As Firm As Their Foundations: Creating Transferable Adversarial Examples Across Downstream Tasks with CLIP

Anjun Hu[1,2]
anjun.hu@eng.ox.ac.uk
Jindong Gu[1]
jindong.gu@outlook.com
Francesco Pinto[1]
francesco.pinto@eng.ox.ac.uk
Konstantinos Kamnitsas[2]
konstantinos.kamnitsas@eng.ox.ac.uk
Philip H. S. Torr[1]
philip.torr@eng.ox.ac.uk

[1] Torr Vision Group
University of Oxford
Oxford, UK

[2] Biomedical Image Analysis Group
University of Oxford
Oxford, UK

Abstract

Foundation models pre-trained on web-scale vision-language data, such as CLIP, are widely used as cornerstones of powerful machine learning systems. While pre-training offers clear advantages for downstream learning, it also endows downstream models with shared adversarial vulnerabilities that can be easily identified through the open-sourced foundation model. In this work, we expose such vulnerabilities among CLIP's downstream models and show that foundation models can serve as a basis for attacking their downstream systems. In particular, we propose a simple yet alarmingly effective adversarial attack strategy termed Patch Representation Misalignment (PRM). Solely based on open-sourced CLIP vision encoders, this method can produce highly effective adversaries that simultaneously fool more than 20 downstream models spanning 4 common vision-language tasks (semantic segmentation, object detection, image captioning and visual question-answering). Our findings highlight the concerning safety risks introduced by the extensive usage of publicly available foundational models in the development of downstream systems, calling for extra caution in these scenarios.

1 Introduction

Foundation models that combine both vision and language modalities [13, 31, 33, 43, 74] are becoming increasingly popular, with CLIP [64] standing out as a prime example. CLIP is extensively used by various downstream models, offering comprehensive cross-modality semantics to enhance a wide range of tasks such as open-vocabulary segmentation (OVS) [46, 87, 92], open-vocabulary object detection (OVD) [82, 98, 101], image captioning (IC) [34, 53, 59] and visual question answering (VQA) [4, 38, 72].

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

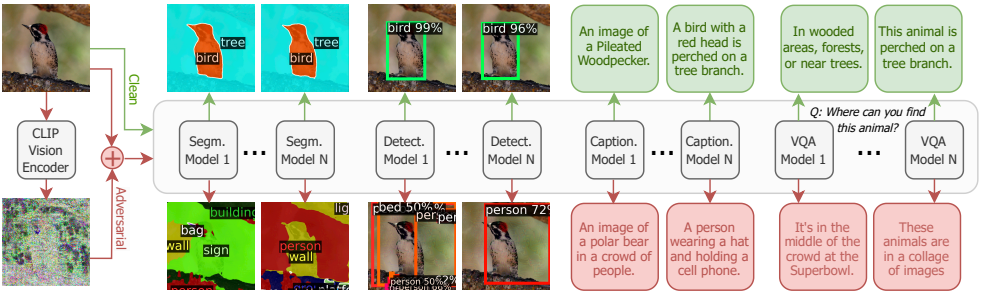

Figure 1: Given a clean image (top left), attackers can leverage the open-sourced CLIP vision encoder to find imperceptible input perturbations (bottom left, magnified by 30× for visualisation) that distort CLIP's intermediate features. These perturbations are added to the original image to construct an adversarial sample that can simultaneously fool many downstream models intended for various tasks. Highly performant downstream models suffer significant performance degradation (bottom row) under such attacks.

While the usage of CLIP yields unprecedented performance improvements across many tasks, in this work, we show that such practice also introduces additional safety risks in downstream systems: since downstream models rely heavily on the semantics acquired during contrastive vision-language pretraining, adversarial perturbations that can disrupt such pre-trained semantic knowledge could significantly hamper these models' performance regardless of their architectural design or intended tasks. This implies that these models tend to share vulnerabilities to the same semantics-breaking adversarial examples, or that these adversaries demonstrate high *transferability* across these downstream models. These adversaries pose significant safety risks as they make the downstream systems more prone to transfer-based attacks (i.e. more likely to be fooled by adversarial examples devised on models other than themselves). Such risks are exacerbated by the fact that these highly transferable adversaries can be easily found through the publicly available CLIP model, rendering the downstream systems particularly susceptible to malicious exploitation.

To this end, we demonstrate the extent of such susceptibility by designing an adversarial attack strategy that creates highly transferable cross-task adversaries using only publicly available CLIP vision encoders. Notably, we show that an attacker with access to an off-the-shelf version of the CLIP vision encoder can substantially compromise a wide range of downstream models, including those that use an encoder with a different architecture than the one accessed by the attacker. Our contributions can be summarised as follows:

- We show that adversarial examples built with off-the-shelf CLIP vision encoders can fool a wide range of downstream models in a task-agnostic manner, evidencing the propagation of adversarial vulnerability from open-sourced foundation models to downstream systems as a significant safety threat.
- We introduce Patch Representation Misalignment (PRM), a simple yet alarmingly effective adversarial attack strategy which seeks input perturbations that distort CLIP vision encoders' intermediate representations to undermine the predictive capability of downstream systems.
- Comprehensive experiments are conducted to show that this technique induces significant performance degradation in more than 20 target models spanning four common vision-language tasks, outperforming competing baselines by a significant margin.

