725 | 0.2819 | 0.5888 | 1.2912 | 0.2255 |
| **Co-Attack** [48] | 0.6654 | 0.5220 | 0.3925 | 0.2895 | 0.2380 | 0.5264 | 0.9810 | 0.1790 |
| **SGA** [29] | 0.6635 | 0.5012 | 0.3676 | 0.2653 | 0.2244 | 0.5071 | 0.8806 | 0.1644 |
| **Ours** ($\mathcal{L}_{atn}$) | 0.4022 | 0.2025 | 0.1005 | 0.0600 | 0.0938 | 0.3039 | 0.0788 | 0.0261 |
| **Ours** ($\mathcal{L}_{emb}$) | 0.4815 | 0.2766 | 0.1571 | 0.0934 | 0.1272 | 0.3527 | 0.1936 | 0.0590 |
| **Ours** ($\mathcal{L}_{comb}$) | **0.4000** | **0.2004** | **0.0983** | **0.0558** | **0.0878** | **0.2949** | **0.0716** | **0.0252** |
| **FLICKR Dataset** | | | | | | | | |
| **Clean** | 0.6997 | 0.5392 | 0.3970 | 0.2874 | 0.2179 | 0.5058 | 0.6741 | 0.1567 |
| **Co-Attack** [48] | 0.6499 | 0.4693 | 0.3295 | 0.2287 | 0.1903 | 0.4563 | 0.5198 | 0.1330 |
| **SGA** [29] | 0.6169 | 0.4266 | 0.2900 | 0.1948 | 0.1723 | 0.4308 | 0.4180 | 0.1127 |
| **Ours** ($\mathcal{L}_{atn}$) | 0.3846 | 0.1770 | 0.0757 | 0.0388 | 0.0713 | 0.2632 | 0.0350 | 0.0197 |
| **Ours** ($\mathcal{L}_{emb}$) | 0.4892 | 0.2889 | 0.1765 | 0.1135 | 0.1173 | 0.3385 | 0.1361 | 0.0583 |
| **Ours** ($\mathcal{L}_{comb}$) | **0.03598** | **0.1640** | **0.0705** | **0.0359** | **0.0662** | **0.2463** | **0.0332** | **0.0183** |

Table 3. Impact of adversarial attacks on image captioning metrics for the BLIP-2 model[2] with $\epsilon = 8/255$. Evaluated on 1,000 random images from COCO (val) and FLICKR datasets. Adversarially perturbed images are generated using the CLIP-ViT-L model.

text and image modalities, their focus is exclusively on perturbing the embedding space. In contrast, we attack both embeddings and the attention mechanism simultaneously. This dual disruption results in a more severe degradation of model performance. In Fig. 4, we show the effects of adversarial perturbations for the captioning task using COCO and the BLIP-2 model. The first row displays adversarial examples generated with a perturbation magnitude of $\epsilon = 8/255$. These adversarial examples are visually indistinguishable from the original images but induce significant errors in the model's predictions. The second row shows the original captions generated by BLIP-2 for clean images, where the model accurately describes the content. In contrast, the third row presents the adversarial captions generated after applying the attention-based perturbations, which lead the model to produce inaccurate and unrelated descriptions. We also report captions generated using SGA Attack and CO-Attack.

**Zero-Shot classification** In Tab. 4, we evaluate the performance of different adversarial attack methods on zero-shot classification tasks. Following prior work [42], we test

[1] We used the "BLIP-2-t5/pretrain-flant5xl-vitL" variant.

[2] https://huggingface.co/Salesforce/blip2-flan-t5-xl

| | | | | | |
|---|---|---|---|---|---|
| Predicted caption (unattacked) | A person on skis standing on a snow covered slope | two girls holding donuts up to their eyes | A pizza with arugula and prosciutto on top | A plate of chicken, broccoli and rice | A tray of cinnamon rolls in an oven |
| Ours ($\mathcal{L}_{atn}$) | a photograph of a glass of ice with a red light in it | A picture of a person in a pool of water | A bunch of grapes in a glass of water | A painting of a woman with a flower in her hair | A black and white image of a black and white image of a black and white image of a |
| SGA Attack [29] | A group of people on skis on a snow covered slope | A man and a woman drinking a donut | A plate of food on a wooden cutting board | A plate of food on a table | A piece of bread in an oven |
| Co-Attack [48] | A person on skis on a snowy slope | A girl wearing a pink shirt | A sandwich on a plate | A plate of food on a table | a black and white picture of a tooth brush in a dark room |

Figure 4. Comparison of adversarial and original COCO images with predicted and adversarial captions from BLIP-2. Adversarial examples were crafted with $\epsilon = 8/255$. The first row shows the predicted captions, the remaining rows show captions generated by attacking with: Our Method (second row), SGA (third row) and Co-attack (fourth row).