While our experiments explore this phenomenon for CLIP, the extensiveness of our results and the alarming effectiveness of the proposed method calls for further efforts to understand whether this phenomenon extends to other foundational models. Our observations indicate that extreme caution should be taken to address the vulnerability introduced by the usage of foundational models in the development of novel systems.

2 Related Works

Adversarial Transferability. Adversarial attacks introduce imperceptible perturbations to models' input data with the intention of causing degradation in model performance [73]. Early attack strategies such as [22, 50] operate under the assumption of having white-box access to the target model. More recent works [39, 60] point out that such assumptions about target model accessibility are often unrealistic and subsequently show that it is possible to craft adversarial perturbations that *transfer* across multiple models in a black-box fashion (i.e., without requiring access to the target's parameters or gradients). Such phenomenon of adversarial transferability, being the basis for black-box transfer attacks, has thus gained substantial research interest due to its significant implications for model safety [23, 100]. Crafting transfer-based attacks typically requires access to a surrogate model that shares some similarity with the target model, which could be in terms of model architecture, optimisation objectives, or training data. Having significant differences between the surrogate and the target model in these aspects may result in limited attack transferability [3, 81] though recent efforts have been made to foster stronger transferability despite these differences [11, 56, 63, 79, 95]. In this work, we show that the extensive use of open-sourced foundation models in the development of other systems creates vulnerabilities that can be maliciously leveraged to improve attack transferability among downstream systems that differ significantly in architectures and training paradigms.
Inspired by prior works on cross-task attacks [44, 55, 68] and feature-based attacks [21, 27, 28, 30, 69, 69, 80], we design a simple yet highly effective attack strategy that creates adversarial examples by finding input perturbations that induce maximal distortions in the surrogate model's internal representations. Different from prior works that focus on distorting the magnitude or variance of model features, we introduce a more refined objective that treats each patch independently and emphasises patch-wise directional misalignment between clean and adversarial features. This design serves to induce stronger semantic distortion in the CLIP feature space that densely covers all image regions and thereby produces more destructive adversaries compared to baseline methods.

Foundation Models as Attack Surrogates. Recently, there has been a growing focus on investigating and improving the robustness of vision-language foundation models [5, 7, 18, 58, 62, 70, 76, 102, 104]. In particular, several studies have proposed strategies that use vision-language foundation models to devise attacks that transfer to other foundation models [42, 71, 94, 99, 103]. While existing works primarily concern the robustness of foundation models themselves, we focus on the fact that foundation models can serve as a basis for attacking their downstream systems [47]. Our study extends the investigation on foundation model robustness to their downstream systems, which frequently engage in direct interactions with end users and can have a substantial negative impact on real-world applications if their integrity is compromised.

Robustness of Pre-training and Downstream Models. Several existing works have explored the inheritance of certain robustness properties in various transfer learning settings [25, 68, 78, 90, 96]. In particular, how various forms of weaknesses and limitations inherited from pre-training can influence downstream tasks [10, 57] have garnered growing research interest. Building upon this line of investigation, our study demonstrates the propagation of adversarial vulnerabilities from foundation models to downstream models, remarking the trade-off between utility and safety when using foundation models in the development of downstream systems.

3 Methods

Preliminaries, Notations and Threat Model. Given a clean input image sample x, a prediction ground truth y and a target model $\mathcal{M}$ which is under attack, we consider an untargeted adversarial attack which aims at producing an adversarial sample $x' = x + \delta$ by adding an imperceptible perturbation δ to x such that the target model $\mathcal{M}$ would produce an incorrect output prediction when given the adversarial sample x' as input, i.e. $\mathcal{M}(x') \neq y$. To ensure that the attacks remain imperceptible, δ is often constrained by a perturbation budget ε such that the adversary stays within the ε-neighbourhood of the original input. In this work, we specifically consider L_∞-norm bounded attacks (i.e. $\|x - x'\|_\infty \leq \varepsilon$). We assume that an attacker cannot query $\mathcal{M}$ nor access the parameters or gradients of $\mathcal{M}$. The attacker, however, has thorough knowledge about another model $\mathcal{F}$, referred to as a *surrogate* model, which is expected to share some common characteristics as $\mathcal{M}$. This allows the attacker to design algorithmic processes that can identify input perturbations that change the target network's prediction. Note that this threat model does not exclude the possibility of $\mathcal{F}$ and $\mathcal{M}$ having overlapping submodules (e.g. due to shared modules acquired from publicly available foundational models).