| **Attacks** | **ViT-B/16** | **ViT-B/32** | **T** | **I** |
|---|---|---|---|---|
| **NES*** [17] | 42.80% | 41.00% | - | ✓ |
| **SPSA*** [41] | 43.50% | 40.50% | - | ✓ |
| **FGSM** [14] | 6.70% | 10.20% | ✓ | ✓ |
| **DeepFool** [32] | **0%** | **0%** | ✓ | ✓ |
| **BIM** [23] | **0%** | **0%** | ✓ | ✓ |
| **MIM** [9] | **0%** | **0%** | ✓ | ✓ |
| **Ours** ($\mathcal{L}_{atn}$) | 1.00% | 1.20% | - | ✓ |
| **Ours** ($\mathcal{L}_{emb}$) | 0.10% | 0.10% | - | ✓ |
| **Ours** ($\mathcal{L}_{comb}$) | **0%** | **0%** | - | ✓ |

Table 4. Zero shot accuracy on ImageNet using $\epsilon = 8/255$. Black box attacks denoted with $^*$.

the zero-shot capability of CLIP with different visual backbones, namely ViT-B/16 and ViT-B/32. In the tables, we report the classification accuracy after the attack. The evaluations are performed on 1,000 random images from the ImageNet 2012 [37] dataset, with an adversarial perturbation budget of $\epsilon = 8/255$. Notably, our proposed approach effectively attacks all images, yielding a 0% classification accuracy, as well as DeepFool, BIM and MIM. The only white-box attack that is not able to achieve a perfect misclassification is FGSM (6.7% and 10.2% on ViT-B/16 and ViT-B/32). Black-box attacks instead are less effective as they do not manage to lower the accuracy below 40%. It has to be noted that all white-box competitors have access to both images and text during the attack, contrary to our method that only attacks images, without any knowledge about text. We specify whether a model has access to text (**T**) and/or images (**I**) in Tab. 4.

**Depth Estimation** For depth estimation, we investigate the impact of the proposed attack on models[3] that use DINOv2. In Tab. 5, we present the results on the NYU-Depth v2 dataset [40] for our method compared to the task-specific attack DepthPGD [22], which attacks the model's head pixel-wise. We report the post-attack Root Mean Squared Error (RMSE), which significantly increases compared to the one of the unattacked models. Although our approach does not outperform DepthPGD, it is agnostic to the model architecture and, as shown in Fig. 5, it introduces significant distortions. It is interesting to notice that our attention-based attack obtains the best results compared to our other proposed variations. In fact, the attack that targets only the embedding does not prove to be as effective, even in combination with the attention loss. We impute this to the fact that depth estimation is a very fine-grained task, where rather than compressing information into a latent vector (as in, say, classification), the purpose is to capture spatial relations among pixels. Attacking attention hinders the capabilities of the model to perform such reasoning.

| **Attacks** | **ViT-S/14** | **ViT-B/14** | **Task-agnostic** |
|---|---|---|---|
| *No attack* | 0.49 | 0.46 | - |
| ***DepthPGD*** [22] | **2.60** | **2.74** | - |
| **Ours** ($\mathcal{L}_{atn}$) | 1.50 | 1.35 | ✓ |
| **Ours** ($\mathcal{L}_{emb}$) | 1.44 | 1.07 | ✓ |
| **Ours** ($\mathcal{L}_{comb}$) | 1.46 | 1.34 | ✓ |

Table 5. RMSE for depth estimation for different attacks on the NYU-Depth v2 dataset. Attack budget $\epsilon = 8/255$.

**Semantic Segmentation** Similarly to depth estimation, for semantic segmentation, we use ViT models[4] that utilize DINOv2. In Tab.6, we compare our method against SegPGD [22], a task-specific attack designed to disrupt segmentation performance by directly targeting the model's output. We evaluate on the ADE20K dataset [51], measur-

[3] ViT-S/14 and ViT-B/14 (1 layer)

[4] ViT-S/14 and ViT-B/14 (linear)

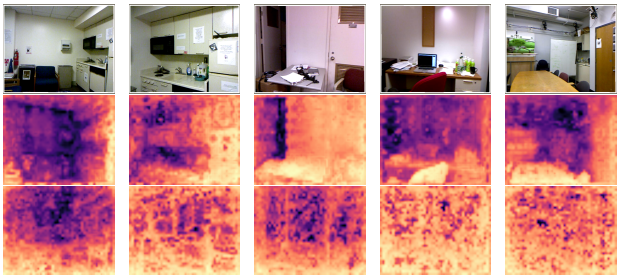

Figure 5. Estimated depths with unattacked (second row) and attacked (third row) images using DinoV2 ViT-S/14 model.

| Attacks | ViT-S/14 | ViT-B/14 | Task-agnostic |
|---|---|---|---|
| *No attack* | 0.420 | 0.450 | - |
| ***SegPGD*** [22] | 0.010 | 0.010 | - |
| **Ours** ($\mathcal{L}_{atn}$) | **0.008** | 0.010 | ✓ |
| **Ours** ($\mathcal{L}_{emb}$) | 0.026 | 0.044 | ✓ |
| **Ours** ($\mathcal{L}_{comb}$) | **0.008** | **0.009** | ✓ |

Table 6. mIoU for semantic segmentation for different attacks on ADE20K. Attack budget $\epsilon = 8/255$.