Attack via Patch Representation Misalignment (PRM). As illustrated in Figure 2, our approach uses off-the-shelf CLIP vision encoders as surrogates and creates adversarial examples by iteratively optimising for input perturbations that induce maximal distortion of the surrogate model's internal representations. Given a surrogate vision encoder $\mathcal{F}$, we denote intermediate features of a clean sample from encoder layer l as $\mathcal{F}_l(x)$ and those of an adversarial sample as $\mathcal{F}_l(x')$. For transformer-based vision encoders [19] built with a series of residual attention blocks, $\mathcal{F}_l(x), \mathcal{F}_l(x') \in R^{1+\lceil HW/d^2 \rceil \times N}$. Each of them can be expanded as a set of N-dimensional embeddings for individual tokens $\mathcal{F}_l(x) = \{f_l^0, f_l^1, \ldots, f_l^{\lceil HW/d^2 \rceil}\}$, corresponding to representation of image patches of size d and one global `CLS` token. We treat each token (both the patch tokens as well as the `CLS` token are involved in the adversarial example generation process) as an independent sample and minimise the cosine similarity between the adversarial embedding f' of each token and its clean counterpart f, which drives adversarial token representations away from their clean counterparts and thereby misaligns them with the correct semantics:

$$\mathcal{L}_{\text{PRM}} = \sum_{l \in L} \sum_{p=0}^{\lceil \frac{HW}{d^2} \rceil} \frac{f_l^p \cdot f_l'^p}{\|f_l^p\| \|f_l'^p\|} \tag{1}$$

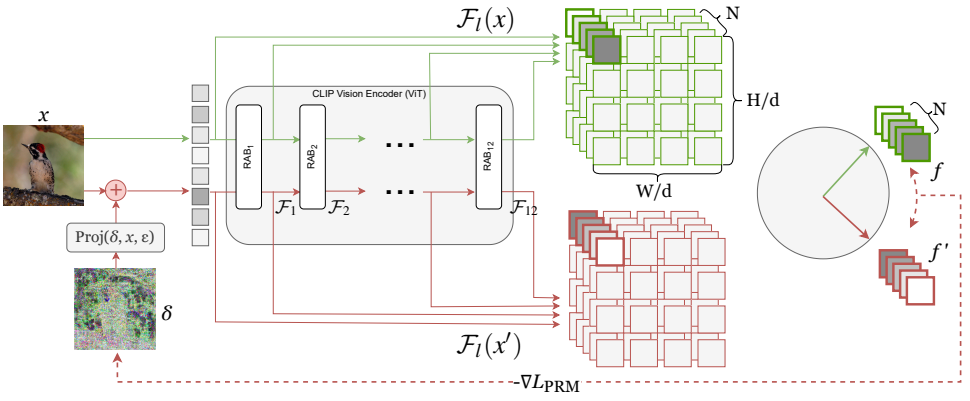

Figure 2: Overview of our attack pipeline. A normal forward pass with clean input is marked in green whereas the forward pass of the adversarial sample is marked in red. Dashed line indicates the flow of loss gradients which are used to update the injected adversarial perturbation. The loss objective minimises the cosine similarity between the adversarial representation of each patch (token) f' and its clean counterpart f along the embedding (ViT) or channel (CNN) dimension of the features. This approach individually diverts each patch representation (indicated by the reversed intensity of the top-left patch representation) to induce semantic distortions in all image regions.

It is straightforward to adapt this loss for convolutional vision encoders built upon ConvNeXt [40] or ResNet [24] blocks. In this case, clean and adversarial intermediate features from layer l are $\mathcal{F}_l(x), \mathcal{F}_l(x') \in R^{\lceil \frac{H}{d} \rceil \times \lceil \frac{W}{d} \rceil \times N}$ where N is the number of filters (channels) and d is a downsampling factor. Rather than viewing the latent features as N *feature maps*, we regard them as a collection of N-dimensional *descriptors* for each spatial element in the features. A similar minimisation of cosine similarity between clean and adversarial descriptors can be done via Eq.(1) where the fs are indexed by their spatial locations.

We include pseudocode in Appendix A and remark that our attack does not require access to task-specific surrogate models or ground truth annotations. It can be applied alongside other transferability-enhancing techniques based on input diversity [17, 36, 84, 105], optimisation [63, 79, 85, 88, 95] and surrogate refinement [9, 26, 51, 56, 97] to further amplify its effectiveness (e.g. the effect of input scale diversity is examined in Appendix C.3).

4 Experiments and Evaluations

4.1 Experimental Setup

We select three baseline adversarial loss objectives, make minor adaptations to suit our problem settings, and test each of them on two surrogate options.

1. **Task-Specific Baseline: Training Loss Maximisation** is the conventional approach taken by the original PGD [50] method which requires a task-specific surrogate and access to ground truths for computing the surrogate loss. We choose two open-vocabulary segmentation models SAN [87] and DeOP [1] as the task-specific surrogate model

and use the corresponding negated surrogate training loss (denoted as $-\mathcal{L}_{\text{SAN}}$ or $-\mathcal{L}_{\text{DeOP}}$ in tables) as the adversarial loss objective.

2. **Cross-Task Baseline 1: Neural Representation Distortion (NRD)** [55] is a cross-task attack strategy based on maximising perceptual dissimilarity between clean and adversarial samples, measured by L_2 distance between VGG-16 conv3.3 features [32]. To make it a comparable baseline applicable to CLIP vision encoders, we simply adapt their objective $\mathcal{L}_{\text{NRD}} = -\sum_{l \in L} \|\mathcal{F}_l(x) - \mathcal{F}_l(x')\|_2$ to ViT and ConvNeXt intermediate features and attack features from all encoder layers (setting L to be all layers).

3. **Cross-Task Baseline 2: Dispersion Reduction (DR)** [44] is a cross-task attack strategy that manipulates adversarial features from one selected CNN layer to become "less dispersed" by minimising their standard deviation. To make it a comparable baseline, we similarly adapt their loss objective $\mathcal{L}_{\text{DR}} = -\sum_{l \in L} \sigma(\mathcal{F}_l(x'))$ to ViT and ConvNeXt intermediate features and attack features from all layers.

We perform attacks across a total of 20 downstream models spanning four common vision-language understanding tasks. Further details can be found in Appendix B.2.

Target Task 1: Open Vocabulary Semantic Segmentation (OVS). Attacks on OVS target models are evaluated using two semantic segmentation benchmarks: Pascal Context (59 classes) [54] and COCO-Stuff (171 classes) [6]. We attack eleven SoTA OVS models of which eight are densely supervised by pixel-level annotation (SAN [87], DeOP [1], ZSSeg-Baseline [86], ZegFormer [16], CAT-seg [14], MasQCLIP [89], SED [83], FC-Clip [92]) and three are weakly supervised by image-level labels or text (SegCLIP [46], TCL [8], SimSeg [91]). Target model performance on the two validation sets is measured with the mean of class-wise intersection over union (mIoU) metric. Results are reported in Table 1 for PC-59 and Table 9 (Appendix E) for COCO-Stuff.

Target Task 2: Open Vocabulary Object Detection (OVD). For evaluating attack efficacy on OVD target models, we use the OV-COCO [37] dataset. We attack six OVD target models (VL-PLM [98], Detic [101], CORA [82], CoDet [48], Object-Centric OVD [49, 66], OV-DETR [93]). Target model performance is measured on the OV-COCO validation set using the AP50 (the average precision at an IoU of 50%) metric. We report model performance on base and novel classes separately in Table 2.

Target Task 3: Image Captioning (IC). We generate adversarially perturbed versions of COCO Captions [12] dataset and evaluate with on three target models (CapDec [59], OpenFlamingo [4], LLaVA [38]) using the 5000-sample test set in Karpathy splits. As shown in Table 3, we evaluate caption quality with several commonly adopted metrics: BLEU score (B@1) [61], METEOR (M) [15], ROUGE-L (R-L) [35] and CIDEr (C) [77].

Target Task 4: Visual Question Answering (VQA). We generate adversarially perturbed versions of OK-VQA [67] and evaluate attack efficacy on two victim models (OpenFlamingo [4], LLaVA [38]), as shown in Table 3. Target model performance is measured with 5046 validation questions from OK-VQA using the VQA accuracy metric.

We set a perturbation budget of $\varepsilon = 8/255$ across all experiments and consider untargeted attacks. All experiments are boosted with random resizing as input augmentation [84, 105] with scaling factors ranging in (0,75, 1.25) and optimized for $N = 250$ (converged) iterations with AdamW [41] optimizer at a learning rate of $\alpha = 0.5/255$. We present our strategy

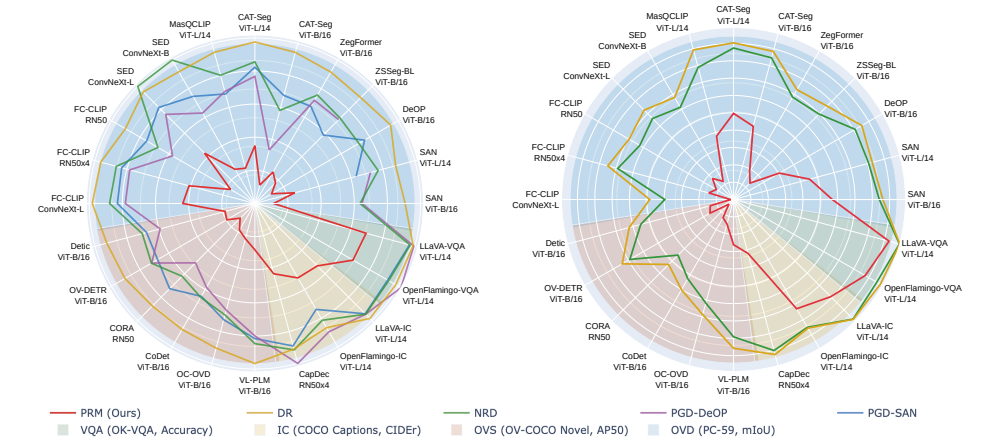

Figure 3: Normalised target model performance (model performance metrics under adversarial attacks divided by metrics on clean samples) of various attack strategies.
Left: using ViT-B/16 (or task-specific baselines that use ViT-B/16 backbone) as surrogates.
Right: using ConvNeXt-L as surrogates. PRM (red line) outperforms baseline methods by a significant margin across all tasks with both surrogate choices.
Each attack strategy corresponds to a line. Each task is indicated by a differently coloured sector (datasets and metrics used for each task are detailed in the 2nd line of the legend).

using either ViT-B/16 or ConvNeXt-L CLIP vision encoders as surrogates with the same loss objective in Eq.1. Attackable layers L are set to all 12 Residual Attention Blocks for the former or all ConvNeXtBlocks in the 3 ConvNeXt stages (excluding stem) for the latter. Using two surrogates helps us to (1) validate the efficacy of our loss design on different architectures; (2) examine transferability in scenarios where the vision encoder architecture of the surrogate matches or mismatches that of the target models' vision encoder.