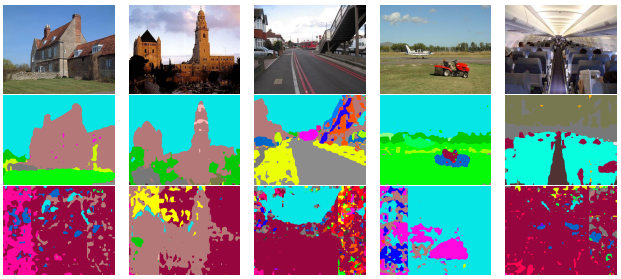

Figure 6. Semantic segmentations with unattacked (second row) and attacked (third row) images with the DinoV2 ViT-S/14 model.

ing performance in terms of mean Intersection over Union (mIoU). Although SegPGD achieves near-complete degradation, reducing mIoU to 0.01, our approach is able to slightly improve, lowering it to 0.008. Again, attention-based attacks prove to be more effective than embedding-based ones, confirming the trend of Tab. 5. In Fig. 6, we show qualitative results of semantic segmentation by the DinoV2 model for clean and attacked images.

**Attacking Attention vs. Attacking Embeddings** From the experiments, it emerges that different downstream tasks are affected in different ways depending on which parts of the attention layers we target. It is interesting to notice that "semantic" tasks, such as classification (Tab. 1) and cross-modal retrieval (Tab. 2) are more easily disrupted by attacking embeddings rather than attention. This is understandable, as token-to-token relationships are not so important for the task as instead the overall representation of the image is. On the contrary, more fine-grained tasks such as captioning (Tab. 3), depth estimation (Tab. 5) and segmentation (Tab. 6), suffer more when attacked using attention since spatial relations between objects or pixels have a direct effect on the outputs. Using a combination of both results to be effective throughout most downstream tasks.

| Target | Method | Image-to-Text | | | Text-to-Image | | |
|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| $\text{CLIP}_{\text{ViT}} \rightarrow \text{ALBEF}$ | Sep-Attack | 2.50 | 0.40 | 0.10 | 4.93 | 1.44 | 1.01 |
| | Co-Attack | 2.50 | 0.60 | 0.20 | 5.80 | 1.78 | 1.11 |
| | SGA | **3.86** | **0.70** | **0.30** | **7.69** | **2.73** | **1.52** |
| | Ours ($\mathcal{L}_{atn}$) | 2.61 | 0.20 | 0.20 | 5.29 | 1.29 | 0.99 |
| | Ours ($\mathcal{L}_{emb}$) | 2.92 | 0.40 | 0.20 | 5.05 | 1.54 | 0.99 |
| | Ours ($\mathcal{L}_{comb}$) | 2.40 | 0.40 | 0.10 | 5.10 | 1.54 | 0.97 |
| $\text{CLIP}_{\text{ViT}} \rightarrow \text{TCL}$ | Sep-Attack | 4.85 | 0.20 | **0.20** | 8.17 | 2.27 | 1.46 |
| | Co-Attack | 5.27 | 0.40 | **0.20** | 9.12 | 2.75 | 1.48 |
| | SGA | **6.43** | 0.60 | **0.20** | **10.93** | **3.47** | **2.05** |
| | Ours ($\mathcal{L}_{atn}$) | 5.42 | 0.60 | **0.20** | 10.55 | 2.93 | 1.78 |
| | Ours ($\mathcal{L}_{emb}$) | 5.63 | **0.70** | 0.10 | 10.59 | 2.91 | 1.76 |
| | Ours ($\mathcal{L}_{comb}$) | 5.53 | **0.70** | 0.10 | 10.76 | 2.93 | 1.66 |
| $\text{CLIP}_{\text{ViT}} \rightarrow \text{CLIP}_{\text{CNN}}$ | Sep-Attack | 5.36 | 1.16 | 0.72 | 8.44 | 2.35 | 1.54 |
| | Co-Attack | 7.66 | 1.90 | 1.44 | 9.37 | 3.90 | 2.53 |
| | SGA | **11.24** | **5.39** | **2.68** | **15.68** | **6.88** | **5.08** |
| | Ours ($\mathcal{L}_{atn}$) | 6.13 | 1.48 | 1.24 | 8.30 | 3.30 | 2.12 |
| | Ours ($\mathcal{L}_{emb}$) | 6.64 | 1.59 | 0.82 | 8.75 | 3.06 | 1.83 |
| | Ours ($\mathcal{L}_{comb}$) | 6.39 | 1.80 | 0.93 | 8.58 | 2.86 | 2.30 |

Table 7. Transferability experiment on Flicr30K dataset.