4.2 Emperical Results and Discussions

Inheritance of Adversarial Vulnerability. As shown in Table 1-3, our attack transfers to a comprehensive collection of victim models with significantly different training paradigms across all four tasks of interest. Our method (PRM) produces significantly more effective adversaries in comparison to baseline methods with both surrogates, as illustrated by Figure 3 where the red line (PRM) shows a substantial shrinkage in performance across all target tasks and models. This provides strong empirical evidence that downstream models could inherit adversarial vulnerabilities from parent foundation models, rendering them more susceptible to transfer attacks.

Transferability Across Vision Encoder Variants. Despite prior research showing limited transferability of adversarial examples between models with significantly different parameters and architectures [3, 81], we notice that substantial transferability can be attained even when a surrogate has a vision encoder architecture that does not match that of the target model. This suggests a significant overlap of non-robust features [81] across vision encoders of various architectures, which is likely due to the common vision-language alignment pre-training process.

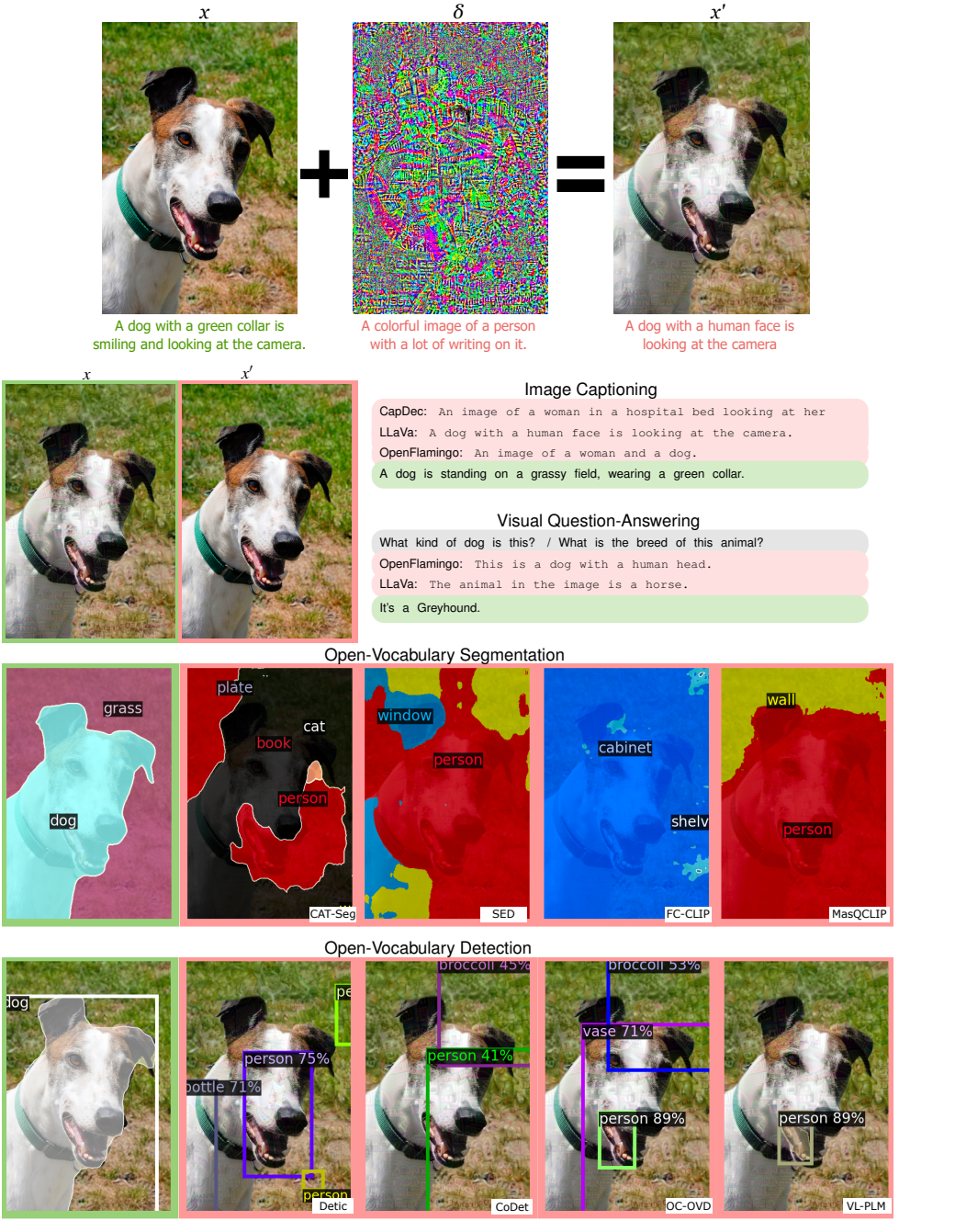

Figure 4: An adversarial example created via PRM (ViT-B/16 CLIP vision encoder as surrogate) can fool various downstream models across various tasks. Clean (x) and adversarial (x') inputs are shown in the top-left corner. Correct predictions are marked with green frames whereas adversarial predictions are marked with red. Downstream models tend to make semantically consistent mistakes (i.e. perceiving a false positive human in the scene).

Table 1: Transfer attack efficacy on OVS target models on Pascal Context 59 Dataset. Model performance is measured in mean Intersection over Union (**mIoU**) where a lower metric indicates higher attack efficacy. Each row (S) represents a surrogate-loss configuration while each column (T) corresponds to a target model. Vision encoder backbones used by each model are specified below model names. Light grey cells indicate scenarios where the source and target models have matching vision encoder architectures. Excluding white-box attacks where source and target models are identical, the **best** attack strategy for each target model within each group is highlighted in **blue**.

	S \ T	$\mathcal{L}_{\text{attack}}$	SAN		DeOP	ZSSeg-BL	ZegFormer	CAT-Seg		MasQCLIP
			ViT-B	ViT-L	ViT-B	ViT-B	ViT-B	ViT-B	ViT-L	ViT-L
		Clean	54.07	57.70	48.65	46.92	41.27	57.4	61.97	58.65
1.0	SAN	$-\mathcal{L}_{\text{SAN}}$	7.12	36.54	37.21	27.42	27.94	38.84	51.00	40.15
	DeOP	$-\mathcal{L}_{\text{DeOP}}$	35.00	41.68	4.57	33.59	29.69	19.25	47.57	41.10
	CLIP ViT-B	DR	49.06	50.78	46.07	42.42	37.83	54.30	60.44	55.44
		NRD	34.34	44.45	35.29	34.17	31.20	33.46	53.01	47.00
		PRM	**6.13**	**14.37**	**5.69**	**8.26**	**9.05**	**6.64**	**21.63**	**12.93**
	CLIP CNeXt-L	DR	46.57	47.76	41.54	35.77	30.18	50.84	56.00	52.39
		NRD	45.45	46.54	39.39	32.71	28.12	48.62	54.23	46.28
		PRM	**30.81**	**26.18**	**14.83**	**6.31**	**7.64**	**25.15**	**30.79**	**22.20**

S \ T	$\mathcal{L}_{\text{attack}}$	SED		FC-CLIP			SegCLIP	TCL	SimSeg	
		CNeXt-B	CNeXt-L	RN50	RN50×64	CNeXt-L	ViT-B	ViT-B	ViT-S	ViT-B
	Clean	57.67	60.87	51.80	55.15	56.99	24.25	33.9	25.81	26.18
SAN	$-\mathcal{L}_{\text{SAN}}$	42.95	49.94	39.07	45.92	47.32	15.45	22.79	21.66	21.89
DeOP	$-\mathcal{L}_{\text{DeOP}}$	36.27	46.31	29.75	43.17	44.58	17.22	13.61	18.21	19.28
CLIP ViT-B	DR	52.81	58.01	46.77	53.14	55.95	12.64	23.94	24.70	25.71
	NRD	57.51	60.87	35.03	47.77	49.99	15.69	15.59	21.50	22.00
	PRM	**13.71**	**26.11**	**8.79**	**22.71**	**24.80**	**1.83**	**4.87**	**11.22**	**11.32**
CLIP CNeXt-L	DR	39.08	44.39	36.04	41.39	27.55	21.75	21.11	23.93	24.50
	NRD	35.33	40.22	32.33	38.14	22.60	21.46	20.29	23.68	24.19
	PRM	**6.99**	**10.44**	**5.97**	**8.17**	**0.93**	**12.64**	**13.10**	**18.94**	**19.38**

Table 2: Transfer attack efficacy on OVD target models on MS-COCO. Model performance is measured in **AP50**. Metrics of Base (B) and novel (N) classes are reported separately.

S \ T	$\mathcal{L}_{\text{attack}}$	Detic		OV-DETR		CORA		CoDet		OC-OVD		VL-PLM	
		ViT-B/16		ViT-B/16		RN50		ViT-B/16		ViT-B/16		ViT-B/16	
		B	N	B	N	B	N	B	N	B	N	B	N
	Clean	51.1	27.8	55.7	30.0	36.8	35.5	52.5	30.6	56.6	40.4	56.4	34.1
SAN	$-\mathcal{L}_{\text{SAN}}$	39.4	18.8	42.8	20.3	26.3	25.8	40.9	19.8	44.1	29.4	46.3	27.8
DeOP	$-\mathcal{L}_{\text{DeOP}}$	38.7	16.5	45.0	21.6	19.8	18.0	39.7	17.9	42.7	26.9	47.0	27.3
CLIP ViT-B	DR	47.5	25.7	52.8	27.2	31.9	31.1	48.7	26.9	52.2	36.5	54.4	32.9
	NRD	40.8	19.6	45.7	45.7	23.0	22.1	42.0	20.1	44.4	28.2	49.1	28.8
	PRM	**20.1**	**5.2**	**21.6**	**5.9**	**7.1**	**4.5**	**21.0**	**5.7**	**21.2**	**8.9**	**22.9**	**9.5**
CLIP CNeXt-L	DR	38.9	17.3	46.0	22.2	21.0	18.8	39.9	18.4	43.5	27.7	48.6	29.2
	NRD	36.8	15.4	44.3	20.7	19.0	16.1	38.0	16.2	40.9	25.0	46.9	27.0
	PRM	**15.4**	**3.9**	**17.1**	**4.7**	**3.4**	**1.4**	**15.6**	**3.7**	**16.0**	**5.9**	**22.0**	**8.9**

Effective and Semantically Consistent Mistakes Across Various Tasks. While significant performance degradation is observed on all four tasks of interest, our attack demonstrates particularly strong efficacy among dense predictors (OVS and OVD models in Tables 1-2), where the impact of full-coverage dense semantic distortion induced by our patch-wise cosine similarity minimisation objective is more prominent. Interestingly, as exemplified by Figure 4, PRM perturbations often incur semantically consistent mistakes across downstream models of various tasks (e.g. in this sample, most models made mistakes by hallucinating a false positive human in this scene). Closer scrutiny reveals that certain perturbations on their own may exhibit semantic meanings [29, 75]: comparing the LLaVA captions of the clean image x, the adversarial sample x' and the perturbation δ, they seem to exhibit an additive relationship in terms of perceived semantic meanings.