**Transferability** Adversarial transferability is a crucial yet challenging aspect of adversarial robustness, particularly important in real-world scenarios. In Tab. 7, we compare our method against existing adversarial attacks, including SGA [29] and Co-Attack [48], for image-to-text and text-to-image retrieval tasks on Flickr30k. We use as source model CLIP with ViT-B/16 backbone and as target model ALBEF, TCL and CLIP with a CNN-based backbone (ResNet101). While SGA achieves the highest transfer success rates, it is important to note that this advantage largely stems from its joint optimization with text, allowing it to perturb both modalities. However, the absolute retrieval degradation remains relatively low, limiting its threat level in real-world scenarios. This suggests that adversarial transferability remains an open problem requiring further investigation.

## 5. Conclusions

We presented a novel task-agnostic adversarial attack method tailored for attention-based vision foundation models. Our approach targets both the attention mechanism and the embedding space within these models, resulting in significant disruptions of performance across a variety of downstream tasks. Our experiments, conducted on widely-used datasets show that our method achieves high attack success rates using only image perturbations, without the need for textual data or joint multi-modal optimization. The simplicity and effectiveness of the proposed methods highlights potential issues of relying on widespread foundation models.

## References


[1] Ahmed Aldahdooh, Wassim Hamidouche, and Olivier Deforges. Reveal of vision transformers robustness against adversarial attacks. *arXiv preprint arXiv:2106.03734*, 2021. 2

[2] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in neural information processing systems*, 34:26831–26843, 2021. 3

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. 3

[4] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 3

[6] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[8] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 5

[11] Stanislav Fort. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, 2021. 3

[12] Stanislav Fort. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations, 2021. 3

[13] Yonggan Fu, Shunyao Zhang, Shang Wu, Cheng Wan, and Yingyan Lin. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*, 2022. 3

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3, 5, 7

[15] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? In *European Conference on Computer Vision*, pages 404–421. Springer, 2022. 3

[16] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. Sa-attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. *arXiv preprint arXiv:2312.04913*, 2023. 3, 5, 6

[17] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. 7

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Tom Duerig, and Claire Song. Scaling up vision-language pretraining. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[19] Gihyun Kim and Jong-Seok Lee. Analyzing adversarial robustness of vision transformers against spatial and spectral attacks. *arXiv preprint arXiv:2208.09602*, 2022. 2

[20] Juyeop Kim, Junha Park, Songkuk Kim, and Jong-Seok Lee. Curved representation space of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13142–13150, 2024. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[22] Antoni Kowalczuk, Jan Dubiński, Atiyeh Ashari Ghomi, Yi Sui, George Stein, Jiapeng Wu, Jesse C Cresswell, Franziska Boenisch, and Adam Dziedzic. Benchmarking robust self-supervised learning across diverse downstream tasks. *arXiv preprint arXiv:2407.12588*, 2024. 3, 7, 8

[23] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 7

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 6

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 5

[27] Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, and Dacheng Tao. Exploring the relationship between architecture and adversarially robust generalization. *arXiv preprint arXiv:2209.14105*, 2022. 3

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[29] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 2, 3, 5, 6, 7, 8

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3, 5

[31] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7838–7847, 2021. 3

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 7

[33] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. *arXiv preprint arXiv:2106.04169*, 2021. 2

[34] David A Noever and Samantha E Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *arXiv preprint arXiv:2103.10480*, 2021. 3

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 7

[38] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 2

[39] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 1

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 7

[41] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International conference on machine learning*, pages 5025–5034. PMLR, 2018. 7

[42] Chenguang Wang, Ruoxi Jia, Xin Liu, and Dawn Song. Benchmarking zero-shot robustness of multimodal foundation models: A pilot study. *arXiv preprint arXiv:2403.10499*, 2024. 6

[43] Youze Wang, Wenbo Hu, Yinpeng Dong, and Richang Hong. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *arXiv preprint arXiv:2308.12636*, 2023. 2, 3

[44] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2668–2676, 2022. 2

[45] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 3

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5

[47] Lu Yuan, Qibin Hou, Zihang Jiang, Zhe Feng, Mingfei Cheng, Abner Wan, Jiadong Xie, Varun Kumar, Hongyu Shi, Dongdong Yu, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[48] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 3, 5, 6, 7, 8

[49] Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16415–16424, 2023. 2

[50] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. 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*, pages 862–871, 2024. 6

[51] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 7