Table 3: Transfer attack efficacy on image captioning (COCO-Captions [12]) and visual question-answering (OK-VQA [52]) target models. Caption quality is measured with BLEU score (B@1) [61], METEOR (M) [15], ROUGE-L (R-L) [35] and CIDEr (C) [77]. VQA performance is measured with VQA accuracy.

S \ T	$\mathcal{L}$	CapDec (RN50×4)				OpenFlamingo (ViT-L/14)					LLaVA (ViT-L/14)				
		B@1	M	R-L	C	B@1	M	R-L	C	VQA	B@1	M	R-L	C	VQA
	Clean	68.3	25.2	51.3	92.4	64.9	22.7	50.5	87.3	30.4	72.3	28.7	55.2	106.9	54.1
	SAN	64.2	23.3	48.7	82.3	54.9	18.8	44.5	64.5	28.1	72.7	27.6	53.4	100.7	52.3
	DeOP	66.6	24.5	50.4	93.1	56.6	20.5	47.2	78.1	29.4	70.4	27.9	54.0	101.1	53.8
V B	DR	66.2	23.9	49.7	84.1	57.1	20.5	47.3	75.6	29.8	72.2	28.4	55.0	105.0	53.6
	NRD	66.2	23.9	49.8	84.6	56.4	19.9	46.4	71.1	28.3	71.1	27.7	54.0	101.1	52.6
	PRM	**50.9**	**16.6**	**39.5**	**40.6**	**51.4**	**15.9**	**39.9**	**45.3**	**20.8**	**57.2**	**20.5**	**43.6**	**57.0**	**37.6**
C L	DR	66.2	24.1	49.9	85.5	56.9	20.4	47.0	74.8	29.7	71.9	28.2	54.7	104.4	53.7
	NRD	65.5	23.9	49.6	83.3	56.9	20.5	47.0	74.3	28.9	72.2	28.2	54.8	104.2	53.6
	PRM	**46.0**	**14.5**	**36.3**	**29.7**	**56.4**	**18.8**	**44.4**	**63.5**	**26.6**	**66.0**	**25.2**	**50.0**	**84.7**	**50.2**

4.3 Discussions and Ablations

Which Features to Attack? As mentioned in Sec. 4.1, we choose to attack all tokens (or patches) from all vision encoder layers. This contrasts with previous studies that argue single-layer attacks (only distorting features from one surrogate layer) on mid-layer features would be sufficient [44, 55, 69]. We attempt several alternatives that attack subsets of features and present results in Table 6, Appendix C.2. and found that attacking the ensemble of all layers is more effective.

The Effect of Input Scale Diversity. In the main experiments, we employ random rescaling augmentation for all competing loss objectives. We ablate this component and show results without input scale augmentation in Table 7, Appendix C.3. While this technique significantly amplifies attack efficacy for all attacks by fostering scale invariance, our strategy still outperforms baselines when all are examined in the absence of such augmentations.

The Role of Pre-Training. Various factors in the pre-training process (e.g. data, training objective and optimisation techniques) can contribute to the generalisability of an encoder's feature space. To examine how different pretraining protocols affect an encoder's efficacy as an attack surrogate, we perform our attack using architecturally identical vision encoders pre-trained and fine-tuned with different protocols (Table 8, Appendix C.4). Notably, adversaries crafted with classification-trained encoders exhibit much lower transferability compared to those crafted with a vision-language alignment-pretrained vision encoder of the same architecture, suggesting that the latter encoder serves as a better tool for identifying common non-robust features on which downstream models rely.

Conclusions. This work aims to raise awareness about a currently overlooked but widespread safety vulnerability introduced by the usage of open-sourced foundation models in downstream applications. We show that the downstream models' reliance on pre-trained features can be exploited by attackers, who may devise extremely effective adversarial attacks using solely open-sourced foundation models, compromising the performance of the downstream models independently of the type of task or architecture they adopt. We remark on the extent of such vulnerabilities by showcasing the effectiveness of our attack on 20 different models across 4 common vision-language tasks. While our investigation focuses on CLIP's downstream systems, the alarming efficacy of PRM warrants additional efforts to determine whether similar phenomena apply to other foundational models. We hope our work will encourage further exploration into the safety implications of using foundation models in downstream learning, as well as inspire the development of effective defence strategies or robust learning approaches.

Acknowledgements

This work is supported by the UKRI grant Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. AH is supported by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems Grant EP/S024050/1. We would also like to thank the Royal Academy of Engineering and FiveAI.

References

[1] *Zero-Shot Semantic Segmentation with Decoupled One-Pass Network*, 2023.

[2] Abhishek Aich, Calvin-Khang Ta, Akash A Gupta, Chengyu Song, Srikanth Krishnamurthy, M. Salman Asif, and Amit Roy-Chowdhury. GAMA: Generative adversarial multi-object scene attacks. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=DRckHIGk8qw.

[3] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 888–897, 2018.

[4] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[5] Yuanhao Ban and Yinpeng Dong. Pre-trained adversarial perturbations. *Advances in Neural Information Processing Systems*, 35:1196–1209, 2022.

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[7] 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.

[8] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4489–4498, October 2023.

[10] Hao Chen, Bhiksha Raj, Xing Xie, and Jindong Wang. On catastrophic inheritance of large foundation models. *arXiv preprint arXiv:2402.01909*, 2024.

[11] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[14] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation, 2023.

[15] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[16] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022.

[17] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[18] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

[19] 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.

[20] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3136–3146, October 2023.

[21] Aditya Ganeshan, Vivek BS, and R Venkatesh Babu. Fda: Feature disruptive attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8069–8079, 2019.

[22] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[23] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, and Yao Qin. Initialization matters for adversarial transfer learning. *arXiv preprint arXiv:2312.05716*, 2023.

[26] Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20514–20523, 2023.

[27] Lifeng Huang, Chengying Gao, and Ning Liu. Defeat: Decoupled feature attack across deep neural networks. *Neural Networks*, 156:13–28, 2022.

[28] Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2019.

[29] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[30] Nathan Inkawhich, Kevin Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *Advances in Neural Information Processing Systems*, 33:20791–20801, 2020.

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[32] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[34] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023.

[35] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.

[36] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[39] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[42] 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, 2023.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[44] Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 940–949, 2020.

[45] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[46] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *ICML*, 2023.

[47] Yiqiang Lv, Jingjing Chen, Zhipeng Wei, Kai Chen, Zuxuan Wu, and Yu-Gang Jiang. Downstream task-agnostic transferable attacks on language-image pre-training models. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2831–2836, 2023. doi: 10.1109/ICME55011.2023.00481.

[48] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

[49] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer, 2022.

[50] 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.

[51] Hashmat Shadab Malik, Shahina K Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Adversarial pixel restoration as a pretext task for transferable perturbations. 2022.

[52] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[53] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[54] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[55] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*, 2018.

[56] 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.

[57] Laura F Nern, Harsh Raj, Maurice André Georgi, and Yash Sharma. On transfer of adversarial robustness from pretraining to downstream tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[59] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022.

[60] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[61] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[62] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[63] Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *arXiv preprint arXiv:2210.05968*, 2022.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[65] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023.

[66] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *36th Conference on Neural Information Processing Systems (NIPS)*, 2022.

[67] Mengye Ren, Ryan Kiros, and Richard Zemel. *Advances in neural information processing systems*, 28, 2015.

[68] Shahbaz Rezaei and Xin Liu. A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. *arXiv preprint arXiv:1904.04334*, 2019.

[69] Mathieu Salzmann et al. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962, 2021.

[70] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

[71] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models, 2023.

[72] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[73] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[74] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[75] Guanhong Tao, Shiqing Ma, Yingqi Liu, and Xiangyu Zhang. Attacks meet interpretability: Attribute-steered detection of adversarial samples. *Advances in Neural Information Processing Systems*, 31, 2018.

[76] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

[77] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[78] Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. With great training comes great vulnerability: Practical attacks against transfer learning. In *27th USENIX security symposium (USENIX Security 18)*, pages 1281–1297, 2018.

[79] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[80] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021.

[81] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1360–1368, 2023.

[82] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. *ArXiv*, abs/2303.13076, 2023.

[83] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation, 2023.

[84] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[85] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14983–14992, 2022.

[86] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.

[87] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[88] Xiaojun Xu, Jacky Y Zhang, Evelyn Ma, Hyun Ho Son, Sanmi Koyejo, and Bo Li. Adversarially robust models may not transfer better: Sufficient conditions for domain transferability from the view of regularization. In *International Conference on Machine Learning*, pages 24770–24802. PMLR, 2022.

[89] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masqclip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 887–898, 2023.

[90] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9215–9224, 2022.

[91] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7071–7080, 2023.

[92] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023.

[93] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. 2022.

[94] 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.

[95] 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.

[96] Ziqi Zhang, Yuanchun Li, Jindong Wang, Bingyan Liu, Ding Li, Yao Guo, Xiangqun Chen, and Yunxin Liu. Remos: reducing defect inheritance in transfer learning via relevant model slicing. In *Proceedings of the 44th International Conference on Software Engineering*, pages 1856–1868, 2022.

[97] Anqi Zhao, Tong Chu, Yahao Liu, Wen Li, Jingjing Li, and Lixin Duan. Minimizing maximum model discrepancy for transferable black-box targeted attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8162, 2023.

[98] Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *ECCV*, pages 159–175. Springer, 2022.

[99] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[100] Zhengyu Zhao, Hanwei Zhang, Renjue Li, Ronan Sicre, Laurent Amsaleg, Michael Backes, Qi Li, and Chao Shen. Revisiting transferable adversarial image examples: Attack categorization, evaluation guidelines, and new insights. *arXiv preprint arXiv:2310.11850*, 2023.

[101] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

[102] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023.

[103] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[104] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

[105] Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*, pages 563–579. Springer, 2020.